International Journal of Nursing Education Scholarship

Volume 1, *Issue* 1 2004 *Article* 10

Beyond Student Ratings: Peer Observation of Classroom and Clinical Teaching

Ronald A. Berk* Phyllis L. Naumann†

Susan E. Appling‡

*Johns Hopkins University, rberk@son.jhmi.edu

†Johns Hopkins University, pnaumann@son.jhmi.edu

‡Johns Hopkins University, sappling@son.jhmi.edu

Beyond Student Ratings: Peer Observation of Classroom and Clinical Teaching

Ronald A. Berk, Phyllis L. Naumann, and Susan E. Appling

Abstract

Peer observation of classroom and clinical teaching has received increased attention over the past decade in schools of nursing to augment student ratings of teaching effectiveness. One essential ingredient is the scale used to evaluate performance. A five-step systematic procedure for adapting, writing, and building any peer observation scale is described. The differences between the development of a classroom observation scale and an appraisal scale to observe clinical instructors are examined. Psychometric issues peculiar to observation scales are discussed in terms of content validity, eight types of response bias, and interobserver reliability. The applications of the scales in one school of nursing as part of the triangulation of methods with student ratings and the teaching portfolio are illustrated. Copies of the scales are also provided.

KEYWORDS: peer review, peer observation, faculty evaluation, clinical teaching, student ratings

The mere mention of *faculty evaluation* can be very disturbing and threatening to many professors in schools of nursing. There has been mounting evidence of faculty hostility and cynicism toward student ratings (Nasser & Fresko, 2002; Schmelkin-Pedhazur, Spencer, & Gellman, 1997). Despite this negative image, a large percentage of faculty in all disciplines exhibit moderately positive attitudes toward the validity of student ratings and their usefulness for improving instruction; however, there is no consensus (Nasser & Fresko, 2002). Although there is still a wide range of opinion on their value, student ratings has emerged as the dominant method for evaluating teaching over the past 30 years in the United States (Seldin, 1999) and has become the most influential measure of performance used in promotion and tenure decisions at institutions that emphasize teaching effectiveness (Emery, Kramer, & Tian, 2003). McKeachie (1997) noted that "student ratings are the single most valid source of data on teaching effectiveness" (p. 1219). In fact, there is little evidence of the validity of any other sources of data (Marsh & Roche, 1997).

Considering all of the polemics over the merits of student ratings, they still provide only *one source* of information on teaching effectiveness. There are several other strategies that can be used to augment the heavy reliance on student ratings: peer review, teaching portfolios, self-evaluation, administrator evaluation, teaching scholarship, and student outcomes. Among these alternatives, peer review is becoming more prominent, such that more than 40% of liberal arts colleges use peer observation for summative evaluation (Seldin, 1999), compared to 88% that use student ratings. Schools of nursing are exploring different forms of peer review in the evaluation of both classroom and clinical teaching (Appling, Naumann, & Berk, 2001; Costello, Pateman, Pusey, & Longshaw, 2001; Ludwick, Dieckman, Herdtner, Dugan, & Roche, 1998; Martsolf et al., 1999).

PEER REVIEW

Rationale

In the early 1990s, Boyer (1990) and Rice (1991) redefined *scholarship* to include teaching. After all, it is the means by which discovered, integrated, and applied knowledge is transmitted to the next generation of scholars. Teaching *is* a scholarly activity. In order to prepare and teach a course, faculty must complete the following:

- Conduct a comprehensive up-to-date review of the literature.
- Develop content outlines.

- Prepare a syllabus.
- Choose the most appropriate print and nonprint resources.
- Write and/or select handouts.
- Integrate instructional technology (IT) support (e.g., audiovisuals, Web site).
- Design learning activities.
- Construct and grade evaluation measures.

Webb and McEnerney (1995) argued that these products and activities can be as creative and scholarly as original research.

If teaching performance is to be recognized and rewarded as scholarship, it should be subjected to the same rigorous peer review process to which a research manuscript is subjected prior to being published in a refereed journal. In other words, teaching should be judged by the same high standards applied to other forms of scholarship: *peer review.* Shoffner, Davis, and Bowen (1994) indicated that this perspective on teaching is of special consideration in nursing because of the discipline's focus on the application of nursing theory and the dissemination of knowledge to clinical practice. This type of scholarship can be coupled with the requirement of original research as significant evidence of scholarship.

The American Association of Higher Education's (AAHE) project, "From Idea to Prototype: The Peer Review of Teaching" (Hutchings, 1995), involved 12 universities, including the Kent State University College of Nursing. This project and the increased attention on the value of peer review by higher education leaders, such as Shulman (2004) and Palmer (1998), have moved peer review to the forefront.

Two Components

Peer review of teaching is composed of two activities: peer observation of in-class teaching performance and peer review of the written documents used in a course. *Peer observation* requires a rating scale that covers those aspects of teaching that peers are better qualified to evaluate than students. The scale items typically address the instructor's content knowledge, delivery, teaching methods, learning activities, and the like. The ratings may be recorded live with one or more peers on one or multiple occasions or from videotaped classes. *Peer review of teaching materials* requires a different type of scale to rate the quality of the course syllabus, instructional plans, texts, reading assignments, handouts, homework, and tests/projects. This review is less subjective and more cost-effective, efficient, and reliable than peer observations. However, the

DOI: 10.2202/1548-923X.1024

observations are the more common choice because they provide direct evaluations of the act of teaching. Both forms of peer review should be included in a comprehensive system, where possible.

