Sources of Error in Student Evaluation of Teaching

Paul G. Grussing
College of Pharmacy, M/C 871, University of Illinois at Chicago 833 South Wood Street, Chicago IL 60612

Because of the importance of assessing teaching effectiveness based on reliable and valid instruments, a study was conducted to identify dimensions of pharmacy teaching. Sources of rating error which were the object of the study are summarized together with examples of error-reducing approaches. The importance of reduction of error in processes which contribute to periodic performance reviews and the promotion and tenure process is stressed.

INTRODUCTION

However teaching is valued in the cultures within colleges of pharmacy, it is necessary for peers and administrators to make judgments about the effectiveness of teaching. These judgments affect the outcomes of annual performance reviews as well as of the promotion and tenure process. More importantly, if instructors are to have confidence in instructional performance rating systems and expect to improve teaching performance based on feedback provided by such systems, error in ratings should be minimized. Student ratings of instructor performance provide but one type of data for evaluation of teaching. Potential for imprecision in this source alone can be traced to at least eight sources of performance rating error.

SOURCES OF RATING ERROR

Error in Instrument Content

Most rating instruments developed for student evaluation of instruction are based on classroom teaching. It is possible that teaching in the laboratory or in the practice environment is often unrated. Moreover, dimensions of classroom teaching behavior, and specific rating items within those dimensions, may have been inappropriately applied to evaluation of teaching in other settings. Without a logical linkage between classroom teaching behaviors and those used in other environments, it can not be assumed that the lecture rating instruments are generalizable. Where the teaching dimensions differ, separate rating scales for each pharmacy teaching environment; classroom lecture, laboratory and practicum, are called for. For dimensions of teaching which are similar across teaching environments, common scales may be used, as previously reported(1). A summary of scales for teaching in three settings is shown in Table I. Faculty committees may elect to assign weights to the various dimensions and scales, yielding an overall student rating of instruction.

Content errors may also occur in ratings of conditions which are beyond the instructor's control. Examples include asking students to: (*i*) rate curricular issues, *e.g*., appropriateness of prerequisites to an instructor's course; (*ii*) rate the learning environment *e.g*., classroom lighting, temperature or ventilation; or (*iii*) report the students' self-rated learning interests and accomplishments. These may be useful areas for curriculum and program evaluation but they are based on institutional and student variables, not instructor performance. For supervisor use of all ratings, teaching behaviors, *e.g*., "The instructor followed a course outline", should be separated from student behaviors, *e.g*., "I sought help when I didn't understand the material", or physical environment descriptors, *e.g*., "The classroom space provided a suitable environment."

Error in Interpretation of the Meaning of Ratings

Use of norms for interpreting student ratings of instruction is essential for meaningful and fair comparisons. However, the nature of the norm group may contribute error to interpretations. For example, because studies have shown that instructors in science courses receive lower ratings than those teaching applied life studies and education, the meaning of campus-wide norms is questioned(2,3). Beyond the variables of curricular content, numerous other variables may contribute to errors in normative interpretation; levels of student maturation, levels of education, elective vs. core requirements, and instructional variables(4). Rating scale norms should be developed for several norm groups to make comparisons more meaningful, *e.g*., elective/core, educational level, department, lecture/laboratory, practicum, college and university-wide. The consequences of any type of error may be more pervasive than interpretations of single ratings. Negative experiences with rating systems may serve to reduce confidence in the faculty evaluation process, discouraging both faculty and administrators.

Showmanship

The impact of an instructor's presentation style on students' learning and on their perceptions of instructor satisfaction has been shown(5). Presentations which are remembered for their seductive, entertaining features may produce error in assessment of some aspects of teaching effectiveness, such as content "coverage." However, aspects of an instructor's "personality" and style which are observable in terms of caring, empathic support of students' learning and motivating behaviors contributing to a stimulating, wholesome learning environment do contribute to learning and thus may appear as components of factors associated with effective teaching(6,7). Rating scales should avoid student ratings of instructor "personality," "charisma" or similar attributes. Only those instructor traits which have been shown to be related to effective teaching should be emphasized, *e.g*., "student-teacher interaction," or "concern for students' learning."

Table I. Rating scale content for ten dimensions of pharmacy teaching in three environments

Teaching dimensions	Teaching environments		
	Classroom	Laboratory	Practice
1. Teaching ability—Lecture	X		
2. Teaching ability—Laboratory		X	
3. Teaching ability—Experiential			X
4. Course organization	X	X	X
5. Selection and use of media	X	X	
6. Student performance evaluation	X	X	X
7. Student-instructor interaction	X	X	X
8. Workload/course difficulty	X	X	X
9. Enthusiasm/motivation	X	X	X
10. Knowledge of subject area	X	X	X

Source: Reference #1.

