Validating Dental and Medical Students’ Evaluations of Faculty Teaching in an Integrated, Multi-Instructor Course

Terry D. Stratton, Ph.D.; Donald B. Witzke, Ph.D.; Mary Jane Freund, M.A.E.; Martha T. Wilson, M.A.; Robert J. Jacob, Ph.D.

Abstract: As more students from various health professions are combined into integrated courses, evaluating the teaching quality of individual faculty in these typically large, multi-instructor contexts becomes increasingly difficult. Indeed, students who lack sufficient recall of a given faculty member or are not committed to the evaluation process may respond by marking identical responses to all evaluation items (e.g., 3-3-3-3-3), regardless of the specific content of the items on the faculty evaluation questionnaire. These “straight-lining” behaviors—more formally referred to as monotonic response patterns (MRPs)—often reflect students’ inattention to the task at hand or lack of motivation to be discriminating, which may result in invalid data. This study examines the prevalence of MRP ratings in relation to indicators reflective of students’ lack of attention to evaluating the quality of faculty teaching. Dental and medical students in a required, second-year (medicine) basic science course conducted by the medical school and taught primarily by medical school faculty completed seven-item faculty evaluation forms, along with an anonymous questionnaire measuring their need to evaluate, attitudes toward faculty evaluation, and recall of instructors. MRP ratings failed to correlate significantly with students’ need to evaluate or their attitudes toward faculty evaluation. However, among medical students, MRP “straight-line” responses were more prevalent for raters who recalled faculty members “very well” \( (p=.04) \). For dental students, MRPs were associated with less accurate recall \( (p=.01) \). As such, the validity of faculty evaluations within integrated, multi-instructor courses may vary when students rate distinct aspects of a teacher’s performance identically. In this case—in which medical students’ greater recall of instructors coincides with MRPs—ratings may suffice as global, holistic assessments of an instructor’s teaching. For dental students, similar ratings may be less viable. Individual item analysis is cautioned under any circumstances.

Dr. Stratton is Assistant Dean, Student Assessment and Program Evaluation and Assistant Professor, Department of Behavioral Science, University of Kentucky College of Medicine; Dr. Witzke is Associate Professor, Department of Pathology and Laboratory Medicine, University of Kentucky College of Medicine; Ms. Freund is Educational CQI Manager, Office of Medical Education and Curricular Affairs, Indiana University School of Medicine; Ms. Wilson is Curricular Consultant, Office of Academic Affairs, University of Kentucky College of Medicine; and Dr. Jacob is Associate Professor, Department of Microbiology, Immunology, and Molecular Genetics, University of Kentucky College of Medicine and Department of Oral Health Practice, University of Kentucky College of Dentistry. Direct correspondence to Dr. Terry D. Stratton, Office of Academic Affairs, University of Kentucky College of Medicine, 800 Rose Street, MN104-UKMC, Lexington, KY 40536-0298; 859-323-2785 phone; 859-323-2076 fax; tdstra00@uky.edu.

Key words: faculty evaluation, medical-dental integration, interprofessional education

Submitted for publication 2/1/05; accepted 4/14/05

In response to the 1995 Institute of Medicine (IOM) report urging greater curricular integration of basic and clinical content, dental educators have increasingly crossed disciplinary boundaries—often with medicine—in an effort to more rigorously prepare their students in the biological sciences (e.g., microbiology, biochemistry, immunology, and anatomy). Moreover, as nutrition, genetics, and other scientific areas of overlap become more evident, and the effects of the aging population on the health care system continue to be felt, the integration of medicine and dentistry in the basic sciences and, to a lesser extent, in clinical training will likely continue to increase.

Concurrent with the movement toward more integrated curricula, however, are changes in the academic culture that are driving a renewed emphasis on performance assessment and institutional accountability. Medical and dental faculty are both expected to document the quality of their performance to various stakeholders (e.g., funding agencies, accrediting boards, promotion/tenure committees, department chairs, and students) in each of the tripartite missions of research, service, and teaching—although how these faculty roles are to be measured, weighted, or interpreted varies greatly and is often subject to debate.