Faculty Resistance

Despite the current state of the art of peer review (PR), there is considerable resistance by faculty to accept it as a complement to student ratings. Relative unpopularity of peer review stems from the following concerns:

1. It is biased because the ratings are personal and subjective (PR of research is blind and subjective).
2. One observer is unfair (PR of research usually has two or three reviewers).
3. In-class observations take too much time (PR of research can be time-consuming, but the time is distributed at the discretion of the reviewers).
4. One or two class observations do not constitute a representative sample of teaching performance for an entire course.
5. The results probably will not make any difference in teaching.
6. Only students can really evaluate what an instructor does for an entire course.
7. Available rating scales do not measure important characteristics of teaching effectiveness.
8. Teaching is not valued as much as research, especially at large research-oriented universities.
9. Observation data are inappropriate for summative evaluation (e.g., merit pay, promotion, or tenure decisions) by administrators.

Most of these reasons or perceptions are legitimate based on how different institutions execute a peer review system. A few can be corrected to minimize bias and unfairness and improve representativeness of observations.

However, there is consensus by experts on the preceding concern 9: Peer observation data should be used for formative evaluation to improve teaching rather than for summative evaluation on which personnel decisions are based (Aleamoni, 1982; Arreola, 2000; Centra, 1999; Cohen & McKeachie, 1980; Millis & Kaplan, 1995). In fact, 60 years of experience with peer assessment in the military and private industry led to the same conclusion (Muchinsky, 1995). Employees tend to accept peer observations when the results are used for constructive diagnostic feedback instead of as the basis for administrative decisions (Cederblom & Lounsbury, 1980; Love, 1981).

What's Needed?

The faculty in our school of nursing were interested in testing a peer review system. However, the greatest initial deterrent to building this system and convincing our faculty was finding an appropriate rating scale for not only traditional classroom observations, but also for observations of clinical instructors. Since student rating scales of course coordinators and clinical instructors were already being administered (Appling et al., 2001), parallel peer observation scales for both classroom and clinical instructors were essential in order to complement the student data. Faculty evaluation experts recommend the old "adopt or adapt" strategy to produce a peer rating scale; that is, collect a range of forms and either adopt one of those or adapt one to fit your faculty and courses. Unfortunately, a comprehensive review of the peer review literature reveals there is not a single book, chapter, or article on peer review that explains how to "adapt" a scale, write items, or correctly build a rating scale for peer observation. Further, there is no structured scale available to rate the performance of clinical instructors.

The remainder of this article presents the step-by-step procedure undertaken by our faculty to develop two peer observation scales to evaluate classroom and clinical teaching. This is followed by the applications of these scales to the school of nursing courses, which describe the actual execution of this observational system. Copies of the scales are appended.

SCALE DEVELOPMENT

Developing a scale for peer observation is a nontrivial task. First, the most frequently cited volumes on faculty evaluation do not describe how to develop rating scales (Arreola, 2000; Braskamp & Ory, 1994; Centra, 1993; Chism, 1999; Seldin & Associates, 1999). Some provide item banks for student rating scales, but none for peer review. Second, the next logical source would be generic books on rating scale development in the measurement literature. Recent ones (De Vellis, 2003; Netemeyer, Bearden, & Sherman, 2003) as well as popular works within the past millennium describe about a half a dozen rules, or only list rules, for writing statements for Likert-type scales (Nunnally & Bernstein, 1994; Streiner & Norman, 1995). Even the classic sources on scaling just list rules for writing items or give brief descriptions (Likert, 1932; Thurstone & Chave, 1929; Wang, 1932). Finally, the most comprehensive books on writing questions for questionnaires by Payne (1951) and Sudman and Bradburn (1982) cover numerous rules for developing items for survey instruments, including a few

DOI: 10.2202/1548-923X.1024

rating scale item formats, but do not address scaling issues, various types of anchors, or any application to higher education, much less peer observation.

For more than 75 years during which these books were published, and since the first research study on student ratings was published (Remmer & Brandenburg, 1927), it would seem that explicit instructions for item construction would have appeared. Unfortunately, there are few or no explanations of the rules presented, few examples of good and bad items, no lists of the types of anchors that could be used, and no procedures for selecting and matching anchors to the items.

Drawing from all of the sources on scale construction cited previously, a five-step process was adopted: (1) domain specification, (2) item generation, (3) scale structure, (4) faculty review and field-testing, and (5) validity and reliability. Each step is described next.

Domain Specification

The best road map to item generation is to create a comprehensive list of teaching behaviors, skills, and characteristics that define the domain of teaching effectiveness. These specifications can guide the development of the words, phrases, or statements that comprise the heart of the rating scale. In general, the content of a peer observation scale should complement that already measured by the student rating scale. What characteristics would faculty be in a position to rate which a student could not? The domain should be different for both types of scales, although there may be some behaviors where overlap is appropriate to confirm or disaffirm student and faculty observations.

Two techniques were employed to define the behaviors: (1) research evidence and (2) focus groups. First, the results of several studies indicate that faculty may be the best judge of content expertise, pedagogy, and related dimensions (DeZure, 1999). Cohen and McKeachie (1980) identified 10 criteria of effective teaching that colleagues are in the best position to evaluate: mastery of course content, selection of course content, course organization, appropriateness of course objectives, instructional materials, evaluative devices and methods used to teach specific content areas, commitment to teaching and concern for student learning, student achievement, and support for departmental instructional efforts. Keig and Waggoner (1994) isolated five categories: goals, content, and organization of course design; methods and materials used in delivery; evaluation of student work; instructor's grading practices; and instructor's adherence to ethical standards.