Error in Instrument Reliability

When multiple items appear in traditional, numerically-anchored rating instruments, those items of similar, or related meaning which are summated or averaged as scales to measure performance dimensions should show highly intercorrelated student ratings. Estimates of internal consistency should be reported to instrument users to demonstrate the reliability of such multi-item scales. In addition, the use of rating instruments should show stability across subsequent administrations. Test-retest reliability data permit additional assurances of instrument quality when the traits or behaviors measured are known to remain stable over time. Student ratings for scales should also be shown to correlate highly with those purporting to measure the same dimensions of performance in similar well-established instruments of high reliability and validity. This kind of parallel forms reliability estimate is related to the concept of establishing concurrent validity of the meaning of scores obtained using similar scales. Confidence in the meaning of ratings can be enhanced by development of scales (clusters of related rating items) based on validated dimensions of teaching performance.

Mixed Purposes of Evaluation

Instructors inherently seek self-improvement and utilize evaluation methods to obtain feedback about the effectiveness of innovations or other adjustments in their instruction. This formative approach may or may not be prompted by colleague mentors or department administrators[1]. Independent of the initiative, such self-evaluation is a sensitive and personal process, requiring that instructors have freedom to select and interpret the "results" of such mid-or end-of-course measurements. Instructors also participate cooperatively in institutionally-required evaluations which are summative in nature and used for periodic departmental performance reviews or for making promotion and tenure decisions[2]. If summative and formative data are mixed by administrators and both are considered on a summative basis in personnel decisions, errors in rating occur because of the inappropriate, confounding use of personal formative data, perhaps lowering total performance ratings(8,9). Instructional rating systems should provide for both purposes: a fixed set of items and scales for summative evaluation, and a set of instructor-selected items for the purpose of personal introspection in formative evaluation.

Inconsistent Methods of Instrument Administration

Numerous administrative practices contribute to faulty student ratings of instruction and to lack of confidence in the data collected. Such practices include lack of standardization of administration time, *e.g*., collecting ratings too early in the term, too near the final examination period, without sufficient time for instrument completion, and administration at different times in the term for each instructor. Because errors in rating may also be attributed to lack of stated purpose of the evaluation process, formative vs. summative purposes should be explained to raters and ratees, indicating what specific use is to be made of the data and by whom(10-12). If instruments are introduced and distributed by instructors and completion is supervised in their presence, questions of inappropriate influence may occur. Student administration of rating instruments has been recommended as a remedy to inappropriate instructor influences(13). Finally, errors of measurement have occurred because instruments have not been collected in a systematic, secure manner. Students have been reported to collaborate in the rating activity after leaving the classroom and before turning in the instrument to a designated student representative. When completion of the rating is not accomplished in a standardized classroom situation, accuracy is also compromised because of lack of representativeness when only a portion of students complete and turn in the ratings. A final threat to the integrity of ratings occurs when multiple course instructors are rated using the same instrument. For such "team-taught" courses, insufficient length of instructor exposure to students, and inconsistent time periods between times of presentation and rating also contribute to rating error. Error attributable to administration methods can be minimized by use of standardized instructions, given by a designated college specialist, who states the purpose and utility of the rating process while administering the scales at regular end-of-instruction times for all courses.

Common Rating Error Effects

At least five kinds of rating effects may cause rating errors. The term "Halo effect" is applied to situations in which specific ratings are not based on observation of relevant performance but upon some antecedent experiences

[1] Formative evaluation refers to evaluation of a process or product to provide feedback for the purpose of making possible mid-process refinements or improvements.

[2] Summative evaluation is conducted to examine the quality or impact of a final, completed process or product.

or attitudes(14). This effect typically occurs when impressions of successful performance in one area are translated to overrating in another. The opposite, "Reverse halo effect," may occur when students are predisposed to issuing a low instructor performance rating in one area because they may have unfavorable attitudes about another, *e.g.*, about the "department's reputation." ''Leniency effect" has been reported as contributing to rating error, as has its opposite, "Harshness (or strictness) effect"(15). These effects explain consistently high, or low, ratings without regard for differences in the particular performance being rated. Students may feel compelled to rate their instructors highly or they may have a perception that giving low ratings would reflect negatively on them personally, relating to their learning effort or instructor-perceived ability. Giving consistently low ratings, when not based on valid comparative criteria, might be driven by a student's belief in "instructional rigor" or by motivational effects. Instructors have also observed evidence of the "Central tendency effect," resulting in neither high or low ratings based on performance, but on the tendency to provide middle-of-the-scale ratings(16,17). Such a rating approach may be supported by student fears that low, or high, ratings might require additional justification, or that low ratings might interfere with an instructor's personal professional relationship with students. Appropriate standardized instructions to student raters can minimize such common rating error effects.