Assessing teaching quality arguably remains one of the most difficult and contentious tasks that
While research productivity and service activities are more easily quantifiable, evaluations of individual faculty teaching continue to consist largely of student ratings, in spite of their documented limitations and potential for misuse in decision making. Faculty often view these subjective assessments as little more than popularity contests, while students question the utility of continually providing feedback on behaviors that are only rarely demonstrated to change.

Nowhere is the task of eliciting valid ratings of faculty teaching more challenging than in the large, integrated, expert-taught courses that use multiple instructors to convey voluminous amounts of detailed, complex information to an increasingly diverse group of learners. This “parade of stars” format is commonly employed in basic science courses and makes it difficult to assess individual faculty performance in a class that may span several months and involve only limited exposure of instructors to students.

As with all rating data, the potential sources of error in students’ evaluations of faculty are vast, ranging from rater bias to general instrumentation effects (e.g., question wording, scaling) to timing of administration. Within integrated (multidisciplinary or interprofessional) courses, another potentially confounding variable is added—namely, the presence of students who perhaps feel marginalized by being combined in classes with other students who they do not know or who are apathetic about the value of evaluating faculty from colleges in which they are not heavily invested. In this sense, dental students are faced with whether (and how) to evaluate medical faculty with whom their sole contact is likely restricted to this single exposure.

In all faculty evaluations (and all survey research in general), a fundamental prerequisite to issuing a valid response to any question is the ability of subjects to sufficiently access and recall the necessary information. Unfortunately, even when afforded a “don’t know/can’t rate” option, lacking access to this pertinent information does not preclude subjects from offering responses. This phenomenon, known as “satisficing,” occurs when subjects—for a variety of reasons—are compelled to offer acceptable responses rather than optimal ones. One frequent result of satisficing is a “failure to differentiate among a set of diverse objects in ratings” (p. 213), termed a “monotonic response pattern” (MRP) or, more informally, “straight-lining.” An MRP occurs, for example, when students, in assessing a given faculty member, assign identical ratings across numerous items that measure discrete and different aspects of instructional performance (e.g., 3-3-3-3).

Internal consistency, one common measure of reliability, is the interrelationship among scale items that indicates the degree to which they appear to measure the same thing. Under certain conditions, the effect of MRPs on the internal consistency of responses can be dramatic. Barnette, in an experimental simulation, found that a mere 5 percent prevalence of “mono-extreme” respondents (those marking all items at the absolute highest or lowest end of the response set) inflated the sample reliability to 0.89, markedly above the known population value of 0.70. Extending this line of inquiry, Stratton et al., examining actual ratings of faculty teaching, found that the impact of MRPs on the internal consistency of scale scores fluctuated across instructors being rated.

The goal of this study was to examine the validity of MRPs by examining selected rater attributes reflecting nonattending behaviors in the context of dental and medical students’ evaluations of teaching faculty. Within this framework, the focus is on the contextual factors leading to the cognitive formulation of a given response or pattern of responses. By understanding the type of rater responsible for MRPs, some insights may be gained into the validity of medical and dental students’ ratings when identical responses are issued across all items.

Thus, the following research questions were addressed: 1) Do dental and medical students differ in their ratings of faculty teaching? 2) Does the prevalence of MRPs in faculty ratings vary between dental and medical students? 3) Do MRPs appear to be the result of dental and medical students’ inattention to the evaluation task? and 4) Under what circumstances, and toward what ends, do MRPs represent valid or invalid measures of faculty teaching, and does this differ between dental and medical students?

**Methods**

This study setting was a required, second-year basic science course at the University of Kentucky College of Medicine (UKCOM) titled “Immunity, Infection, and Disease” (MI822). This intensive, expert-taught course meets four hours daily for twelve weeks (~240 total contact hours) and involves twenty
to thirty highly specialized basic scientists and clinicians covering such diverse topics as microbiology, immunology, infectious diseases, pathology, internal medicine, and pediatrics. Second-year medical and second-year dental students attend this course together and complete the same requirements, with two exceptions: 1) dental students participate in two additional small-group sessions focused on oral health issues; and 2) dental students do not participate in the laboratory portion of the course (~37 contact hours). The bulk of the instruction is didactic in nature, and the vast majority of instructors are UKCOM faculty.