DeZure's (1999) synthesis of the research produced six major dimensions of teaching behaviors that can be observed: class environment; indicators of student involvement and engagement; instructor's ability to convey the course content; instructional methods; indicators of student-instructor rapport; and global rating of overall effectiveness. These dimensions are consistent with the research on instructional effectiveness (presentation style, enthusiasm, sensitivity to students' levels of knowledge, openness to questions, and clarity of organization) (Murray, 1997). Feldman's (1989) study of faculty's highest ratings included: knowledge of subject; enthusiasm; sensitivity to class level of knowledge and progress; preparation and organization of class; and clarity.

The aforementioned categories of behavior furnished a preliminary structure within which to group a wide range of teaching behaviors. These behaviors were drawn from the research on peer review (Hutchings, 1995; Morehead & Shedd, 1997; Muchinsky, 1995; Webb & McEnerney, 1995). These behaviors were used to develop the items for the classroom observation scale.

During this process of domain specification, it became clear that there were several types of behavior that could not be directly observed in the classroom, but were no less important: fairness, grading practices, ethics, and professionalism (Braxton, Bayer, & Finkelstein, 1992). These behaviors would have to be assessed by a peer review of course materials, graded work, and peer references. This review was beyond the scope of the current project.

Once all potential categories or dimensions of teacher behaviors and the behaviors themselves were identified in the literature, a focus group of faculty volunteers, representing undergraduate and graduate nursing programs, was assembled to critically review those lists. After several sessions, a draft list of categories and behaviors was compiled. The next iteration of meetings involved brainstorming the specific teaching behaviors that should be covered on the scale. This series of meetings produced a comprehensive inventory of behaviors under the most important dimensions.

Item Generation

Historically, faculty rating scales have mostly been developed using the approach of *dust bowl empiricism,* which means: get a bunch of items about teaching and see what works (McKeachie, 1997). However, instead of rewriting every item and reinventing the peer observation scale wheel, our first strategy was to review existing scales published in the literature and reprinted in books on

DOI: 10.2202/1548-923X.1024

faculty evaluation (Arreola, 2000; Braskamp & Ory; 1994; Centra, 1993; Chism, 1999). The items on these scales provided prototypes for both content and structure so they could be tailored to our specific nursing faculty and courses. Despite the diversity of scales and items currently being used (e.g., Willis & Kaplan, 1995; Richlin & Manning, 1995), the available tools served as an excellent springboard for building our own class observation scale. In total, our faculty committee reviewed 17 scales and 535 items. Unfortunately, there were no prototypes for structured clinical observation scales.

The second step toward item generation was to determine the item format. Most scales contain two types of items: structured and unstructured.

Structured items. Virtually all structured items consist of a *stimulus*, usually a word, phrase, or declarative sentence, and a set of *response options*, which are a series of descriptors indicating one's strength of feeling toward the stimulus.

Unlike student rating scales which usually use a Likert-type format with a set of declarative sentences and response options of one or two words expressing varying degrees of agreement and disagreement, peer observation scales usually list teaching behaviors as single words or short phrases which can be checked quickly while they are being observed. The response options ask a faculty peer to evaluate the quality of each behavior according to a particular dimension, such as excellent–poor, effective–ineffective, and satisfactory–unsatisfactory. These represent unipolar scales, ranging from the highest level of quality to the lowest. The options or anchors provide a graduated scale of levels of quality.

The format for the structured items consisted of words or phrases and five anchors: "Excellent," "Very Good," "Good," "Needs Improvement," and "Not Applicable". The last-named anchor was added to capture all behaviors that may not be observed during a single observation period.

A draft pool of 75 words/phases was developed by adapting items from available scales and writing others to match all of the behaviors in the domain specifications. There was at least one item per behavior to assure coverage of the domain. The domain categories into which all items were sorted included: content and organization, communications skills, questioning skills, critical thinking skills, rapport with students, learning environment, and teaching methods. The majority of the items were much more specific than the statements on the student rating scale.

Unstructured items. These items permit the faculty peer to provide answers in his or her own words. These answers are designed to supply information not tapped by the responses to the structured items. Overall, the structured and unstructured items should be complementary in content and format.

The unstructured items can be very broad or specific. Each item can be one or two words, a phrase, or imperative or interrogative sentence. The committee's decision on our peer observation scale was to add a "Comments:" column after the anchors on the structured section of the scale so specific comments could be noted quickly as each item was being rated, and two unstructured items at the end of the scale: "Strengths:" and "Areas for Improvement:".

Faculty Review and Field Testing

Once a draft of the instrument was completed, it was distributed to the entire faculty and the Appointments, Promotion, and Tenure (APT) Committee for review. Their content and format feedback was incorporated into the next revision. This time the length was also shortened to 61 items.

The next draft was sent back to faculty for field testing. Guidelines for using the checklist were provided. Over the past year, several faculty tested the scale. Confidentiality was maintained throughout the entire process.

Validity and Reliability

The preceding steps in scale construction as well as the collection of validity and reliability evidence related to the specific score uses and inferences are required by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME Joint Committee on Standards, 1999.) The salient evidence for the formative evaluation uses of this peer rating scale was examined next. In particular, the response biases and limitations on gathering reliability data are discussed.