Errors in Data Implementation

Some academic administrators may feel uncomfortable when giving feedback to employee-colleagues(18,19). Perhaps such supervisors fear using and interpreting instructional performance data because they lack management training or experience, or because they do not consider the rating scales (or indeed the entire faculty performance evaluation process) to be objective, thus devaluing both the data and the process. Errors may manifest themselves in incomplete or erroneous feedback, or lack of any useful feedback at all. Part of the developmental process for inexperienced administrators should include training in provision of performance feedback, using reliable and valid performance data.

SUMMARY

The challenge to reduce error occurs for instructors, departmental peers, departmental administrators, deans and campus instructional and evaluation services alike. Instructors' inherent interest in self-improvement requires that formative teaching performance feedback be reliable, valid and confidential. Faculty peers benefit from enhanced objectivity in their mentoring of colleagues. Departmental and college administrators require performance measures which meet professional and legal standards for validity and reliability. University service agencies involved in supervision and administration of the instructional evaluation process exist to provide the most reliable and valid measures possible. Emphasis on "publish or perish" demands quality calibration in "bench" and social science research processes. Perhaps a balanced emphasis, respecting a "teach or perish" paradigm, might enhance emphasis on accuracy in measurement and evaluation of the instructional process. The teaching landscape is ever-changing. Weil-developed measures of ability to teach may, where problem-solving or other "new" instructional approaches are in place, require modification of rating systems to reflect a broader continuum of teaching attributes in all contemporary teaching modes.

Am. J. Pharm. Educ., **58**, 316-318(1994); received 1/21/94, accepted 6/2/94.

References

(1) Grussing, P.G., Valuck, R. and Williams, R.G., "Development and validation of behaviorally-anchored rating scales for student evaluation of pharmacy instruction", *Am. J. Pharm. Educ.*, **58**, 25-37(1994), *N.B.*p.33.
(2) Centra, J.A. and Creech, F.R., "The Relationship Between Student, Teachers, and Course Characteristics and Student Ratings of Teacher Effectiveness," Project Report 76-1, Educational Testing Service, Princeton NJ (1976).
(3) Measurement and Research Division, Office of Instructional Resources, "ICES Norms," Unpublished Report, University of Illinois, Urbana IL (1977-1983).
(4) Braskamp, L.A., Brandenburg, D.C. and Ory, J.C., *Evaluating Teaching Effectiveness*, Sage Publications, Newbury Park CA (1984) pp. 29-76.
(5) Ware, J.E. and Williams, R.G., "The Dr. Fox effect: A study of lecture effectiveness and ratings of instruction," *J. Med. Educ.*, **50**, 149-156 (1975).
(6) Hildebrand, M., Wilson, R.C. and Dienst, E.R., *Evaluating University Teaching*, Center for Research and Development in Higher Education, Berkely CA (1971) pp. 18-20.
(7) Dickinson, T.L. and Zelliger, P.M., "A Comparison of the behaviorally anchored rating and mixed standard scale formats." *J. Appl. Psychol.*, **65**, 147-154 (1980).
(8) *Op. cit.* (4) pp. 25-26, 49-50.
(9) Manning, R.C., *The Teacher Evaluation Handbook*, Prentice-Hall, Englewood Cliffs NJ (1988) pp. 4-5.
(10) *Ibid.*, *p.* 6-9.
(11) Centra, J.A., *Determining Faculty Effectiveness*, Jossey-Bass. San Francisco CA (1979) pp. 7-11.
(12) *Op. Cit.*, (4), p. 51.
(13) *Op. Cit.*, (4), p. 51-52.
(14) *MacMillan Dictionary of Psychology.* (edit. Sutherland, S.,) MacMillan, London (1989) p. 183.
(15) *Encyclopedia of Psychology*, Vol. 3, (edit. Corsini, R.), John Wiley and Sons, New York, NY (1984) p. 205.
(16) Smith, P.C., "Behaviors, results and organizational effectiveness: The problem of criteria," in *Handbook of Industrial Psychology*, (edit. Dunnette, M.) Wiley and Sons, New York, NY (1983) p. 757.
(17) *Op. cit.*, (17), p. 205.
(18) *Op. Cit.*, (4), p. 80.
(19) Ivancevich, J.M., *Human Resource Management*, 5th ed., Irwin, Homewood IL, (1992) pp. 327-327.