Faculty who taught in the first half of the course during 2002-03 and 2003-04 were divided alphabetically into two equal groups (Groups A and B) of six and seven, respectively (this slight difference in the number of faculty rated in 2003-04 reflects merely an administrative need to evaluate different instructors). Along with teaching evaluations of faculty instructors, anonymous packets containing a one-page, thirty-one-item questionnaire measuring selected rater attitudes and attributes (described in the following section) were randomly distributed to dental and medical students immediately following the midterm examination. By intervening at the course midpoint (~7 weeks), students’ recall was bounded within a shorter, more proximate time frame, thus limiting the number of teaching faculty to whom they were exposed.

To summarize, separate halves of each class evaluated two subsets of six (2002-03) or seven (2003-04) faculty across seven dimensions of teaching which were: 1) overall quality, 2) organization, 3) preparation, 4) stimulation, 5) respectfulness, 6) understandability, and 7) clarity. The full text of the faculty evaluation items are listed in the Appendix. Items were measured on a four-point, Likert-type scale with responses consisting of “Outstanding” (4), “More Than Adequate” (3), “Adequate” (2), “Less Than Adequate” (1), and “Unable to Rate” (NR). Per institutional review board approval, a cover letter detailed the study and students’ voluntary participation, along with instructions to consider the questionnaire only after completing the evaluation forms.

Monotonic response patterns were defined as those forms containing invariant ratings across all seven evaluation items (excluding “Unable to Rate”)—that is, when a student rated a given faculty member identically on all evaluation items (e.g., ranking all items as a “2,” “adequate”). These MRP ratings were represented as the percentage of forms per student. For example, if a student returned five completed faculty evaluation forms, one of which contained identical responses to all evaluation items, then the prevalence of MRP forms for that student was 20 percent. Three criterion measures were used to examine the validity of MRPs in students’ ratings.

Anonymous packets containing a one-page, thirty-one-item questionnaire were administered to participating students. (Note: The questionnaire length in the 2003-04 cohort was thirty-two items to accommodate the evaluation of an additional faculty member.) Three measures were included in the questionnaire: 1) the sixteen-item Need to Evaluate Scale (NES); 2) the eight-item “students’ attitudes toward evaluation of faculty teaching” scale; and 3) a single question that asked students to indicate their level of recall of each faculty (six or seven) that they were requested to evaluate.

First, the sixteen-item Need to Evaluate Scale (NES) assessed “individual differences in the propensity to evaluate” (p. 172) and consisted of items such as “I form opinions about everything” and “I often prefer to remain neutral about complex issues.” The NES is premised on the theory that individuals are variably compelled to assess “the positive and/or negative qualities of an object” (p. 172) independent of situational factors. Using a five-point, Likert-type scale, subjects rated the extent to which each statement is characteristic of themselves: “Extremely Characteristic” (5), “Characteristic” (4), “Uncertain” (3), “Uncharacteristic” (2), or “Extremely Uncharacteristic” (1). NES scores have been shown to be reliable and valid in undergraduate college samples. This criterion was hypothesized to be negatively correlated with MRPs.

Second, since respondents’ interest in the subject matter should bolster their attentiveness to the task at hand, students’ attitudes toward evaluation of faculty teaching were measured. This eight-item scale contained statements such as “It is important for students to provide feedback to faculty on their teaching” and “Faculty seem genuinely interested in receiving feedback from students,” and were rated on a five-point, Likert-type scale ranging from “Strongly Agree” (5) to “Strongly Disagree” (1). Items were reverse-coded as necessary so that higher scores represented more positive views toward faculty evaluations. This criterion was expected to be negatively correlated with MRPs.
Third, we asked students to indicate how well they recalled each of the faculty being rated using a scale that consisted of “Very Well” (4), “Pretty Well” (3), “Somewhat” (2), or “Not at All” (1). This measure was also anticipated to be negatively correlated with MRPs. T-tests for independent samples, Chi-square tests, and Pearson product moment correlation coefficients were used to examine relationships between MRPs and criterion variables. A critical alpha of ≤.05 was specified for all analyses.

