Do Repeated Clinical Competency Ratings Stereotype Students?

David W. Chambers, Ed.M., M.B.A., Ph.D.

Abstract: A stereotype is a generalization that protects itself from critique by limiting or distorting new information. The potential for stereotyping exists where faculty members repeatedly rate students’ clinical competency. Stereotyping is difficult to study because of methodological problems. If, for example, a student’s score remains low over repeated ratings, it may be because the faculty member has pegged the student as a poor performer or because the student is in fact a consistently low performer. An existing dataset of clinical competency ratings for almost 300 students was divided so that ratings given by faculty members who had evaluated students previously could be compared to ratings of the same students by faculty members who had not previously evaluated these students. This study supports the following conclusions: 1) repeating faculty members use both current information and carryover information from previous rating periods; 2) the amount of information carried over increases from quarter to quarter; and 3) faculty evaluators who use more carryover information are more accurate in predicting students’ graduation competency level than are faculty members on their initial ratings of students. In conclusion, there is no evidence in this study that repeated clinical competency ratings promote stereotyping of students by faculty members.

Key words: evaluation, competency, ratings, clinical, stereotyping, bias

Submitted for publication 8/23/04; accepted 10/13/04

Stereotypes are generalizations.1 That means they are approximately true statements, omitting the nuance and detail of comprehensive and individual description. We need generalizations to make communication efficient, and sometimes a very rough idea is all that is needed. “That is a research-intensive school,” “she is a strong candidate for the graduate program,” and “the class seems fidgety today” are generalizations that would raise few eyebrows and would even likely continue to be held in the face of a detail or two here and there not exactly expected.

Not all generalizations are sound. Some are too ambiguous to be useful: “The patient might be fine.” Some are too precise to travel well: “If you really want to know how good the student is, I will photocopy all his patient charts.” Perhaps they are based on insufficient evidence: “I only worked with her once.” Some generalizations are misleading; they are off center and will lead to surprises if acted upon. We call these biased generalizations: “We never had any really good applicants from that feeder school.”

Stereotypes are a special class of generalizations, usually thought of as being negative, biased, and based on insufficient evidence. But they can also be overly positive and well grounded in experience, such as the feelings we usually have for our friends. The distinguishing characteristic of a stereotype is that it distorts new information. Stereotypes are self-protecting in the sense of making confirmatory evidence more readily available and erecting a barrier against disconfirming observations.1

Although generalizations are unavoidable in dental education and are largely useful, we should work to eliminate stereotypes. A faculty member who is slow to understand the subtleties of teaching and has a few rough classes in the first years can be branded as a poor teacher and later commendations get discounted. A school with an established reputation may not be scrutinized carefully by applicants.

The research reported here addresses the possibility of stereotypes in clinical grading. In cases where faculty members rate students repeatedly, the potential exists for initial ratings to dominate the perception of later performance.

Humphris and Kaney2 found that examiners judging medical students’ communication ability preserved the same standards across extended peri-
ods of time. Previous ratings did not dull their attention to the evaluation task. In a study of family practice residents, Bornsten et al. found a willingness to change treatment decisions based on new information, regardless of whether the previous treatment decision was the resident’s own or one that had been made by another resident. Ryan et al. performed a complex statistical analysis on the ratings of residents in emergency department rotations. They found that leniency (defined as the higher difference between ratings by particular staff compared with the average of all ratings) was positively correlated \( r = .52 \) with overall ratings. They interpreted these findings as being consistent with the existence of a positive stereotype. Interrater agreement in the assessment of neurological signs was found to be higher when background on the case was made available to the evaluator—again a case of positive stereotyping. In a generalizability study of judging social dance, the dancer-by-judge-by-performance interaction was found to account for 15 percent of the variance. Regrettably, in the study design used, this component is contaminated with error variance, so it is impossible to determine how much smaller this interaction effect really is.

Although there is a small body of literature that suggests the existence of a small and positive stereotyping in repeated evaluations of performance, all of these studies are suggestive at best. Each has one or more methodological limitation. There is an inherent difficulty in studying stereotyping in repeated measures. Assume that data show that Rater A judges Student X to be very good on two or three successive occasions. There are at least two plausible explanations: 1) Rater A has stereotyped Student X and discounted information about later performance, and 2) Student X is consistently good and that fact has been correctly registered by Rater A. On the strength of the data themselves, there is no way to distinguish between the two explanations.

