

Summer 1996

Meanings Underlying Student Ratings of Faculty


Carolyn Ridenour

University of Dayton, cridenour1@udayton.edu

Stephen J. Blatt

University of Dayton, blattste@udayton.edu

Follow this and additional works at: https://ecommons.udayton.edu/eda_fac_pub

 Part of the [Educational Assessment, Evaluation, and Research Commons](#), [Educational Leadership Commons](#), [Higher Education Commons](#), and the [Higher Education and Teaching Commons](#)

eCommons Citation

Ridenour, Carolyn and Blatt, Stephen J., "Meanings Underlying Student Ratings of Faculty" (1996). *Educational Leadership Faculty Publications*. 101.

https://ecommons.udayton.edu/eda_fac_pub/101

This Article is brought to you for free and open access by the Department of Educational Leadership at eCommons. It has been accepted for inclusion in Educational Leadership Faculty Publications by an authorized administrator of eCommons. For more information, please contact frice1@udayton.edu, mschlangen1@udayton.edu.

The Review of Higher Education

Summer 1996, Volume 19, No. 4, pp. 411–433

Copyright © 1996 Association for the Study of Higher Education

All Rights Reserved (ISSN 0162-5748)

Meanings Underlying Student Ratings of Faculty

Carolyn R. Benz

Stephen J. Blatt

PURPOSE OF THE STUDY

The purpose of this study was to examine how undergraduate students interpret the items on a faculty evaluation instrument. Most research on faculty evaluation is quantitative (Marsh and Bailey 1993). Our first study was also quantitative. After we produced a profile of quantitative ratings of faculty by students across all departments in our university in an earlier study, we wanted to go beneath the numbers to their meaning. We designed the present qualitative study to investigate what the items on that form meant to students.

JUSTIFICATION FOR THE STUDY

We began this study with three assumptions. First, institutional strategies to evaluate professors in their role as teachers are increasingly important. Public pressure to account for faculty time and renewed emphasis

Carolyn R. Benz is Associate Professor in the Department of Educational Administration at the University of Dayton in Ohio. She teaches research design and educational statistics and has published in the *American Educational Research Journal* and the *Journal of Educational Research*. Stephen J. Blatt is Associate Professor in the Department of Communication in the same university. He studies the role of communication in organizational settings—including the function of communication in new employee socialization processes and the role of metaphors and narratives in organizational life and relationships—and investigates organizational culture through interaction patterns.

on teaching rather than research continue to escalate. Student ratings are the prevalent teaching evaluation tool. Second, many studies have shown common findings about the characteristics of faculty in classrooms as rated by students. We cite several of these studies and suggest that it is time to delve into the underlying meanings of the items on those rating forms. Third, and very important to our purpose, was the validity question. How valid are student ratings of faculty for institutional purposes of promotion, tenure, and merit pay raises? It was this question that motivated our studies in the first place. The validity issue is the major focus of the study reported in this paper. We justify each of these assumptions next, with references to what others have done and said.

Our interest was driven by the current political and economic scrutiny of faculty evaluation. Policies are being initiated by state legislatures to increase the emphasis on teaching and increase the scrutiny of its effectiveness in public universities (Braskamp and Ory 1994). Ernest Boyer, president of the Carnegie Foundation for the Advancement of Teaching, issued a report recently on the evaluation of faculty work in higher education in which he describes pressure on universities to increase faculty teaching load and to more rigorously account for faculty time and effectiveness (cited in Magner 1994). In times of financial restraint, university administrators and faculty become more diligent in monitoring who gets what in the decreasing pool of money for merit pay raises. Given this heightened emphasis on teaching, our concern was: To what extent are the ways we measure teaching effectiveness valid?

John Centra explains the current use of student evaluation systems as growing out of the "golden age of research on student evaluations" in the 1970s (1993, 52). Questions of validity, reliability, and correlation with learning outcomes have been examined in numerous studies. In Centra's 1993 book, he claims that faculty continue to use questionnaires developed by their institutions. Our university does, too. Denise Magner (1994) reports that student ratings continue to be the only form of faculty teaching evaluation in most institutions. Peer review and portfolio review are much less common, she says.

At our university, student evaluation is a summative assessment activity. Ratings are collected at each term's end, and the results are used for decisions about promotion, tenure, and merit pay. For a summative purpose, student evaluation seems to adopt one of two perspectives: Either teaching effectiveness is assessed globally, using a single overall measure, or teaching is multidimensional and assessment must address many individual dimensions (Ryan, Harrison, and Zia 1993).

With these two perspectives, however, higher education institutions use no universal strategy in applying student ratings to decisions about

faculty teaching effectiveness. Faculty rating forms come in many versions. First, the items range from those that describe specific observable behaviors (sometimes referred to as “low-inference” items) to items alluding to nonobservable qualities (sometimes referred to as “high-inference” items). Harry Murray (1985), for one, contends that teaching effectiveness in college classrooms is predictable from low-inference classroom behaviors by the instructor. He speculates whether characteristics of teaching (like fairness and rapport) can be understood in terms of specific observable behaviors. He further claims that teaching can more likely be improved if the feedback focuses on observable changeable behaviors (low-inference) rather than on “intractable generalities” (high inference) (33).

Despite the wide variety of evaluation forms used from one institution to the next, there are strong commonalities. Bringing together the results of numerous factor analyses studies, Centra identified a set of qualities commonly assessed on student rating forms. They are:

1. Organization, planning, or structure
2. Teacher-student interaction or rapport
3. Clarity, communication skill
4. Work load, course difficulty
5. Grading and examinations, assignments
6. Student learning, self-ratings of accomplishments (1993, 57)

Most research on faculty ratings has been driven by the quantitative research paradigm; qualitative studies appear much less frequently in the literature. Studies like one by Tiberius and his associates (1989) on the discussion mode of student feedback demonstrate one of the few qualitative approaches to evaluation that used a qualitative approach to analysis. From the many studies that have been conducted, some conclusions are warranted.