Results

From the two cohorts, criterion data were returned from 155 medical students and eighty dental students, comprising an overall response rate of 80.8 percent. Response rates did not differ significantly between the 2002-03 (90.4 percent) and 2003-04 (93.1 percent) academic cohorts or between dental (89.1 percent) and medical (93.1 percent) students. Internal consistency of both the NES (=.87, n=207) and the attitudes toward faculty evaluation measure (=.76, n=218) was good, suggesting the individual items comprising each were measuring the same construct, respectively. Due to low internal consistency, however, a single item was dropped from the “attitudes toward faculty evaluation measure,” reducing it to seven items. Twenty-seven subjects failed to complete all sixteen NES items. To allow calculation of a composite score, mean substitutions were made on up to two missing items per subject (n=17). Similarly, eight students had two or fewer missing responses on the attitudes toward faculty evaluation measure. In these cases, a comparable mean replacement strategy was used. All reliability estimates were made prior to mean substitutions.

With a maximum possible score of 80, student scores on the sixteen-item NES ranged from 26 to 80 (n=224, Mean=50.0, Median=50.0, SD=10.0), with individual item means varying from 2.1 (SD=0.9) on “I am pretty much indifferent to many important issues” to 3.6 (SD=1.1) on “I form opinions about everything.” Summated scores on the faculty evaluation attitudes measure ranged from 11 to 35 (n=226, Mean=24.2, Median=25.0, SD=4.5). Means across items ranged from 2.9, SD=1.2 (“I feel that my evaluations of teaching faculty really make a difference” and “Teaching students simply isn’t a priority for most faculty”) to 4.4, SD=0.8 (“It is important for students to provide feedback to faculty on their teaching”). Neither item nor scale means differed significantly between medical and dental students. Table 1 contains a descriptive summary of scale means, standard deviations, etc., for dental and medical students.

Although students reported recalling individual faculty to varying degrees (see Figure 1), the distributions for most faculty being rated were negatively skewed, with larger percentages of students recalling instructors “Very Well.” Subsequently, faculty recall was dichotomized into: “Very Well” (VW) and “<Very Well” (less than “Very Well”); the latter was comprised of “Pretty Well,” “Somewhat,” and “Not at All.” The percentages of students recalling rated faculty “Very Well” varied greatly: from 28.3 percent to 91.8 percent in 2002-03 and from 18.5 percent to 82.7 percent in 2003-04. Level of recall tended to be slightly, though not significantly, greater for medical students. Indeed, when rank ordered, medical and dental students’ recall of teaching faculty was comparable for both the 2002-03 (r=.80, df=12, p=.002) and 2003-04 (r=.84, df=14, p<.001) cohorts, suggesting the most memorable faculty tended to be similar for the two groups.

The 235 students who completed the thirty-one-item questionnaire also returned 1,477 faculty evaluation forms, 1,177 (79.7 percent) of which contained ratings on all seven items. Of these, 441 (37.5 percent), or slightly more than one of every three forms, contained identical ratings (MRP) across all seven evaluation items. (See Figure 2.) The largest proportion of student raters (31.1 percent) marked invariant ratings for one faculty member, while eleven students (4.8 percent) returned all faculty forms with MRPs. Slightly over a fifth (21.9 percent) of students returned forms containing no

Table 1. Descriptive summary of key measures for dental and medical students

<table>
<thead>
<tr>
<th>Measure</th>
<th>Student</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Alpha</th>
<th># Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>% of Faculty Evaluation Forms Containing</td>
<td>Dental</td>
<td>89</td>
<td>39.9</td>
<td>31.9</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Medical</td>
<td>168</td>
<td>37.1</td>
<td>27.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Need to Evaluate Scale (NES)</td>
<td>Dental</td>
<td>73</td>
<td>50.1</td>
<td>9.4</td>
<td>0.87</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Medical</td>
<td>150</td>
<td>49.5</td>
<td>11.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitudes Toward Teaching Evaluation</td>
<td>Dental</td>
<td>75</td>
<td>23.9</td>
<td>4.1</td>
<td>0.76</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Medical</td>
<td>150</td>
<td>26.1</td>
<td>5.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Journal of Dental Education ■ Volume 69, Number 6
MRPs. Dental students returned a significantly higher percentage of completed faculty evaluation forms than did medical students (86.6 percent vs. 77.7 percent), although the percentage of forms with MRPs did not differ between dental and medical students.