In the study reported in this article, this methodological challenge is addressed by using a dataset in which two sets of ratings are compared for students at Time; those where the same faculty members rated these students at Time, and those where the same students were rated for the first time by a new set of faculty members. By comparing the ratings from first-time evaluators with ratings of the same students by faculty members who previously evaluated these students, it should be possible to form an estimate and characterization of the carryover of information in subsequent ratings of students (a measure of stereotyping). This approach resembles the theoretical discussion on regressed reading scores found in Wainer.

Based on the literature and my personal experience with the clinical rating system studied, it is hypothesized that:

1. Faculty members will carry over information from previous ratings of students;
2. The more frequently faculty members rate the same students, the more information will be carried forward; and
3. Carried-over information will improve the accuracy of the evaluation system.

**Materials and Methods**

The University of the Pacific Arthur A. Dugoni School of Dentistry has used a competency evaluation system in its comprehensive care clinic since 1996. Ratings are comprehensive professional judgments of students’ capacity to perform, in this case made by faculty members at the end of each clinical quarter. The scale used at the University of the Pacific includes four action-oriented levels: “qualified”—promote or graduate now; “appropriate progress”—no change in educational program needed; “questionable”—evaluate for remediation; and “dangerous”—evaluate for dismissal. These ratings count for from 20 percent to 100 percent of students’ quarterly grades, depending on the discipline and the quarter. Approximately sixty faculty members rate the roughly 150 second-year students at the end of quarters 5 through 8, and eighty-five faculty members rate the 150 third-year students at the end of quarters 9 through 12 in our four academic years in a thirty-six-month curriculum. Students in the second year receive an average of fifteen ratings from various instructors who worked with them in the disciplines of oral diagnosis, endodontics, fixed prosthodontics, operative dentistry, and periodontics; each faculty member rates in his or her discipline. In addition, faculty members assess the clinical judgment and patient management skills of each student they rate. The same pattern is followed in the third year, except that there are a few more faculty members rating each student and the disciplines rated include endodontics, fixed prosthodontics, operative dentistry, periodontics, and removable prosthodontics. Since faculty members rate only those students they have worked with often enough to form a defensible opinion and since new faculty members are
added to the clinical faculty from time to time, students receive ratings from some of the same instructors over several quarters and from new instructors from time to time.

The datasets for competency ratings for the classes that graduated in 2003 and 2004 were included in this study. Faculty records and clinical evaluation forms were reviewed for each quarter to classify a faculty member in one of three categories: 1) faculty member who had rated the student in the previous quarter, 2) faculty member who was rating the student for the first time, or 3) faculty member who had an inconsistent pattern of rating in some previous quarters and not others. Data from the latter category of faculty members were excluded from the study. Of special interest was the set of ratings made by faculty members who were new to the clinic during students’ twelfth and final quarter. These represented ratings of competency immediately prior to graduation made by faculty members who were rating these students for the first time without having formed prior opinions.

In each of quarters 6 through 12 there were four sets of ratings: 1) new evaluators for the quarter in question, 2) raters who had evaluated the students the previous quarter (or more) for the quarter in question, 3) repeating raters for the previous quarter, and 4) new raters in quarter 12. These relationships are shown schematically in Figure 1. PR|N is the partial correlation between previous (P) and repeat (R) ratings with new (N) ratings factored out. There were six such datasets, three each for the two classes. The three different sets per class were ratings of clinical judgment, patient management, and technical skill within disciplines. All scores were normalized to reduce scale variance across quarters and differences caused by non-random sampling of student-evaluator combinations. Preliminary ANOVA of Fisher z-transformed data revealed no differences across years or across the three rating categories. The dataset was combined across years.

The primary unit of analysis used to measure the carryover effect of stereotyping is correlation. In particular, we are interested in the correlation between Time₁ and Time₂ for faculty who rated students in subsequent quarters and the correlation between these faculty members at Time₁ and new faculty members at Time₂. Also of interest is the correlation between Time₁, from new and repeat evaluators and the new ratings in quarter 12. Statistical tests were performed for differences in rated correlations, partial correlations (Time₁ and Time₂ for repeating evaluators with the scores of new evaluators held constant), and sequential multiple regression analysis with new evaluator scores forced into the equation at the first step.

