● STUDENT AND RESIDENT EVALUATIONS OF FACULTY—HOW DEPENDABLE ARE THEY?

Moderator: John Littlefield, PhD

The Dependability of Students' Ratings of Preceptors

KATHLEEN MAZOR, BRIAN CLAUSER, ANDREW COHEN, E. ALPER, and MICHELE PUGNAIRE

Time spent under the direction of preceptors during the clinical years is an important component of medical students' clinical training. Clerkship directors and other medical school administrators are concerned that the preceptors who work with their students provide high-quality teaching. However, time constraints and the geographic distribution of preceptors make it difficult if not impossible for course directors to directly monitor each preceptor and the quality of the teaching provided. One means of gathering information about teaching quality is to have students rate their preceptors' teaching skills. However, as with any measurement procedure, if such ratings are to be used for decision making, the dependability of the ratings must be assessed. This paper describes an application of generalizability theory to students' ratings of preceptors, with recommendations for the interpretation and reporting of such ratings.

Generalizability theory[1,2] provides a framework for estimating the relative magnitudes of various sources of error in a set of scores. Appropriate estimation of error is necessary for the calculation of appropriate confidence intervals, which in turn facilitate correct interpretation of scores. In addition, estimation of the relative contributions of multiple sources of error can suggest changes in measurement procedures that would increase the efficiency and precision of measurement.

Analyses based on generalizability theory are generally conducted in two stages. The first stage is the generalizability study (G-study), which partitions the variance associated with a set of scores in a manner similar to analyses of variance. Depending on the study design, different variance components can be estimated. The second stage, the decision study (D-study), estimates the impacts of changing various components of the measurement procedure. Generalizability coefficients (g) and phi coefficients, comparable to reliability coefficients in classical test theory, are calculated for different measurement conditions. Using the results from a particular implementation of a questionnaire, it is possible to estimate the impact of changing the number of respondents or the number of items. Confidence intervals can also be calculated, again allowing for different measurement conditions.

While there are many reports in the literature of efforts to use rating scales to assess physicians', residents', or medical students' performances, there is considerably less literature on the psychometric quality of such assessments. One exception is an article by Woolliscroft et al.[3] These authors collected ratings of residents' humanistic qualities from patients, nurses, attending physicians, and program supervisors. They conducted generalizability analyses of these ratings and concluded that for appropriately generalizable results, ratings from more than 50 patients would be needed for each resident. Approximately 20 to 50 attending physicians, or 10 to 20 nurses would be required to achieve this level of reliability.

Also noteworthy is a study by Kreiter and colleagues,[4] who assessed the generalizability of preceptors' evaluations of students. Using generalizability theory to estimate variance components, generalizability coefficients, and standard errors of measurement for various numbers of raters and items, these authors concluded that using three or more preceptors to rate a student's performance resulted in acceptable reliability. They report generalizability coefficients between approximately .20 to .50 for three raters and 20 items, depending on the clerkship.

The current study describes an application of generalizability theory to students' ratings of preceptors' teaching skills. Variance components and g and phi coefficients for different conditions and confidence intervals are estimated, and recommendations for practice are made.

Method

At the University of Massachusetts Medical School (UMMS), all students participate in six clerkships during their third year and two required clerkships during their fourth year. The third-year clerkships are medicine, surgery, pediatrics, psychiatry, family medicine, and ob–gyn. The fourth-year clerkships are the medicine subinternship and neurology. At the end of each clerkship, each student is asked to rate the preceptors with whom he or she worked during that clerkship. Depending on the duration and the structure of the clerkship, students are asked to rate between two and seven different preceptors. All items analyzed here were on a four-point scale, with 4 being the most positive rating. Nine items were chosen as the focus of this study. These items addressed perceived interest in teaching, availability, attention to student-specific learning needs, respect for students' opinions, use of constructive feedback, and effectiveness in teaching physical exam skills, interviewing skills, patient management, and general medical knowledge.