Criterion and MRP data were unable to be accurately matched for forty-one subjects from the 2002-03 medical student cohort, reducing to 194 the sample size for these analyses. In examining criterion-related correlates of MRPs in faculty ratings, we found no relationship with students’ orientations toward evaluating faculty teaching ($r = 0.01$) or NES scores ($r = 0.03$). However, when dental and medical student data were examined separately by year, some noticeable variations emerged. For example, in the 2002-03 group, the correlations of MRPs with orientation toward faculty evaluation and NES scores were markedly larger ($r = -0.25$, $r = 0.35$, respectively) among dental students. Similar correlations for medical students, on the other hand, were largely unremarkable ($r = 0.04$, $r = 0.13$, respectively). Among both dental and medical students in the 2003-04 cohort, all correlations were sufficiently small to be attributable to chance.

Finally, using individual faculty forms as the unit of analysis, we examined the relationship between MRPs and students’ recall of the instructor, the final criterion measure (see Figure 3). Based on a cross-tabulation of dichotomous measures of MRPs (yes/no) and instructor recall (VW/<VW), a significantly greater percentage of medical student forms contained MRPs when the instructor was recalled “Very Well” (44.7 percent vs. 36.2 percent). Conversely, dental students were significantly more likely to issue MRP ratings to instructors they remembered “<Very Well” (52.9 percent vs. 40.3 percent).

For both dental and medical students, MRP ratings were significantly more likely to be situated at the extreme ends of the scale for those faculty instructors they recalled “Very Well.” Indeed, 43.1 percent of dental and 57.8 percent of medical students’ MRP ratings were either “<Adequate” or “Outstanding,” compared with 16.8 percent and 14.0 percent of their classmates who recalled the faculty member “<Very Well.” Conversely, then, students’ MRP ratings based on less accurate recall (i.e., <Very Well) were significantly more likely to fall in the middle of the scale (i.e., Adequate or >Adequate).

**Figure 1.** Dental and medical students’ recall of faculty being evaluated, 2002-03 and 2003-04

**Figure 2.** Students’ faculty evaluation forms containing identical ratings for all items (MRPs)
Discussion

Using a rater-centered approach, this study examined the validity of dental and medical students’ ratings of faculty teaching in an integrated, multi-instructor course by correlating MRPs with indicators of inattentiveness to the evaluation task. While the findings do not provide definitive evidence of the validity of these data, they do suggest a complex model of why some raters fail to distinguish among various aspects of teaching quality—and how valid these ratings may be in assessing instructional performance.

Like clinical performance ratings,22 ratings of teaching performance are subject to a host of influences ranging from extraneous, environmental factors (e.g., the acoustics of the room) to internal, rater characteristics (e.g., lack of sufficient recall), all of which add to the collective measurement error. Of course, unlike assessing students’ clinical competence—which is explicitly part and parcel of being an academic educator—it is unclear whether or not providing feedback on faculty instruction is seen as an obligatory component of the dental or medical student role. More specifically, are students able or motivated to provide valid evaluations of faculty outside their home colleges?

Although the finding that MRPs were largely unrelated to students’ NES and attitudes toward faculty evaluation scores does not ensure validity, neither does it provide evidence of invalidity. Moreover, it cannot be presumed that recalling an instructor “Somewhat” or “Pretty Well” is insufficient to issuing valid ratings of teaching quality. Yet, since medical students were more likely than dental students to issue identical ratings to faculty they recalled “Very Well,” there is the suggestion that these may suffice as valid global, holistic ratings of teaching. Of course, if medical students’ recall is based on prior exposure to faculty outside of the course, then ratings may be compromised even as global assessments. Under any circumstances, however, individual item analysis should be discouraged.