For instance, evidence points to commonalities in student perceptions across studies. However, this pattern may be merely an unsurprising result of the common qualities being rated by university evaluation systems. Kenneth Feldman (1976) reported that when students came up with characteristics of their effective teachers without being given a list of qualities to choose from, they focused on concern and respect for students and impartiality. When given lists of descriptors, students chose intellectual challenge and sensitivity to class level and progress. Janet Donald (1985) reports on differences in desired teaching qualities across disciplines and fields of study. The three characteristics mentioned most frequently were among the top five in all fields: subject mastery, being well prepared and orderly, and encouraging student questions and opinions. She also concludes that organization and clarity, instructor knowledge, and enthusiasm

and stimulation are consistent across many studies. Herbert Marsh (1986) reports that two qualities (clear well-organized presentations and enthusiasm) correlate with faculty effectiveness ratings across international cultures. Ten years earlier, Robert Wilson et al. (1975) found that, in addition to classroom behaviors, faculty interaction with students outside of class was related to high student ratings.

On both Feldman's (1976) and Donald's (1985) lists is subject mastery. However, Peter Seldin's (1980) review of research found that instructor knowledge was less frequently listed on ratings forms. He suggests that instructor knowledge can perhaps be better judged through peer evaluation than student evaluation. Murray (1985) reports on a study by Alan Tom and Cushman in 1975 that showed significant correlations between student ratings of the amount learned and several teacher behaviors, including the use of real-life examples. In an earlier study of 28,000 student ratings we found that three items were the most significant predictors of students' high ratings on having "learned a lot." They were that the instructor was "interesting," the course "met the objectives," and the instructor was "well prepared." (Benz and Blatt 1995).

Student ratings on the same professor over several classes have shown high reliability, according to Murray (1985); but, he maintains, validity is quite another issue. The validity issue is one we set out to address in this study.

Because student ratings are a *measure* of teaching effectiveness, we need to be assured of the validity of that measurement. Validity, in essence, answers the question: Are we measuring what we purport to measure? (Newman and Newman 1992) There are numerous definitions of validity in educational research. Measurement experts select a type of validity that serves the purpose of their measure, i.e., consistent with why they are measuring in the first place. For example, a strong estimate of "content validity" would provide assurance that the instrument accurately represents the content it is supposed to measure. On the other hand, assessing "criterion-related validity" would indicate how well a teacher would perform on some criterion, either now or in the future (Popham 1990). Learning outcomes would seem to be a reasonable criterion against which to validate student ratings of faculty. Do student ratings of faculty predict (correlate with) learning outcomes?

Learning outcomes can be measured in a variety of ways: test scores, course exams, gain scores from the pretest to the posttest, plans for further study in the subject matter, and observed classroom performance, as well as GPA. From her review of studies, Donald (1985) concluded that the relationship between students' evaluation of instruction and student learning (no matter how it is measured) is not strong enough to permit student

evaluations to be the only measure of teaching effectiveness. The criterion-related validity is not, therefore, assured. Seldin would concur. In his 1988 article, he summarizes ten things we know from a decade of student evaluation research; and at the top of his list is that no one source of feedback on faculty is sufficient.

These types of validity (content-related, criterion-related) are among the several most frequently used in quantitative validation studies. We wanted to examine a different type of validity—the underlying meaning to students of the items on the rating form. Our purpose might be closest to “construct validity,” a validity that, when assured, gives evidence that the appropriate hypothetical constructs are being measured. In the quantitative paradigm, factor analyses are commonly used to estimate construct validity, a mathematical way to calculate underlying dimensions or constructs (Rummel 1970).

Rather than using the traditional notion of construct validity, we opted for a more phenomenological approach to validity—one from the qualitative paradigm and one grounded in human experience. We selected Harry Wolcott's (1990) offer of the term “understanding” to parallel validity in this study. We were interested in the *understanding* students have of the items on which they evaluate faculty. Perhaps better still is the notion set forth by Judith Goetz and Margaret LeCompte that validity for qualitative researchers often “represents empirical reality . . . or assesses . . . whether constructs devised . . . measure the categories of human experience that occur” (1984, 210). In other words, we asked: Do the responses on the evaluation form measure some agreed-upon reality of classroom experience from the students' point of view?

METHODS

This study is part of a program of four studies of faculty evaluation we have undertaken (Benz and Blatt 1994, 1995; Blatt and Benz 1993, 1994). Undergraduate students at the University of Dayton during two academic years (1992–94) were the research participants. The University of Dayton is a private coeducational school administered by the Society of Mary (the Marianists), a Roman Catholic teaching order, located in Dayton, Ohio. It is primarily an undergraduate institution and includes the College of Arts and Sciences and four professional schools: business administration, education, engineering, and law. Approximately six thousand undergraduates, nearly all residential, are enrolled. There are graduate programs (primarily master's degrees with a few Ph.D. programs) in each college/school, attended by approximately three thousand students who come primarily from the Dayton region.

The Student Evaluation of Faculty instrument contains 27 items (answered on a Likert-type scale) designed to collect three categories of responses: demographics, course ratings, and instructor ratings. We were interested in only eight of these items, the "instructor" ratings: items 10–17 (Table 1); and specifically, whether this instrument is valid for the purposes to which it is put. No validity or reliability studies on the form had been done previously, to our knowledge, although it has been used in most undergraduate classes for almost twenty years.

Data Collection and Analysis

Data collection for this qualitative study took place simultaneously with the regular end-of-term faculty evaluation for fall term 1993. In addition to the standard faculty rating form, we distributed a second form to collect responses for this study. The top of this form contained this explanation: "We ask about your reasoning in giving ratings on the evaluation form. We are interested in how students interpret these items." Below this statement were copied the eight items (10–17) related to the instructor. We asked students to copy their ratings on these items and, after each rating, to answer the question: "Why did you rate this item as you did?"