Content validity. The most important validity evidence relates to the content of the scale. The items on the rating scale must be congruent with the domain specification structure of major dimensions and teaching behaviors described previously. A formal judgmental review by the focus group of five faculty members evaluated the representativeness of the sample of teacher behaviors, their relevance to nursing faculty and courses, and the comprehensiveness of coverage of the *a priori* dimensions. Revisions were

DOI: 10.2202/1548-923X.1024

completed as necessary for the seven dimensions and 61 items until unanimity was attained. Since the domain of teaching behaviors became the actual items on the rating scale, with only minor changes, the congruence between the behaviors and items was virtually built into the structure.

Response bias. When peers rate an instructor's teaching performance, it is assumed that they will read each behavior carefully and make their honest rating with scrupulous impartiality. Unfortunately, this assumption is not always justified. The problem is that there are human tendencies or factors that may contaminate their responses, rendering them less than honest and impartial. These tendencies are known as *response sets* or *biases*. The most common biases in peer observation scales are halo effect, end-aversion bias, extreme-response bias, acquiescence or yea-saying bias, and gender, racial/ethnic, and sexual-orientation bias. There are also three other biases that particularly afflict peer ratings: incompetence bias, buddy bias, and back-scratching bias. All eight biases are described briefly below:

1. Halo effect: This is the extent to which a peer's overall impression of an instructor will affect his or her rating. For example, if the global impression is positive, the faculty member may simply mark "Excellent" or "Very Good" to every item. Thorndike (1920), who coined the name for this effect, defined it as follows: "The judge seems intent on reporting his [or her] final opinion of the strength, weakness, merit, or demerit of the personality as a whole, rather than on giving as discriminating a rating as possible for each separate characteristic" (p. 447).

2. End-aversion bias: This bias refers to the tendency of faculty to ignore the extreme anchors on the scale. They may be viewed as too strong. Instead, they choose the middle ratings of the scale, which restricts the range of responses.

3. Extreme-response bias: This tendency is the opposite of the one above. In this case, peers mark the extreme anchors rather than those in between. This bias is difficult to detect because the reason for the choice of the extremes may also be due to their honest ratings or the halo effect.

4. Acquiescence/yea-saying bias: This bias is the tendency to give positive responses to items irrespective of their content (Couch & Keniston, 1960). In our culture, we are socialized to be agreeable, to say "yes" instead of "no," and when asked, "How are you?", we answer "fine," whether we honestly mean it or not. Faculty may select "Excellent," "Very Good," and "Good" more often than negative anchors or "Needs Improvement" on the observation checklist. In fact,

peer ratings tend to be very generous (Root, 1987). This response set tends to inflate the ratings and skew the distribution toward the upper end of the scale so that an instructor's performance appears much better than it really is.

5. Gender, racial/ethnic, and sexual-orientation bias: In the interest of fairness and nondiscrimination, this type of bias must be addressed. Since such bias exists in the forms of salary inequity, differential hiring and promotion rates, and available benefits/privileges at the different ranks, cross-gender, cross-race, or straight-gay peer ratings can exhibit conscious or unconscious bias. Although this type of bias has not been formally assessed in peer observations in colleges and universities, industrial research has indicated that same-race peer evaluations were more positive than cross-race evaluations (Schmitt & Lappin, 1980). The most serious implications of this bias are in uses of peer observation data for administrative decisions.

Given the format of peer observation scales, it is very difficult to detect these biases in response patterns. Although they may decrease the validity of the ratings, there is little one can do in structuring the scale to minimize those sources of bias. In fact, there are three other sources, formally coined here, that can be added to the list:

6. Incompetence bias: This is the tendency to assign high ratings because of a lack of competence and/or confidence in rating teaching behaviors (Root, 1987). A peer may have limited or no knowledge of teaching methods or simply lack experience. When peer observers are incompetent on the characteristics being rated, they tend to give more positive ratings, rather than penalize the faculty member for his or her own shortcomings.

7. Buddy bias: Friendship and degree of acquaintance can inflate peer ratings. Early studies of peer assessment in the military (Wherry & Fryer, 1949) and private industry Freeberg, 1969; Hollander, 1956; Love, 1981) suggest this type of bias may be just as applicable to academia. This bias can be eliminated if the peer rater is chosen by someone other than the peer's "buddy," such as an administrator.

8. Back-scratching bias: This bias occurs when a faculty member gives high ratings to peers on the exchange assumption that he or she will then receive high ratings, kinds of a "mutual admiration society" mentality. This mutual back scratching is most common when faculty select their own peer evaluators. If these observers are selected by an administrator, such as a department chair,

DOI: 10.2202/1548-923X.1024

associate dean, or dean, and they are trained in teaching observation, back-scratching bias can be minimized and even eliminated.

These three sources of bias were not directly addressed with formal observer training and selection of peer observers by our program directors or associate dean. This level of commitment was not evidenced by the faculty and administrators. As an adjunct to the student rating scale and teaching portfolio, the peer observations were to be used for formative decisions only; that is, for teaching improvement. Therefore, confidentiality of the process, and the resistance of faculty toward the intrusion on their teaching by anyone other than a colleague of their choosing, precluded the elimination of incompetence and back-scratching biases.