Exactly why dental students appear more susceptible to issuing MRPs of faculty that they remembered less well is unclear. To conjecture, some research has found that a faculty member’s departmental affiliation is among factors deemed to be salient by raters evaluating instructional quality.23 If this is the case, perhaps dental students are simply less invested in providing discerning evaluations if they perceive the class to be a medical school course taught by medical school faculty and thus view the process as less consequential. Although our institution has had a combined M.D./D.M.D. track in place since the mid-1990s, the dental and medical curricula and their associated faculties have understandably remained largely distinct. This, too, may have contributed to an “in-group/out-group” characterization among dental students taking required coursework within an existing medical curriculum.

This study is limited by several factors. First, the sample is limited to two consecutive cohorts of students in a single program, so the generalizability of the findings is unknown. As a result, it should be viewed as a pilot study, with findings that are largely preliminary in nature. Second, the exact nature of the criterion measures’ relationships to respondent inattention is unclear. The measure devised to reflect students’ attitudes toward teaching evaluation was not rigorously validated, and the use of the NES in this context is a relatively novel application. Third, the fact that dental students did not participate in the laboratory portion of the course is noteworthy. However, since these sessions were coor-
ordinated by a single instructor who was not among those faculty evaluated, we believe the impact of this to be negligible. Lastly, although we explicitly instructed students to complete the questionnaire only after filling out the accompanying evaluations, some noncompliance is possible. If prevalent, the criterion measures may have unintentionally exerted some bias on subsequent faculty ratings.

As raters, students are rarely given any guidance or instruction about what constitutes good teaching—much less how to recognize and assess its concomitant indicators and behaviors. Of course, while students are perhaps ideally situated to rate selected aspects of teaching quality, evaluating the performance of individual instructors is largely peripheral to students’ central roles as learners. Moreover, faced with ever-tightening time pressures in medical and dental curricula, students may feel overwhelmed with the added task of assessing numerous individual instructors. Still, by carefully balancing the desired need for sound, comprehensive measures with the cognitive effort required to provide valid responses, ratings could be garnered that are adequately useful for many (but not necessarily all) purposes.

An important but often overlooked aspect of rating data is its consequential validity or the appropriateness of ends to which it is directed. If the intent is simply to offer a summative assessment of general teaching prowess at a broad, global level for the sole purpose of documenting or rank-ordering individuals’ instructional effort, then MRPs—if provided by raters with reasonable recall of the faculty member being rated—will probably suffice. If, on the other hand, the ratings are to be used to assess discrete areas or aspects of instruction, perhaps from which to direct faculty development efforts, MRPs from any source are inherently problematic. To the extent that ratings of faculty instruction are seriously weighted in tenure and promotion decisions or other high stakes applications, individual institutions and evaluators must decide what level of validity and rigor is sufficient for the purpose at hand and whether MRP ratings or ratings from certain student cohorts should be included in these applications.

At our institution, we have historically excluded from official compilations all ratings of medical faculty provided by nonmedical students, although this decision was based more on bureaucratic protocol than empirical evidence. Nonetheless, dental students are considered equal participants in integrated courses and, as such, represent a viable and potentially useful source of information on the quality of teaching. Course directors of integrated classes should reinforce to all students, but particularly those from companion colleges, their roles as curricular stakeholders in the educational process. Dental educators, too, should drive home this very fact with students: describing their roles, rights, and obligations in helping to ensure the quality improvement of integrated courses residing outside their home college. In the interim, as the number of multi-instructor courses accommodating integrated student bodies increases, attention should be given to concurrently examining the validity of each group’s ratings.

REFERENCES

APPENDIX

Dental and medical students in this study used the scale below to rate instructor quality.

Faculty Evaluation Items

- The organization (e.g., clarity, pace) of this instructor’s teaching was:
- This instructor’s preparation for educational activities was:
- The amount of stimulation to learn provided by this instructor was:
- The degree of respect this instructor demonstrated toward me was:
- The ability of this instructor to make complex material easy to understand was:
- The ability of this instructor to present detailed information in a clear manner was:
- The overall quality of this instructor’s teaching was:

Scale: 1=Less Than Adequate, 2=Adequate, 3=More Than Adequate, 4=Outstanding, 5=Unable to Rate