We sent a memo requesting cooperation to all departments from the provost's office and thence to the department chairs. Ten full-time faculty volunteered their classes. All were tenured or tenure-track faculty, represented eight departments, and provided a total of eighteen classrooms and

TABLE 1
RIGHT ITEMS (10–17) ON THE
STUDENT EVALUATION OF FACULTY FORM

-
10. The instructor prepared well for classes.
 11. The instructor spoke clearly and audibly.
 12. The subject matter was clearly presented by the instructor.
 13. The instructor put material across in an interesting way.
 14. Students were able to express themselves freely as a result of the instructor's openness to their ideas.
 15. The instructor was willing to help students who experienced difficulty in the course.
 16. The instructor respected students as persons.
 17. The instructor was fair in grading examinations and assignments.
-

Response options on bubble sheet: A = strongly agree, B = agree, C = neutral, D = disagree, E = strongly disagree

389 students. (See Table 2 for a description of the participating classrooms.) We collected the data between 30 November and 7 December 1993.

Each of us conducted a narrative analysis of half the responses. First, all completed forms were photocopied so that each researcher could conduct an initial coding of four of the eight items. Using analytic induction, one of us analyzed all 389 student responses to items 10, 11, 12, and 13 while the other analyzed all 389 student responses to items 14, 15, 16, and 17. We attempted to "identify subjective participant constructs," in the words of Goetz and LeCompte (1984), who describe this research analysis strategy.

Analyzing one item at a time, each of us read the students' response and summarized the response in one or more "codes"—words or phrases to characterize the response and retain as much of the original language of the student as possible. These labels were descriptive, not interpretive. As Matthew Miles and Michael Huberman suggest, "These [labels] entail

TABLE 2
PARTICIPANTS BY DEPARTMENT AND INSTRUCTOR

<i>Department</i>	<i>Instructor (and gender)</i>	<i>Course level</i>	<i>Number of Students</i>
Teacher education	Dr. X (m)	200	18
	Dr. X (m)	200	28
	Dr. T (f)	200	27
	Dr. T (f)	200	11
Foreign language	Dr. Q (m)	300	11
	Dr. Q (m)	300	13
Political science	Dr. F (m)	200	37
	Dr. F (m)	200	28
	Dr. L (m)	300	41
Business	Dr. E (m)	200	24
Psychology	Dr. S (f)	400	32
	Dr. S (f)	400	19
Religious studies	Dr. N (m)	400	4
	Dr. N (m)	100	25
	Dr. N (m)	300	17
Engineering	Dr. V (m)	400	21
English	Dr. Z (f)	200	17
	Dr. Z (f)	200	16

no interpretation, but simply the attribution of a class of phenomena to a segment of text" (1984, 56). These codes, too, were not unlike the step of initial coding as described in grounded theorists' work (Glaser and Strauss 1967). We wrote the codes in the right-hand margin of the response form next to the student's answer. Interpretation was a later step.

As a check on reliability, each of us recoded one class of the other's set of responses to determine if wide discrepancies occurred in our interpretations. We selected the set of student responses that was the largest in each group of data. Between the two of us, there were minor differences in the codes assigned and themes defined. Using the rule of staying close to the student's language seemed to account for most differences between us.

We considered taking the numerical ratings into account as we coded the responses but decided against it. To reiterate, students were asked to rate the faculty member on a five-point scale: A B C D E (strongly agree to strongly disagree). We asked why they rated each item as they did. We were interested in the language they used; what words they used in describing their reasoning. We were convinced that, while the reasons for agreeing or disagreeing with a statement like "the instructor prepared well" would probably differ, the substantive content of the words would lead us to a meaning of the item independent of the direction of the rating. The language of all students across the total rating spectrum was our objective. During the initial coding process, we paid little attention to each item's numerical rating, looking instead at the students' language. Our ultimate goal was to interpret the meaning of that language.

A second aggregation of data followed the initial coding of all responses to all items. Here we grouped the responses of each set of students to each of the ten individual instructors. The product here became ten code lists related to item 10, ten lists of codes for item 11, etc. Each list within the stack contained responses to an individual instructor; the number of columns on the sheet represented the number of classrooms used. Combining the codes across all student responses to an item revealed similar repeated meanings from more than one student. In some ways this second step of analysis had similarities to the constant comparative method of qualitative analysis described by Glaser and Strauss (1967). While they suggest noting the code in the margins of each unit of analysis, they are more explicit in that the analyst should compare each incident coded a certain way with previous incidents coded that same way, if any. We were not purists in our procedure; however, we experienced the analysis process as they describe:

Since coding qualitative data requires study of each incident, this comparison can often be based on memory. Usually there is no need to refer to the actual note on every previous incident for each com-

parison. The constant comparison of the incidents very soon starts to generate theoretical properties of the category. The analyst starts thinking in terms of the full range of types or continua of the category, its dimensions, the conditions under which it is pronounced or minimized, its major consequences, its relation to other categories, and its other properties. (106)

We divided the items between us (each analyzing four items) because, in part, we felt that we could develop themes more substantively if we were working with a field of four, rather than eight items. The constant-comparative method of coding relies on the coder's memory for salient themes to emerge, as Glaser and Strauss describe here.

We experienced, too, the dynamic that Glaser and Strauss describe: As categories emerge, we discovered both researcher-constructed categories and also those abstracted from the language of the data. We attempted to maximize the second type. We agreed that the codes closest to the original data (student constructs) tend to be the theoretical components to be explained and the codes constructed by the researcher (our codes) tend to be the explanations.

Each of us wrote memos periodically to capture the emerging coding patterns, writing down all patterns and insights. Following grounded theorists' methods yet further, we discussed the themes emerging from the data. In some instances the patterns were common in both researchers' sets of items—students in both groups mentioned either falling asleep in class (boring) or not falling asleep (interesting), for instance. In other instances, the patterns seemed characteristic of only a single item—for instance, the students' desire to be liked.

Results at this stage can be shown in an example. To item 13 (the instructor presents “the material in an interesting way”), eleven students in one of Dr. S's classes mentioned “videos, visuals, films, slides.” In a second class of Dr. S's, four students used such terms. This process allowed patterns to emerge. This was the extent of our data analysis plan.