Overall, seven of the eight potential types of bias (excluding gender, racial/ethnic, and sexual-orientation bias) described above may have weakened the validity of the ratings by inflating the diagnostic profile to some extent. Hopefully, the most serious teaching deficiencies were still identified in the structured or unstructured sections of the scale.

Reliability. The research on peer classroom observation indicates that interobserver reliability, when it is feasible to estimate, tends to be weak (Centra, 1993). Even when a common set of behaviors is assessed using standardized procedures with at least two faculty observers rating the same instructors, interobserver reliability of the scores is low. Although it is preferable to have more than one observer because of personal response biases or idiosyncrasies, which are difficult to measure, two different faculty bring different expertise to their observations. Even with formal observer training, there are many uncontrolled factors that contribute to the inconsistency in peer ratings.

When the feedback is diagnostic, the instructor can either accept or reject the evaluation to improve his or her teaching. If the peer ratings are used summatively for tenure, promotion, reappointment, or salary decisions, then low reliability can have an adverse impact on the instructor's career. What renders this reliability issue as an intractable problem is the logistics of even obtaining multiple, standardized observations on the same instructor. The sheer practicality of executing an interobserver reliability study of peer ratings with appropriate controls may be prohibitive.

Given the infancy of our peer observation system and the faculty professional milieu within which all faculty teaching evaluations take place, the formative diagnostic purpose of the peer reviews and confidentiality of results

precluded the estimation of interobserver reliability. This issue, however, will be re-examined over the next year.

APPLICATIONS OF SCALES

Classroom Observation

A draft version of the scale was assembled. Preceding the structured and unstructured items were identification information (instructor, observer, class topic, etc.), a statement of purpose, and directions for rating the items. The structured items were grouped into subscales based on the aforementioned seven domain categories. The unstructured, open-response items followed to complete the scale. A copy of the scale is presented in Appendix A.

Administration. Since the scale was an observational checklist, it could be used on one or several occasions by the same peer or different peers. Faculty could request a peer to rate the content and organization subscale only, teaching methods subscale only, any other subscale, or the total scale. That decision would depend on the expertise of the faculty observer and the diagnostic information needed by the instructor.

Scoring. The scale was designed to be used as a diagnostic tool. The responses, which are checkmarks and written comments, provided an item-by-item profile of strengths and weaknesses (Needs Improvement). No scores were computed for total scale and the subscales because the items rated by one peer may be different from those rated by a different peer. Further, total scores do not furnish diagnostic information.

Practice. Several faculty at both the undergraduate and graduate levels have requested peer observations of specific sections of the scale. In practice, this flexibility in completing the scale has been effective. Peers with content expertise may only rate relevant items on the scale whereas others who have knowledge of a variety of teaching techniques may rate the methods section only. This assured validity of the ratings by those peers who possess competence in the areas rated. Although not every peer completes the total scale for every instructor, those sections that are rated do not produce inflated scores due to “incompetence bias.”

Confidentiality has been maintained throughout all of the peer observations, although faculty have been encouraged to report areas of improvement and course changes from their peer reviews to the curriculum

DOI: 10.2202/1548-923X.1024

committees when the courses are evaluated. The approach permits faculty to voluntarily integrate the feedback from peers along with the student evaluations for course and teaching improvement. The actual peer observation results and the name of the observer are never revealed.

Clinical Observation

A sound and effective peer evaluation process is important to assure that the highest educational standards are applied within and across clinical courses. Peer observation of clinical faculty can serve several functions. Ludwick et al. (1998) indicated that it can provide opportunities for reflection and introspection on each instructor's clinical teaching skills. When areas of weakness are identified, faculty development strategies can be recommended. Conversely, when areas of strength are identified, they can be communicated to other instructors to provide true peer-to-peer mentoring opportunities. Peer observation also decreases the sense of isolation felt by clinical faculty and can create a sense of community as effective teaching methods are explored and shared. Students also benefit from observing the peer review process and discussing its role in professional development.

Once the classroom observation scales were in operation, undergraduate course coordinators requested a scale to evaluate their clinical instructors' performance. This time a committee was convened of undergraduate faculty only to execute the same procedure used to develop the previous scale. The first question was: What is different about this scale compared to the one we already constructed? Answer: Just about everything.

Key differences. The peer in this case was the course coordinator and the faculty being rated were clinical instructors. The behaviors to be rated went beyond those in the standard classroom scale to include a range of course management outcomes, such as: follows clinical course guidelines, attends clinical course-related meetings, and develops a written, systematic plan for student improvement, as appropriate.

Scale structure. The aforementioned five-step scale construction procedure was followed. The product was a two subscale structure: *teaching methods* and *course management.* The first subscale contained 18 items drawn primarily from sections of the classroom observation scale; the second subscale consisted of 9 items assessing adherence to course behaviors. The 27 structured items used a different set of response anchors from the other scale. The peer coordinator checked "Agree," "Disagree," " Not Applicable," or "Not Observed."

The dichotomous response format of agree-disagree streamlined the ease and time to complete the ratings, while not compromising the information required. A "Comments:" column after the anchors appeared after each item to note specific comments. Two unstructured items at the end of the scale were "Strengths:" and "Areas for Improvement:". A copy of this scale is shown in Appendix B.