Results

Two hundred fifty-eight students had sufficient data sets in Q06 for analysis, and this increased to 261 in Q07, 297 in Q08, and 298 thereafter. As shown in Table 1, the average number of faculty members making previous quarter ratings rose from about four to about twelve during the two years of students’ clinical education. During the same period, repeating evaluator ratings increased from about 5 to about 12, and new faculty ratings averaged about 3, with a spike to 5 at the transition from second to third year.

The correlations between Time₁ and Time₂ ratings (PR) increased in each subsequent quarter, reaching a high of about r = .750 for the transition between Q11 and Q12. The initial carryover from Q05 to Q06 was about r = .300 for ratings of discipline competency (ratings of skill in oral diagnosis, endodontics, fixed prosthodontics, operative dentistry, periodontics, and removable prosthodontics) and for clinical judgment. The carryover for ratings of patient management began slightly higher at r = .416.

By contrast, carryover correlations involving different sets of evaluators varied across quarters, but neither increased nor decreased. These two correlations are shown in Table 1 as PN (previous quarter ratings from faculty members who also rated in the current quarter versus evaluators new in the current quarter) and RN (ratings during the current quarter).
Table 1. Correlation coefficients for combinations of ratings of clinical competency

<table>
<thead>
<tr>
<th>Qtr</th>
<th>P</th>
<th>R</th>
<th>N</th>
<th>PR</th>
<th>PN</th>
<th>RN</th>
<th>PN-RN</th>
<th>PR</th>
<th>PN</th>
<th>RN</th>
<th>PN-RN</th>
<th>PR</th>
<th>PN</th>
<th>RN</th>
<th>PN-RN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q06</td>
<td>3.73</td>
<td>5.31</td>
<td>3.34</td>
<td>.2894</td>
<td>.2891</td>
<td>.2599</td>
<td>.3043</td>
<td>.1606</td>
<td>.3208</td>
<td>.4156</td>
<td>.2796</td>
<td>.3650</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q07</td>
<td>5.18</td>
<td>7.45</td>
<td>2.60</td>
<td>.2771</td>
<td>.3013</td>
<td>.4772</td>
<td>.2865</td>
<td>.2898</td>
<td>.5519</td>
<td>.3197</td>
<td>.2503</td>
<td>.4972</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q08</td>
<td>6.09</td>
<td>8.44</td>
<td>2.35</td>
<td>.4370</td>
<td>.2026</td>
<td>.2389</td>
<td>.05</td>
<td>.5678</td>
<td>.3213</td>
<td>.3757</td>
<td>.05</td>
<td>.4855</td>
<td>.2810</td>
<td>.2917</td>
<td>.05</td>
</tr>
<tr>
<td>Q09</td>
<td>6.48</td>
<td>6.07</td>
<td>4.99</td>
<td>.4376</td>
<td>.3137</td>
<td>.3716</td>
<td>.5989</td>
<td>.2865</td>
<td>.3645</td>
<td>.001</td>
<td>.4985</td>
<td>.2655</td>
<td>.4161</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>Q10</td>
<td>8.80</td>
<td>11.43</td>
<td>3.43</td>
<td>.5276</td>
<td>.2499</td>
<td>.4625</td>
<td>.05</td>
<td>.5141</td>
<td>.3439</td>
<td>.4633</td>
<td>.01</td>
<td>.5358</td>
<td>.3523</td>
<td>.4729</td>
<td>.05</td>
</tr>
<tr>
<td>Q11</td>
<td>12.53</td>
<td>12.23</td>
<td>1.49</td>
<td>.6101</td>
<td>.2869</td>
<td>.3762</td>
<td>.001</td>
<td>.5755</td>
<td>.3245</td>
<td>.3927</td>
<td>.001</td>
<td>.6102</td>
<td>.2735</td>
<td>.3991</td>
<td>.001</td>
</tr>
<tr>
<td>Q12</td>
<td>11.70</td>
<td>12.54</td>
<td>2.99</td>
<td>.7756</td>
<td>.3183</td>
<td>.3866</td>
<td>.001</td>
<td>.7678</td>
<td>.3922</td>
<td>.4344</td>
<td>.001</td>
<td>.7273</td>
<td>.3020</td>
<td>.3040</td>
<td>.001</td>
</tr>
</tbody>
</table>

P = Ratings from previous quarter, R = Ratings in current quarter from faculty who rated students in previous quarter (“repeat-
erers”), N = Ratings in current quarter from faculty members who are new this quarter; PR is correlation between previous and
current quarters for repeating evaluators (same evaluators), PN is correlation between previous quarter’s ratings and new
ratings in current quarter (different evaluators), RN is correlations between repeating and new raters in current quarter
different evaluators); PN-RN is p-value for test of differences between correlations sharing a common variable.