We agree with Miles and Huberman that when the qualitative researcher shifts from words to solely number counts, the attention also shifts “from substance to arithmetic, and thereby throws out the whole notion of qualitiveness” (1984, 56). Not wanting to limit ourselves to mere frequency counts, our procedure permitted the patterns to guide our analysis from this point. Those processes are interwoven into the discussion of our results.

RESULTS

Analyzing Student Thinking about Teaching

Thematic interpretations aligned with the eight items make up the bulk of our findings. We discuss each in the section following this initial

discussion of student thinking. The data suggested four interesting patterns of student thought. These patterns included: (1) the fact that students use a variety of evidence in making their ratings, (2) the attributions students make about their ratings, (3) students' understandings about the teaching process, and (4) the students' frequent ambiguity, and what we characterized as their being "pulled into doubt" as they reflected on their classroom experience.

First of all we were struck by the evidence students reported. Students cited the absence of some phenomena as their reason for a rating, e.g., "I was never bored." "She never purposefully [sic] confused us." "I had no problem taking notes." "There was no lesson plan." "He was never late." These are not all similar in kind, however. Never being late is a positive trait, as is lecturing with such a clear structure that students have no problem taking notes. "I was never bored," however, means something different. It suggests something much less positive—not "It was very interesting and exciting" but rather, "At least it wasn't boring." From this meaning one could not conclude that the instructor was interesting.

To what or to whom students attributed their ratings was quite diverse. In some cases, students attributed their rating to the subject matter. ("Statistics cannot be made interesting.") In other cases, students attributed all responsibility to the instructor, both the idea contained in the item and well beyond. For example, to item 10's query about the instructor's preparation, a student responded: "He prepared well but sometimes he prepared too much for one class period. That meant we would have to learn things on our own." To item 13 about whether the instructor presented in an interesting way, two students responded: "Overall, I'd like to give an 'A' rating, but there were two days I fell asleep in class." "Although this is a history class, he did a poor job of getting the class involved and giving the material in a way that can be understood in a fun way." This language seems to mean that the instructor controls all results within the classroom. It is the instructor's "fault" that students had to learn on their own, did not keep them awake, and didn't make it fun. One student remarked that he got "an A, so it must have been made interesting."

In contrast, students sometimes attributed the rating to evidence of their own behavior, not the instructor's behavior. In a number of cases, students based their rating on whether they had fallen asleep or not. The reasoning seemed to be: "If I fell asleep it must have not been interesting; but if I stayed awake all the time, it must have been interesting." Other examples are: "We always knew what was going to be on our test," and "There was never a time I didn't understand what was going on." At times the attribution was to the students in general, e.g., "Unfortunately many days class was not lively. We sat in rows or circles, having a hard time

discussing, maybe that's the class's fault, though." This interesting dichotomy of attribution warrants further examination. To what extent does how we present ourselves as faculty and our subject matter influence students' attribution? To what extent are the ways in which we construct the classroom experience and manage classroom activities related to students' attribution?

Students revealed some interesting, perhaps naive and odd, understandings about teaching. "Because [of] the way class was structured, he did not really have to prepare much for the class. The class was all discussion." Another wrote: "The instructor facilitated many different ways—assignments, readings, and prompted discussions so that everyone clearly and fully processed the information." This second student has a very different sense of the teacher's role vis à vis "discussions." One student, responding to item 13 on clarity of presentation, commented, "Dr. A did not tell us everything though, because I think he wanted us to come to and form our own conclusions. If he told us everything, we wouldn't have had anything to do!" Another student responded to the same item: "The discussion format got bogged down. This is also the fault of the students, but he did little effectively to ameliorate the situation." "We did cover some complex ideas and he was able to break it down to more simpler concepts." Another wrote: "The subject matter was clearly presented but at times it seemed that a stronger background in Spanish history was needed to fully understand the significance of what was taught."

Are there interpretations here suggesting that faculty should engage students in what we called "meta-teaching"? If instructors explicitly presented their teaching strategies, might more students understand, for example, why a discussion seems the most appropriate learning activity? And if an instructor leaves some conclusions open-ended, would it be helpful to students to know that he or she expects each student to form his or her own conclusions?

A very frequent pattern of response came across as ambiguity—a pattern in which students agreed with the item but then, in their written comments, presented a major contradiction to their agreement. Examples are: "The subject matter was presented clearly, but sometimes his claim was not too clear, or in interpreting excerpts from the Constitution, I was totally confused as to what it really meant." Does this student think the instructor presented material clearly or not? Another wrote: "He knew what he was talking about but he used a lot of stats jargon that lost a lot of students." A third commented, "Pretty easy to understand, but sometimes a little unclear or confusing." And a fourth observed, "At times Dr. A would be talking about a certain thing which had different parts, and sometimes it got confusing because he jumped back and forth between

the two ideas. However, this is not necessarily negative because the different ideas are closely related." "He seemed to have a firm grasp of the subject and clearly explained it to the class. There was one time, however, that it seemed the whole class was confused."

Perhaps what we learn from these reasonings is that students feel ambiguous about these items which are presented as if they are concrete and quantifiable. This logic might also show students' tendency to make an over-quick initial judgment ("the subject matter was clearly presented") and then pull back to set conditions on that quick judgment. While such a pattern may be typical only of their assigned task (identifying their reasoning), it is not clear whether a similar pattern of quick judgment followed by qualifiers drives the normal numerical rating process. To the extent that the numerical rating process involves a first-glance judgment similar to these descriptions, there is no opportunity to diminish that circled number with qualifiers as the students could when presented with a request for written information.

Consistent with our intent to uncover meanings at the item level, we discuss each of the eight items separately in the next section, followed by our general discussion of "what is going on" when students rate faculty on this form.

THE MEANING OF EACH ITEM TO STUDENTS

Item 10: "The instructor prepared well for classes."