Practice. Several course coordinators used this scale to evaluate their clinical instructors. The most noteworthy implementation issue was the time involved to conduct the observations. The time varied significantly as a function of site – school-based practice labs or community-based clinical sites. Observations of clinical instructors in the *practice labs* were completed easily with very little additional time, beyond lab hours. Observations at a variety of *clinical sites,* including long-term care facilities and hospitals throughout the area, required considerable time. One coordinator traveled to 30 sites and spent approximately one hour at each site. Another coordinator observed 12 instructors within one hospital site, but made multiple one-hour visits to five instructors who were new. During these visits, teaching methods and interactions with both staff and students were observed. They were encouraged to provide feedback to the coordinator regarding instructor effectiveness.

Following each visit, the coordinator completed the scale and documented strengths and areas for improvement. He or she later shared the ratings and comments along with the student rating results in a one-on-one conference with each instructor. Specific suggestions were given to improve clinical teaching.

Anecdotal reactions to the scale and to the entire process were quite positive. The coordinators recommended that, prior to visiting the sites, all clinical instructors be sent copies of the scale to provide them with clear, structured behavioral expectations. The clinical instructors, especially those who were new teachers, found the scale informative in clarifying their role. Coordinators found the scale easy to use, although a few requested additional space for item comments. Both students and clinical instructors reported the process to be very helpful. Students recognized the peer observation/evaluation process as an effort to monitor and improve the quality of clinical instruction and, ultimately, their educational experience. The clinical faculty considered the feedback and guidance essential to their growth and confidence as clinical educators.

The peer observation of clinical instructors requires administrative support and adequate time to implement. Visiting each clinical faculty member, spending time observing teaching, providing insightful feedback, and, finally, sharing

DOI: 10.2202/1548-923X.1024

strategies to address areas for improvement are all time intensive, but necessary to improve clinical education.

FUTURE DIRECTIONS

Over the past two years, faculty who have requested peer observations have been satisfied with the process and usefulness of the information provided by the scale. Similarly, course coordinators have not suggested any substantive changes in the scale designed to rate clinical instructors, although the need for additional space for item comments was noted.

Once the scales were e-mailed to all faculty, implementation was low key. It is anticipated that an increase in administrative support will produce a spike in the level of faculty participation. The faculty committee involved in the project from its inception would also like to expand the peer observation process to include the following:

1. Formal peer training, beyond the instructions provided with the scales
2. Assembling peer-observer teams of two or three faculty to estimate interobserver reliability
3. Scheduling multiple observer visits to increase class sampling per instructor
4. Preparing guidelines for a post-observation conference to maximize its value to the instructor
5. Developing a scale and procedure for peer review of course materials

Furthermore, feedback from classroom and clinical instructors on changes in their teaching as a result of the observation data would be useful. That information will be collected over the next year.

Peer observation has been accepted as a legitimate mechanism for providing meaningful data for teaching improvement. It serves as a valuable addition rather than alternative to student ratings and the teaching portfolio. The triangulation of all three methods to compensate for the inadequacies in each method furnishes a stronger foundation of evidence from which teaching scholarship can be evaluated. Eventually, this evidence may receive the long overdue recognition it deserves alongside research scholarship.

APPENDIX A

JOHNS HOPKINS UNIVERSITY SCHOOL OF NURSING PEER OBSERVATION SCALE

Instructor: _____
Observer:_____
Course/Room No.:_____
Class Topic:_____
Date:_____

PURPOSE: This scale is designed as an observation tool to rate an individual instructor's teaching performance. It is intended to provide a diagnostic profile for teaching improvement.

DIRECTIONS: Using the anchors below, check (✔) your rating for each teaching behavior that's applicable for the specific class observed. Check "NA" for items that do not apply.

E	**=**	**Excellent**
VG	**=**	**Very Good**
G	**=**	**Good**
NI	**=**	**Needs Improvement**
NA	**=**	**Not Applicable**

DOI: 10.2202/1548-923X.1024

	E	VG	G	NI	NA	Comments:
CONTENT AND ORGANIZATION						
Started and ended class on time						
Presented overview of class content/objectives						
Presented rationale for topics covered						
Presented key concepts						
Presented current material						
Presented information in an organized manner						
Demonstrated accurate knowledge of content						
Used relevant examples to explain major ideas						
Used alternative explanations when necessary						
Made efficient use of class time						
Covered class content / objectives						
COMMUNICATION SKILLS						
Varied pace appropriately						
Enunciated clearly						
Varied modulation						
Varied tone						
Spoke with adequate volume						
Demonstrated confidence						
Demonstrated enthusiasm						
Moved easily about room during presentation						
Used speech fillers (um, ok, ah) rarely						
Established and maintained eye contact						
Maintained students' attention						

	E	VG	G	NI	NA	Comments:
QUESTIONING SKILLS						
Encouraged students' questions						
Listened carefully to students' questions						
Answered questions appropriately						
Restated students' questions or comments as necessary						
CRITICAL THINKING SKILLS						
Asked probing questions						
Used case studies or scenarios						
Used small-group discussion						
Encouraged students to answer difficult questions by providing cues or rephrasing						
RAPPORT WITH STUDENTS						
Greeted students at the beginning of class						
Responded appropriately to students' puzzlement or boredom						
Asked students to clarify questions, when necessary						
Requested very difficult, time-consuming, or irrelevant questions be addressed at a later time						
Used humor and/or anecdotes appropriately						
Demonstrated respect for students and their thoughts /concerns						
LEARNING ENVIRONMENT						
Physical characteristics (temperature, lighting, crowding, seating)						
Class Affect						
Conducive to learning						
Relaxed						
Controlled						