These patterns are graphed in Figure 2 for the case of ratings of discipline competency.

The column PN-RN in Table 1 shows the p-
values for the statistical test that the correlation co-
efficient for PN is different from the coefficient for
RN. This is a test for differences in correlations with
a common variable. Generally, this test reached sig-
nificance by the second year of clinic, indicating that
carryover was significant: faculty members who
made repeated ratings used information not available
to faculty members who were rating these stu-
dents for the first time.

Two approaches were taken to characterizing
the magnitude of the carryover on rating informa-
tion. Partial correlation is a method of determining a

Figure 2. Zero-order correlations in ratings of discipline competency—discipline competency
correlation coefficient by “holding constant” a common, measured association with other variables. In this case the PR correlation for repeated raters across subsequent quarters was calculated partialing out the common associations for new raters, PR|N. This can be interpreted as a measure of carryover of rating information across quarters adjusting for information that only new raters were picking up.

Correlation for repeated ratings across quarters, partialing out information identified by new raters, is shown in Table 2 and graphed for ratings of clinical judgment in Figure 3. The data show the same general patterns of increasing r-values across the two years of clinical instruction that were apparent in the unadjusted correlations for repeated raters. The partial correlations are always lower than the zero-order correlations, reflecting the fact that repeating raters used both information carried over from the previous quarter and new information that was also identified by the new raters.

Another way to measure the carryover effect takes advantage of multiple regression techniques. It is possible to compare the difference between the $R^2$ value using both ratings from evaluators in previ-

<table>
<thead>
<tr>
<th>Qtr</th>
<th>Discipline Competency Partial Variance</th>
<th>Clinical Judgment Competency Partial Variance</th>
<th>Patient Management Competency Partial Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q06</td>
<td>.2318 .0501</td>
<td>.2704 .0656</td>
<td>.3508 .1066</td>
</tr>
<tr>
<td>Q07</td>
<td>.1591 .0196</td>
<td>.1587 .0176</td>
<td>.2324 .0406</td>
</tr>
<tr>
<td>Q08</td>
<td>.4087 .1575</td>
<td>.5094 .2229</td>
<td>.4395 .1766</td>
</tr>
<tr>
<td>Q09</td>
<td>.3642 .1381</td>
<td>.5536 .2678</td>
<td>.4426 .1620</td>
</tr>
<tr>
<td>Q10</td>
<td>.4437 .1548</td>
<td>.4264 .1428</td>
<td>.4476 .1556</td>
</tr>
<tr>
<td>Q11</td>
<td>.5658 .2748</td>
<td>.5151 .2245</td>
<td>.5681 .2713</td>
</tr>
<tr>
<td>Q12</td>
<td>.7463 .4737</td>
<td>.7211 .4184</td>
<td>.6998 .4444</td>
</tr>
</tbody>
</table>

Figure 3. Partial correlation and forced-order regression analyses of ratings of clinical judgment competency
ous quarters and new evaluators to predict current quarter ratings with the $R^2$ value for new ratings only. The difference between the two values is a measure of carryover: previous ratings minus new information. Table 2 shows these values for ratings of all three types of competency, and Figure 3 graphs these results for the clinical judgment competency. In general, the proportion of variance in repeated ratings was seen to increase from about 5 percent to 15 percent over the first year and a half of clinical ratings and then to jump to about 25 percent in Q11 and then 45 percent in Q12. Although faculty members who make repeated ratings always rely more heavily on new information than carryover information when making subsequent ratings, the pattern of student competency has begun to be established in the last half of the final clinical year.