By far the strongest theme in responses to why students rated faculty as they did on this item was the students' sense that the instructors knew their subject matter. In fact, this was the only strong pattern of responses. Words that reflect this theme were embedded in responses for all ten faculty members. Some wrote that the instructor "knew what she was talking about" or "really knows his stuff." Many times the reason was stated simply as "knows her subject matter." A subordinate pattern was related to "time." Frequently students would comment that the "instructor used all the class time available," or that there was "never any down time." The language seemed to illustrate the fact that an entire class period was "used up," as some said. Students also had a sense that the instructor "was ready to go when class started." ("Ready" was frequently used.) Many referred to such materials as handouts, overheads, and lecture notes as symbols of preparedness. For example: "He had handouts." "He used overheads." "[He or she] lectured with notes/without notes." It was interesting that different students considered either lecture style as evidence of preparation. They also commented about the instructor's "following a plan" or "sticking to the syllabus." In contrast, students who rated the instructor low on this

item, frequently explained that “he didn’t stick to the syllabus,” or “he always got off the track,” or “he forgot materials.”

Item 11: “The instructor spoke clearly and audibly.”

This item might have posed a problem for students. Perhaps because it is a low-inference item compared to most of the others, generally students merely stated, “because he did.” We considered this remark as reasoning that restated the item. While it was not a response that represented a large number of students, it nevertheless was the strongest pattern of response. A second strong theme was a sense of “understanding.” Students would write, “I could understand her.” Frequently students would write that there was never a time they didn’t understand. There was much negative evidence: “He didn’t mumble.” “He never spoke too softly.” Many merely stated, “I heard her/him.” Some spoke of instructors who “would repeat something if we didn’t understand.”

These reasons might be considering only clarity of articulation and audibility into account, or they may communicate the higher level of comprehension. It was unclear which was intended, but the latter meaning did show up in other evidence. For example, a significant number of responses mentioned the instructor’s vocabulary: “He used real technical terms.” He “has a large vocabulary.” He would stop and define “words we didn’t know.”

Some described the physical arrangement of the room: “We always sat in a circle so it was easy to hear.” “I moved in front to hear.” “[He talked] “to us, not to the board.” The students of one instructor generated a strong pattern of “he speaks in [a] monotone,” a comment made by no other students.

Item 12: “The subject matter was clearly presented by the instructor.”

In contrast to the narrow themes in item 11, the broadest range of meanings were evoked by item 12, which also created the most challenges in interpreting the students’ meaning. No clearly defined pattern emerged for the ten instructors, suggesting that each student had a unique meaning to ascribe to “clearly presented the subject matter.” We had expected two or three strong themes; instead there were many definite themes.

Students frequently used “explaining” and “understanding.” For example, students described “no difficulty understanding” a teacher. “He explained what we didn’t know.” “[He] explained everything clearly to us.” Such descriptions were similar in meaning, it seemed, to “understanding.” For example: “He occasionally had problems explaining so we could understand.” “We didn’t understand her all the time.”

A second pattern was "repetition." Students frequently described an instructor who would "repeat ideas" until they understood. "[He] went over things repeatedly." "[He] repeated the stuff in the textbook." "He rephrased things." "If necessary he went back and explained what we did before."

A third theme of meaning for this item was "orderliness," a concept that went beyond merely being organized. Students said: "We knew where we were." "We got sidetracked a lot." "It was hard to keep our place." "She jumped around." "He followed the syllabus." "He did not confuse us." "He always followed an outline." "It was organized for us." "It was always confusing." "There was a logical pattern to discussion." A few students described an instructor who "broke things down to make them simpler." Taken together, these meanings evoked a construct of orderliness.

A fourth theme was the instructor's "use of examples." Students wrote: "She used practical examples." "He used analogies." The most common phrase was simply: "He used examples." A fifth, closely related, theme was the use of personalized examples: "He used examples from student's lives." "He personalized the material." "She used students' language." "She related it all to our lives."

A sixth theme that emerged seemed to revolve around "questions," which students identified as an important tool for clarity. They responded: "I didn't have to ask questions." "He encouraged our questions." "Few questions for clarification were asked." "She was willing to answer our questions." There were comments about questioning in all ten classroom groups to explain this item.

Item 13: "The instructor put material across in an interesting way."

Related to this item, the strongest theme was what we called "story-telling." Instructors used "stories and personal experiences from real life," related "personal examples," and "used [their] own experiences and examples."

A second theme of somewhat lesser magnitude was the sense of boredom: "I was bored." "I was never bored." "The subject is so boring." In one instructor's class approximately one-third of the students used descriptions like these: "He's so energetic." "A dynamic fun person." "He always spoke with enthusiasm and energy." "Upbeat." "Animated."

Another theme emerging in several, but not all, the groups was diversity in teaching style. Instructors "used videos," "used handouts, videos, slides, lots of things, etc.," and "used a variety of things in class." Finally, sleep was frequently mentioned: "I never fell asleep" or "I couldn't stay awake."

Item 14: “Students were able to express themselves freely as a result of the instructor’s openness to their ideas.”

Students frequently used language such as “open to questions,” “would listen,” and “encouraged participation.” Other comments were: “[He] did not seem open to our ideas.” “His answers were intimidating.” “[He] would say ‘you’re wrong’ to students.” While many students showed little concern for the absence of debate, many others showed a strong concern for the lack of opportunity to voice opinions and make comments.

One common theme among the responses to this item was a description of what did not happen. Students were specific in identifying such instructor nonactions as: “She never condemned anyone.” “[He] never criticized stupid questions.” “I never felt I could express myself.” “He did not exclude anyone from discussion.” These students evaluated “openness” and “express” by personal definitions they ascribed to these terms in this high-inference item.

Interestingly, this item evoked responses that, to us, revealed a potential conflict between openness and credible teaching. Students said, for example: “Never was an opinion considered wrong.” “He let us say whatever we wanted.” “I always make obnoxious comments and feel I can say whatever I want.” These statements showed instructor openness to whatever the students offered. These students might rate a professor with high marks, consistent with the positive affect in these remarks. However, such absolute openness may reflect poor teaching because students are not always right. Here poor teachers get high marks.