DOI: 10.2202/1548-923X.1024

	E	VG	G	NI	NA	Comments:
TEACHING METHODS						
Lecture						
Engagement Techniques						
Q&A						
Discussion						
Small-group activities						
Student individual/panel presentations						
Active learning (e.g., think-pair-share)						
One-minute paper						
Other _____						
Role playing						
Demonstrations/skits						
Simulations						
Games						
Use or integration of technology						
Overheads						
PowerPoint						
Slides						
PC						
CD/ROM						
Course Web site						
Internet						
Videos						
Audiotapes						
Other_____						
Experimental/Innovative techniques						
Specify_____						
Other						

Strengths:

Areas for Improvement:

Observer Signature _____

APPENDIX B

JOHNS HOPKINS UNIVERSITY SCHOOL OF NURSING CLINICAL FACULTY PERFORMANCE APPRAISAL FORM

Course Coordinator:_____

Clinical Faculty:_____

Course Title (No.):_____

Date:_____

PURPOSE: This scale is designed as an evaluation tool to rate each clinical faculty member. It is intended to provide a diagnostic profile for teaching improvement.

PROCEDURES:

Course coordinator will visit new clinical faculty at least once during the semester to evaluate clinical instruction methods. Additional visits may occur as necessary. Following the initial evaluation, course coordinators will evaluate faculty at least bi-annually.

Clinical faculty member will complete a self-evaluation using this tool, including the two open-ended responses.

A ***meeting*** will be scheduled between the course coordinator and the faculty member to discuss the results following the clinical site visit. The meeting should occur as soon as possible after the site visit. The self-evaluation form will be used in conjunction with the course coordinator's appraisal to analyze faculty performance and suggest areas for improvement.

DIRECTIONS: Check (✔) your rating of AGREE (**A**) or DISAGREE (**DA**) for each of the teaching behaviors listed below. For each DA response, please explain in the comments section to the right. For any behaviors NOT APPLICABLE (**NA**) or NOT OBSERVED (**NOB**), check as appropriate.

DOI: 10.2202/1548-923X.1024

	A	DA	NA	NOB	Comments:
TEACHING METHODS					
Demonstrates:					
Professionalism					
Enthusiasm					
Respect for students, staff, and patients					
Current clinical knowledge					
Thorough preparation for clinical experience					
Sensitivity to gender and cultural differences					
Effective communication skills (e.g., constructive, nonthreatening feedback, appropriate verbal and nonverbal responses)					
Assures safe application of clinical care					
Assists students to apply theory to practice					
Assures clinical experiences are appropriate for course level					
Facilitates critical thinking skills					
Provides timely verbal feedback					
Provides constructive verbal feedback					
Organizes clinical experiences to maximize learning					
Varies teaching strategies according to student characteristics and abilities					
Provides insightful written feedback to students					
Provides timely written feedback to students (e.g., nursing care plan comments, e-mail, memos)					
Grading clearly discriminates among different levels of performance					

	A	DA	NA	NOB	Comments:
COURSE MANAGEMENT					
Follows clinical course guidelines					
Adheres to established student clinical hours					
Attends clinical course-related meetings					
Develops a written, systematic plan for student improvement as appropriate					
Provides course coordinator with appropriate feedback on selected students					
Provides course coordinator with timely feedback on selected students					
Seeks assistance from course coordinator as appropriate					
Assists with additional course activities as necessary (e.g., developing care plan guidelines, revising student evaluation tool)					
Submits students' grades promptly					

Strengths:

Areas for Improvement:

_____

Signature

DOI: 10.2202/1548-923X.1024

REFERENCES

AERA (American Educational Research Association), APA (American Psychological Association) & NCME (National Council on Measurement in Education) Joint Committee on Standards. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Aleamoni, L. M. (1982). Components of the instructional setting. *Instructional Evaluation, 7,* 11–16.

Appling, S. E., Naumann, P. L., & Berk, R. A. (2001) Using a faculty evaluation triad to achieve evidenced-based teaching. *Nursing and Health Care Perspectives, 22,* 247–251.

Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system,* (2nd ed.). Bolton, MA: Anker.

Boyer, E. (1990). *Scholarship reconsidered: New priorities for the professoriate.* Princeton, NJ: The Carnegie Foundation for the Advancement of Teaching.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work.* San Francisco: Jossey-Brass.

Braxton, J. M., Bayer, A. E., & Finkelstein, M. J. (1992). Teaching performance norms in academia. *Research in Higher Education, 33*(5), 533–569.

Cederblom, D., & Lounsbury, J. W. (1980). An investigation of user acceptance of peer evaluations. *Personnel Psychology, 33,* 567–580.

Centra, J. A. (1999). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness.* San Francisco: Jossey-Bass.

Chickering, A. W., & Gamson, Z. F. (1994). Seven principles for good practice in undergraduate education. In K. A. Feldman & M. B. Paulsen (Eds.), *Teaching and learning in the college classroom* (pp. 255–263). Needham Heights, MA: Ginn.

Chism, N. V. N. (1999). *Peer review of teaching: A sourcebook.* Bolton, MA: Anker.

Cohen, P. A., & McKeachie, W. J. (1980). The role of colleagues in the evaluation of teaching.*Improving College and University Teaching, 28,* 147–154.

Costello, J., Pateman, B., Pusey, H., & Longshaw, K. (2001). Peer review of classroom teaching: An interim report. *Nurse Education Today, 21,* 444–454.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology, 60,* 151–74.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.