The third hypothesis addressed in this research concerns the quality of the ratings: are competency ratings based on multiple quarters of experience better predictors of graduation competency ratings than are ratings made by fresh sets of faculty members in each quarter? In Table 3 the correlations are shown between ratings made by faculty members who have rated students in previous quarters with the graduation quarter ratings and the correlations between faculty members who are rating students for the first time with graduation quarter ratings. In order to re-

| Table 3. Correlations between ratings of repeating and new raters with ratings from new raters in graduation quarter |
|---|---|---|---|---|---|---|---|
| Qtr | Discipline Competency | Clinical Judgment Competency | Patient Management Competency |
|     | R-Q12   | N-Q12   | R-Q12   | N-Q12   | R-Q12   | N-Q12   |
| Q06 | .2597   | -.0118  | .4113   | .0772   | .2650   | .2215   |
| Q07 | .2829   | .2735   | .2688   | .1651   | .2282   | .1586   |
| Q08 | .1888   | .1444   | .3749   | .2949   | .3848   | .2412   |
| Q09 | .2864   | .2820   | .3765   | .2265   | .3346   | .1651   |
| Q10 | .3039   | .1882   | .2987   | .2002   | .2532   | .0353   |
| Q11 | .2986   | .2565   | .3619   | .3022   | .2958   | .1740   |

![Figure 4. Prediction of graduation quarter competency by repeat and new evaluators—patient management competency](image-url)
duce bias in this comparison, the Q12 ratings were all made by faculty members who had not rated students previously.

Ratings made by faculty members who had the advantage of ratings over more than one quarter are statistically better predictors of graduation competency level (p<.001). The pattern for ratings of the patient management competency is shown in Figure 4.

Discussion

By comparing two sets of ratings for the same group of students, it is possible to get some insight into the way faculty members combine new and previous information about students’ clinical competency. In particular, this study compared the ratings on three clinical competencies given by faculty members who had rated the students in the previous quarter with ratings given to the same students by faculty members who were rating these students for the first time.

The data show that faculty members who are making repeated ratings use both new information and information carried over from previous quarters, and although the carryover information increases as part of the mix, it never achieves a dominant role. By contrast, a succession of new faculty member evaluators seems to rate independently of information available from previous quarters. There is no support in this study for the “hall talk” hypothesis that students’ reputations precede them and influence their ratings. The study also shows that raters who have both current and carryover information are more accurate in their predictions of students’ graduating quarter competence than are faculty member evaluators who have only current quarter contact. The data confirm all three hypotheses: 1) faculty members carry information over from period to period when evaluating student competence, 2) this carryover is cumulative, and 3) carryover information increases rating predictability.

This study found no clear evidence for a stereotyping effect in repeated ratings of clinical competence. Although faculty members carry information forward and deepen their impressions of student competence, there was no evidence that such carryover information systematically dampens or distorts the use of new information. The data show a convergence between repeated ratings and ratings from new faculty members that is attributable substantially to both sets of raters observing students’ current performance. Further, faculty members making repeated ratings were more accurate in predicting the ultimate, graduation quarter competency of students as rated by new evaluators.

Ratings are generally preferred as an evaluation method where complex performance must be assessed in nonstandardized circumstances. Objectivity is sacrificed (professionally competent raters must be used), but reliability and validity are not. The literature is as filled with failed attempts to validate objective checklists as it is with other types of psychometric reports. Ratings have the advantage of allowing for adjustments for a range of circumstances and permitting raters to nuance the weighting of various performances. For example, a system of averaging objective single-encounter scores might overweight a trivial slip that was unanticipated and quickly recognized and corrected by the student. It might underweight a gross, negligent, and costly error. Rating systems permit evaluators to review the repertoire of experiences and judge their overall significance. Raters who have a richer range of exposures to students, including those extending back several quarters, could reasonably be expected to provide more accurate ratings, although the literature generally reports a tendency to give lenient (unjustifiably positive) evaluations in rating systems.

It may be impossible to design a study that perfectly isolates and measures potential stereotyping effects. This study does not claim to have done so. Such a project would require measurements that incorporate all that faculty members know, but require that this knowledge be revealed all at the same time. Some counterhypotheses that might partially explain the results reported in this study include the following: 1) students perform more dentistry in later quarters, thus exposing more of their true competency for evaluation later in the program; 2) new faculty members may be inexperienced at ratings generally; and 3) students may benefit from feedback given by faculty in previous quarters and thus improve performance in line with repeating faculty members’ expectations.

REFERENCES