In contrast, to this same item others stated their reasons for ratings as: “We had to always be exact.” “We had to go with [the instructor’s] opinion.” “He welcomed comments but expressed his view as the right view.” They expressed more negative feelings (and probably low ratings) about openness when there were some standards of right and wrong put forth by the instructor. But such standards are often the mark of good teaching. It is appropriate—essential, in fact—for the instructor to correct students’ erroneous thinking and poor reasoning. This is what teaching is. Here good teachers get low marks.

Item 14 has the further problem of being poorly constructed. It states: “Students were able to express themselves freely as a result of the instructor’s openness to their ideas.” The sentence presents two separate ideas and, as such, is flawed in construction. It creates a dual response set, since it is not possible to determine whether a high rating indicates students’ feeling of being able to express themselves or whether it indicates the instructor’s openness to ideas. The students’ response may rate the first idea, the second, or both. The wide variability of meanings here reflects this ambiguous response set. To what were we asking students to attend?

Item 15: "The instructor was willing to help students who experience difficulty in the course."

Students viewed "willing to help" from two perspectives, the first demonstrating an intention to help and the other demonstrating different types of help. Students often described relatively low levels of intent: "seems helpful," "said she would help," "seemed available," or, neutrally, "he said so in the syllabus." This breadth of judgement about the instructor's intention to help reflects the high inference of the item. Typical of the meager evidence offered by many students was: "She told us her office hours and phone."

Although the item inquired about the instructor's "willingness to help," many students responded with characteristics of the help itself. We classified the types of help students reported into three categories. The strongest pattern of helping behavior occurred in the classroom. Typically these responses did not report one-on-one help but rather included the familiar "he answered our questions," "gave out review sheets," or "gave extra credit." How these classroom techniques, which were applied to all students, particularly benefited the group of students who "experience[d] difficulty" is not clear. The second helping behavior was out-of-class availability. Students reported that the instructor was "always around for extra credit," "required individual interviews," or kept his "office always open." A third helping behavior was personal assistance from the instructor consisting of exceptions to standard operations: "I was sick and he allowed me to make up work." "He reevaluated my paper" or "adjusted my test score."

We discussed the fact that these latter responses were perhaps the closest, in our minds, to the item's intended meaning. The item as stated includes the conditional phrase, "students who experienced difficulty," clearly the case in these latter instances. However, this group of responses does not reflect the major pattern of students' interpretations. Part of the problem, we believe, is that the item assumes that all students either have difficulty in the course or know others who do.

Item 16: "The instructor respected students as persons."

As with several of the items, this statement requires students to infer the instructors' attitude from their own experiences. Themes we constructed from responses included these three: perceived equality, emotionality (overlapping with item 14), and communication behaviors. Students praised teachers who "valued our opinion, were "interested in our well-being," or "respected people." Among negative descriptors were: "Students' ideas [were] less important" than the instructors.' "[He was] rude to us."

"[He] liked to poke fun at us." This item seemed to tap into the emotional domain of student experience.

A number of students responded by describing what did not occur: "[He did] not laugh at us." She made "no disrespectful comments." He was "not biased." He displayed "no sexism." Such descriptions might indicate that students have well-defined standards of respect. They frequently used language similar to the answers for item 14, indicating redundancy in meaning and overlap in measurement.

Some students responded by describing a communication behavior. At times the behavior was descriptive: He "listens to students." She "engaged students in class." Other comments conveyed positive or negative affect: He "joked with us" and "gave good feedback," or he "snapped fingers at us" or "made hurtful comments."

Item 17: "The instructor was fair in grading examinations and assignments."

An important deduction from these responses concerns the specificity of student thinking. Examples of the pragmatic definitions used by students included: He "g[ave] partial credit." He "review[ed] for exams." "He threw out poor test questions." He "allowed me to redo a paper." One emergent pattern was the heavy (four to one) emphasis on exams over assignments. Students seemed most concerned with fairness in testing, a pattern that emerged in language about the level of difficulty of the exam, the vagueness or clarity of the questions, the length of the test, and the clarity of test instructions. Second, the students also identified fairness as an issue in testing procedures. They wrote: He "allowed us to argue [about the] fairness of test questions." He returned the exams within a "reasonable time," or "late." He tried to cover "too much material" in "one exam."

The third characterization of fairness concerned the instructor as grader. A majority of student responses were negative: "too tough," "stiff in grading," "too strict in grading," "never kind," and "picky." Also common was language such as "fair, but harsh," "grading scale unclear," and "could be more lenient." A minority of students perceived that the instructor as a fair grader. This group commented: He was "objective in grading." She "read essay [questions] and gave us the benefit of the doubt." Most frequent was the explanation that the instructor "allowed partial credit [for partially correct answers]."

A fourth theme was grade expectation. Many students mentioned the grade they expected. While correlation of meaning with the value of rating (high or low) was not our intent, a pattern did emerge showing that students satisfied with the grade they expected tended to rate the instructor high on this item while dissatisfied students tended to rate the instructor

lower. In addition, satisfied students used "I" while dissatisfied students more frequently used "he/she" to refer to the instructor. In other words, satisfied students' spoke from their own perspective ("I, me") while dissatisfied students described the instructor's. Comments from the first group included: "I got the grade I wanted." "I get good grades for doing good work." Comments from the second group included: "He never gives high grades." "He screwed me." In other words, students satisfied with their grades took credit for it; students dissatisfied with their grades blamed the instructor. Attribution theory suggests that people tend to see positive outcomes associated with their own dispositions and negative outcomes related to situational factors.

The last, and weakest, theme concerned the fairness of the types of test used. The consensus among students is that an objective test is fairer than a subjective test. Several students commented: "She used a scantron sheet, so answers were fair." He "had to be fair; answers were either right or wrong." "Is there a fair way of grading an essay?"