DeZure, D. (1999). Evaluating teaching through peer classroom observation. In P. Seldin & Associates (Eds.), *Changing practices in evaluating teaching: A practical guide to improving faculty performance and promotion/tenure decisions.* Bolton, MA: Anker.

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of students evaluations of teaching effectiveness. *Quality Assurance in Education: An International Perspective, 11*(1), 37–47.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30,* 137–189.

Freeberg, N. E. (1969). Relevance of rater-ratee acquaintance in the validity and reliability ratings. *Journal of Applied Psychology, 53,* 518–524.

Hollander, E. P. (1956). The friendship factor in peer nominations. *Personnel Psychology, 9,* 435–447.

Hutchings, P. (Ed.). (1995). *From idea to prototype: The peer review of teaching.* Washington, DC: American Association for Higher Education.

Keig, L. W., & Waggoner, M. D. (1994). *Collaborative peer review: The role of faculty in improving college teaching* (ASHE/ERIC Higher Education Report, No. 2). Washington, DC: Association for the Study of Higher Education.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 44–53.

Love, K. G. (1981). Comparison of peer assessment methods: Reliability, validity, friendship bias, and user reaction. *Journal of Applied Psychology, 66,* 451–457.

Ludwick, R., Dieckman, B. C., Herdtner, S., Dugan, M., & Roche, M. (1998). Documenting the scholarship of clinical teaching through peer review. *Nurse Educator, 23*(6), 17–20.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52,* 1187–1197.

Martsolf, D. S., Dieckman, B. C., Cartechine, K. A., Starr, P. J., Wolf, L. E. & Anaya, E. R.(1999). Peer review of teaching: Instituting a program in a college of nursing. *Journal of Nursing Education, 38,* 326-332.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52,*1218–1225.

DOI: 10.2202/1548-923X.1024

Millis, B. J., & Kaplan, B. B. (1995). Enhancing teaching through peer classroom observations. In P. Seldin & Associates (Eds.), *Improving college teaching* (pp. 137–152). Bolton, MA: Anker.

Morehead, J. W., & Shedd, P. J. (1997). Utilizing summative evaluation through external peer review of teaching. *Innovative Higher Education, 22*(1), 37–44.

Muchinsky, P. M. (1995). Peer review of teaching: Lessons learned from military and industrial research on peer assessment. *Journal on Excellence in College Teaching, 6*(3), 17–30.

Murray, H. G. (1997). Effective teaching behaviors in the college classroom. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 171–204). New York: Agathon.

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education, 27*(2), 187–198.

Netemeyer, R. G., Bearden, W. O., Sharma, S. (2003). *Scaling procedures.* Thousand Oaks, CA: Sage.

Nunnally, J., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Palmer, P. (1998). *The courage to teach: Exploring the inner landscape of a teacher's life.* San Francisco: Jossey-Bass.

Payne, S. L. (1951). *The art of asking questions.* Princeton, NJ: Princeton University Press.

Remmers, H. H., & Brandenburg, G. C. (1927). Experimental data on the Purdue ratings scale for instructors. *Educational Administration and Supervision, 13,* 519–527.

Rice, R. E. (1991). The new American scholar: Scholarship and the purposes of the university. *Metropolitan Universities, 1*(4) 7–18.

Richlin, L., & Manning, B. (1995). *Improving a college/university teaching evaluation system: A comprehensive two-year curriculum for faculty and administrators.* Pittsburgh, PA: Alliance.

Root, L. S. (1987). Faculty evaluation: Reliability of peer assessments of research, teaching, and service. *Research in Higher Education, 26,* 71–84.

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin, 63,* 129–156.

Schmelkin-Pedhazur, L., Spencer, K. J., & Gellman, E. S. (1997). Faculty perspectives on course and teacher evaluation. *Research in Higher Education, 38*(5), 575–592.

Schmitt, N., & Lappin, M. (1998). Race and sex as determinants of the mean and variance of performance ratings. *Journal of Applied Psychology, 65,* 428–435.

Seldin, P. (1999). Current practices – good and bad – nationally. In P. Seldin & Associates (Eds.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenured decisions* (pp. 1–24). Bolton, MA: Anker.

Seldin, P., & Associates. (Eds.). (1999). *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions.* Bolton, MA: Anker.

Shoffner, D. H., Davis, M. W., & Bowen, S. M. (1994). A model for clinical teaching as a scholarly endeavor. *Image, 26*, 181–184.

Shulman, L. S. (2004). *The wisdom of practice: Essays in teaching, learning, and learning to teach.* Indianapolis, IN: Wiley/Jossey-Bass.

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use* (2nd ed.) New York: Oxford University Press.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions: A practical guide to questionnaire design.* San Francisco: Jossey-Bass.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*, 25–29.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude.* Chicago: University of Chicago Press.

Wang, K. A. (1932). Suggested criteria for writing attitude statements. *Journal of Social Psychology, 3*, 367–373.

Webb, J., & McEnerney, K. (1995). The view from the back of the classroom: A faculty-based peer observation program. *Journal on Excellence in College Teaching, 6*(3), 145–160.

Wherry, R. J., & Fryer, D. C. (1949). Buddy rating: Popularity contest of leadership criterion? *Personnel Psychology, 2*, 147–159.

DOI: 10.2202/1548-923X.1024