LIMITATIONS

The trustworthiness of our results rests on the quality of our design and its acknowledged limitations. To this end we aligned our research activities with Yvonna Lincoln and Egon Guba's (1985) dimensions of credibility, transferability, dependability, and confirmability as a check of design validity of qualitative research. First, the credibility of our results, we feel rests on the large number of student responses we incorporated. While not strictly parallel to Lincoln and Guba's standard of prolonged engagement, we feel that this large N-size is analogous to extended time in a setting. We feel confident that the initial coding of approximately 389 students' written responses on eight items for ten faculty members provided a strong and stable collection of meanings from which we could deduce students' thinking as they rated faculty at the end of each term. We also used negative case analysis: i.e., we continued to probe the responses to include all cases in our emerging interpretations. We did not use member checks or peer debriefing, at least insofar as a disinterested peer was concerned. We discussed the emerging findings only with each other.

As far as transferability, Lincoln and Guba's answer to the quantitative concern for generalizability, we suggest that our results have possible applications elsewhere. Transferability, of course, is up to others reading our results, and is not a claim we can make. However, to the extent that we found similar rating items and similar rating processes in the literature,

we feel that the underlying meanings suggested by our students may not be that different from undergraduates at other institutions.

The eighteen classrooms in which we gathered evidence were volunteered to us by ten faculty members. Obviously, it was not a random sample. As volunteers, these faculty may be more confident about their teaching and more effective than faculty as a whole. This was not a major concern because we were interested in student thinking as revealed in the language they used. Students' interpretation of the items can be revealed as they rate either effective or ineffective faculty. In other words, the meanings, or constructs that students convey may be stronger or weaker, present or absent, applauded or condemned, with good or bad teaching, but they are likely the same constructs.

The diversity of courses represented by these faculty strengthened the design, and our team effort in reading and studying the narrative increases the dependability of the results. We achieved a type of stepwise replication (Lincoln and Guba 1984) by sharing ideas as our interpretations emerged.

DISCUSSION

Validity: Questions and Answers

Qualitative research presents rich, full, informative details; to summarize it seems reductionistic and antithetical to our purposes. Nevertheless, the rich diversity of student interpretative patterns in the data suggest emerging questions for further study. For example, we found in an earlier study that ratings on "being interesting" (item 13) are significantly predictive of the "overall" rating (item 8) (Benz and Blatt 1994). Because item 8 is frequently used as a sole criterion for faculty evaluation, these results can enlighten faculty on a possible dynamic to which students are attending. If it is the case that story-telling is what it means to "be interesting," as we concluded in this study, then faculty may want to apply this understanding to their teaching, i.e., tell stories. Another finding here was that students think about orderliness and repetition when they're rating the clarity of presentation (item 12). This, too, can inform faculty about the underlying construct in students' minds.

Before applying the results of faculty ratings, understanding the students' interpretations of the items being used is crucial. This concept gets to the heart of validity—namely, are we measuring what we purport to be measuring? For example, "being prepared" to students means that the teacher is knowledgeable, according to our interpretation. In light of the fact that other studies have suggested that students may be ill-equipped to rate an instructor's mastery of the subject, this perception on the part of students is important. They may not, in fact, differentiate the appearance

of being prepared from expertise in subject matter. To faculty, this student perception could be a blessing or a curse. An extremely knowledgeable professor who is world-renowned in her discipline and is successful in enlightening student understanding about her subject matter may still be considered poorly prepared because she gets her overhead transparencies out of order while an intellectually weak and ill-informed professor may be rated as knowledgeable merely because his transparencies are in order. To what extent do our results help answer whether student ratings of faculty are valid?

Clearly, validity can be defined only purposefully. Measures may be valid for particular purposes but not for others. We found that the meanings students ascribe to the items on the rating form vary widely. For some items there were few strong themes; for other items many themes emerged. If nothing else, we concluded that validity is better established for the former items and less well established for the latter set of items. In other words, where there were few strong themes, we perceived that students were more in agreement about the meaning of the item; conversely, we interpreted a diversity of themes as an indication of less agreement among students about the item's meaning. Denis Phillips describes a similar notion of validity (consensual validity) based on Elliot Eisner's suggestions that qualitative research results in no truth, only "what a community believes" (paraphrased in Phillips 1987, 19). In addition, the level of inference of the item (whether high or low) did not seem related to agreement in meaning, i.e., validity.

For example, as already discussed, some students responded to item 10 (being well prepared) by rating their sense of the instructor's expertise in the subject while others were rating his or her punctuality and readiness with overheads. What are we measuring on a relatively low-inference item such as item 11, "spoke clearly and audibly," when some students responded to the professor's diction and others considered vocabulary? On item 12, which dealt with clarity of presentation, we found the least consensual validity. Students were rating all kinds of qualities. For example, some seemed to think about understanding; others thought about the sequencing of material; still others described the personalizing of the material. For item 13, on the other hand, we may claim the strongest validity—the prevalence of story-telling as its meaning above and beyond other lesser themes. Items 14, 15, 16, and 17 were relatively high-inference items as they addressed qualities of being "respectful," "fair," and "helpful." Student meaning enhanced the validity here by employing synonyms for these attitudinal qualities. On the other hand, surprisingly, some attitudinal items evoked stronger consensus on meaning than did the more behavioral ones.

We are left wondering about Murray's concern for "intractable generalities" that we cited earlier. How do we respond to his contention that teaching effectiveness is predictable from low-inference measures when even these seem to evoke meanings that lack consensus among students? Might not the process of student rating confound even low-item clarity? By this we mean (1) the pattern of making a quick positive judgment, then qualifying it with negative examples, (2) reporting what did not happen as well as what did, and (3) over-familiarity with the form leading to an attitude of "I've done this a zillion times, so let's get it over with."

In some instances, the form of the item, not its meaning, is a barrier to validity. Item 14, for example, contains two ideas; students' ability to express themselves freely and the instructor's openness to their ideas. It is impossible to know to which idea a student intends his or her response. Item 15 is similarly poorly constructed, worded in such a way that one who has not experienced difficulty in the course would not have an experiential base from which to respond. Both these items should be eliminated.

Numerical ratings as ultimate meaning are insufficient evidence of how students perceive teaching. Faculty rating systems should be supplemented by other evidence—evidence that gets beneath the numbers. Narrative comments, small discussions in the form of informal feedback sessions, and portfolios are other ways to add depth, richness, and meaning to faculty evaluation. Self-ratings by students—rating their own participation, their own contributions, and their own efforts—in addition to rating the faculty might offer balance to the process as well. Evaluation systems have taught students that nearly all agency in the classroom rests with the instructor.

In this study we examined student language to get a sense of how they interpreted the items; future researchers might want to analyze the meaning of the items correlated with the numerical value of the rating. We chose to focus only on interpreting the language students used as they gave reasons for their ratings.

Examining the meaning students give to these items adds information to the validity of using this faculty evaluation form—but what of the meaning to faculty? We have addressed that question in another study currently in progress. That the evaluation form acts as the intersection between student and faculty perceptions is clear; and to the extent that both sides of the process agree on the meaning of what is being measured, the process has validity. Going beyond the thousands of quantitative studies of student ratings of faculty (March and Bailey 1993), this study attempted to identify the underlying meaning of those ratings, examining the validity of an evaluative process that needs continual study, particularly

in an era of increased monitoring of teaching effectiveness on college campuses.

BIBLIOGRAPHY

- Benz, Carolyn R., and Stephen J. Blatt. "Faculty Effectiveness as Perceived by Both Students and Faculty: A Qualitative and Quantitative Study." Paper presented to American Educational Research Association, April 1994, New Orleans, Louisiana.
- . "Factors Underlying Effective College Teaching: What Students Tell Us." *Midwestern Educational Researcher* 8, no. 1 (Winter 1995): 27–31.
- Blatt, Stephen J., and Carolyn R. Benz. "The Relationship of Communication Competency to Perceived Teaching Effectiveness." Paper presented to the Central States Communication Association Conference, April 1993, Lexington, Kentucky.
- . "Faculty Teaching Evaluations: Behavioral Validation of Student Rating Methodology." Paper presented to the Southern States Communication Association Conference, April 1994, Norfolk, Virginia.
- Braskamp, Larry A., and John C. Ory. *Assessing Faculty Work*. San Francisco: Jossey-Bass, 1994.
- Centra, John A. *Reflective Faculty Evaluation*. San Francisco: Jossey-Bass, 1993.
- Donald, Janet G. *The State of Research on University Teaching Effectiveness: Using Research to Improve Teaching*. New Directions for Teaching and Learning, No. 23. San Francisco: Jossey-Bass, 1985.
- Feldman, Kenneth A. "Grades and College Students' Evaluation of Their Courses and Teachers." *Research in Higher Education* 4 (1976): 69–111.
- Glaser, Barney G., and Anselm L. Strauss. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine deGruyter, 1967.
- Goetz, Judith P., and Margaret D. LeCompte. *Ethnography and Qualitative Design in Educational Research*. San Diego: Academic Press, Inc., 1984.
- Lincoln, Yvonna, and Egon Guba. *Naturalistic Inquiry*. Beverly Hills, Calif.: Sage, 1985.
- Magner, Denise K. "Report to Focus on Standards for Assessing What Professors Do." *Chronicle of Higher Education* 40, no. 23 (9 February 1994): A22.
- Marsh, Herbert W. "Applicability Paradigm: Students' Evaluations of Teaching Effectiveness in Different Countries." *Journal of Educational Psychology* 78, no. 6 (1986): 465–73.
- Marsh, Herbert W., and Michael Bailey. "Multidimensional Students' Evaluations of Teaching Effectiveness: A Profile Analysis." *Journal of Higher Education* 64, no. 1 (January–February 1993): 1–18.
- Miles, Matthew B., and A. Michael Huberman. *Qualitative Data Analysis: A Sourcebook of New Methods*. Newbury Park, Calif.: Sage, 1984.
- Murray, Harry G. "Classroom Teaching Behaviors Related to College Teaching Effectiveness." In *Using Research to Improve Teaching Effectiveness*, edited by Janet G. Donald and Arthur M. Sullivan, 21–35. New Directions for Teaching and Learning, No. 23. San Francisco: Jossey-Bass, 1985.

- Newman, Isadore, and Carole Newman. *Conceptual Statistics for Beginners*, 2d ed. Akron, Ohio: University of Akron, 1992.
- Phillips, Denis C. "Validity in Qualitative Research." *Education and Urban Society* 20, no. 1 (1987): 9-24.
- Popham, W. James. *Modern Educational Measurement: A Practitioner's Perspective*. Englewood Cliffs, N.J.: Prentice-Hall, 1990.
- Rummel, Rudolph J. *Applied Factor Analysis*. Evanston: Northwestern Illinois Press, 1970.
- Ryan, Joseph M., P. D. Harrison, and Yi-Mei Zia. "The Relationship between Individual Instructional Characteristics and the Global Assessment of Teaching Effectiveness across Different Instructional Contexts." Paper presented at American Educational Research Association, April 1993, Atlanta.
- Seldin, Peter. *Successful Faculty Evaluation Programs*. New York: Coventry Press, 1980.
- . *Evaluating College Teaching*. *New Directions for Teaching and Learning*, 33 (1988): 47-56.
- Tiberius, Richard G., H. David Sackin, Joyce M. Slingerland, Kaela Jubas, Mary Bell, and Ann Matlow. "The Influence of Student Evaluative Feedback on the Improvement of Clinical Teaching." *Journal of Higher Education* 60, no. 6 (November/December 1989): 665-81.
- Wilson, Robert C., Jerry G. Gaff, Evelyn R. Dienst, Lynn Wood, and James L. Bavry. *College Professors and Their Impact upon Students*. New York: Wiley, 1975.
- Wolcott, Harry F. "On Seeking—and Rejecting—Validity in Qualitative Research." In *Qualitative Inquiry in Education*, edited by Elliot W. Eisner and Alan Peshkin, 121-52. New York: Teachers College Press, 1990.