Data. During the 1997–98 academic year, a total of 213 students completed third- or fourth-year clerkships. End-of-clerkship evaluation forms were distributed at the end of each clerkship to all students who had completed that clerkship. A total of 790 clerkship evaluation forms were distributed, and 766 were completed and returned, for an overall response rate of 97%. All evaluation forms contained blocks of items related to preceptor teaching skills. While students were asked to evaluate between two and seven preceptors, not all students completed the maximum number of preceptor evaluations. The lowest number of preceptor evaluations completed by a single student was two, the highest number was 28. Deletion of cases with more than one missing value resulted in a total of 2,282 preceptor ratings by 213 students. If only one of the nine items was missing, the missing value was replaced with the mode of the other eight values. All preceptors with three or more ratings within a specific clerkship were selected. For preceptors who had been rated by more than three students, three sets of ratings were randomly selected. The final file consisted of 254 preceptors, each rated on nine items by three students.

Analysis. A G-study was conducted to obtain estimates of variance components for the objects of measurement (preceptors), and for the study facets. The design used here, raters within preceptors by items (R:P × I), allowed the variance to be partitioned in the following manner: preceptor, rater nested within preceptor, item, preceptor by item interaction, and residual. This partitioning of the variance is comparable to the partitioning of variance in analysis of variance (ANOVA). However, while in an ANOVA the F statistic is typically of primary importance, in a G-study the focus is on the relative magnitudes of the variance estimates. When a greater proportion of the variance in ratings is attributable to variation between the objects of measurement (preceptors) compared with the variance associated with the various sources of error (potentially raters, items, and residual), generalizability and phi coef-

ficients will be higher, and ratings can be considered more dependable.

A D-study (using the same design) was conducted to estimate the impacts of varying the numbers of raters and items. In a D-study, this is done by choosing different levels of the facet of interest (for instance, setting the number of raters equal to 1, 5, 10, or 20) and then recomputing the relevant coefficients at each level. This is comparable to using the Spearman–Brown prophecy formula to estimate the impact of increasing the number of test items. Generalizability analyses were conducted using the computer program GENOVA.[4] Order of clerkship and timing of clerkship were ignored.

Results

Overall, students' ratings of preceptors were positive, with item means ranging from 3.08 to 3.58 for the full data set. The means for the subset of ratings used in the analysis (254 preceptors rated by three students each) closely approximated the means for the full sample; the difference between means was not more than .06 for any item. The overall item mean was 3.36 (SD = .58) for the full sample; 3.39 (SD = .54) for the subset used in the analysis.

G-study results are presented in Table 1. The largest percentage of variance was associated with the rater nested within preceptor (R:P) effect, which includes both the rater main effect and the rater-by-preceptor interaction. This implies either differences between students in how they used the scale—that is, some students tending to be more lenient or severe in their ratings than others (a rater main effect)—or differences in how students rank ordered preceptors (a rater-by-preceptor interaction). In either case, the relatively high percentage of variance associated with this effect suggests that preceptor scores based on a small number of ratings are not dependable—they are likely to vary considerably depending on the particular set of students who provide ratings. Obtaining another set of ratings from a different group of students could easily produce a substantially different result. In contrast, relatively little variability is associated with the item facet—students tended to rate a given preceptor the same on all or most items.

It should be noted that in this analysis item was treated as a random variable. While it is possible to argue that item could be treated as a fixed facet, the relatively small percentage of variance associated with the item facet and item-by-preceptor interaction suggests that treating item as a fixed facet would only modestly change the conclusions presented here.

D-study results, specifically the impacts of varying the numbers of raters and items, are summarized in Figure 1. Increasing the number of raters had a greater impact than increasing the number of items with respect to improving generalizability. For the current nine-item set, between 15 and 20 raters were needed to achieve a g coefficient of .80 or greater. Phi coefficients followed the same pattern, but were slightly lower. For instance, for 20 raters rating nine items, g = .83 and phi = .80. Generalizability coefficients are most appropriate when the purpose of measurement is to determine relative standing; phi coefficients are most appropriate when the purpose is to determine absolute level of performance. Increasing the number of items from one to five did improve the generalizability of scores, but increasing that number further from five to ten or more had much less of an impact.

Table 1. Variance Components Estimates, University of Massachusetts Medical School, 1997–8*

Effect	Single Observations		Mean Scores	
	Variance	% of Total Variance	Variance	% of Total Variance
Preceptor (P)	.06562	14	.06562	45
Rater:Preceptor (R:P)	.20771	44	.06924	47
Item (I)	.02521	5	.00280	2
PI	.02207	5	.00245	2
Residual	.14764	32	.00547	4

* Third- and fourth-year students evaluated clerkship preceptors on nine items related to teaching skills, using a four-point scale. The generalizability analysis was conducted on ratings of 254 preceptors, each of whom was rated by three students.

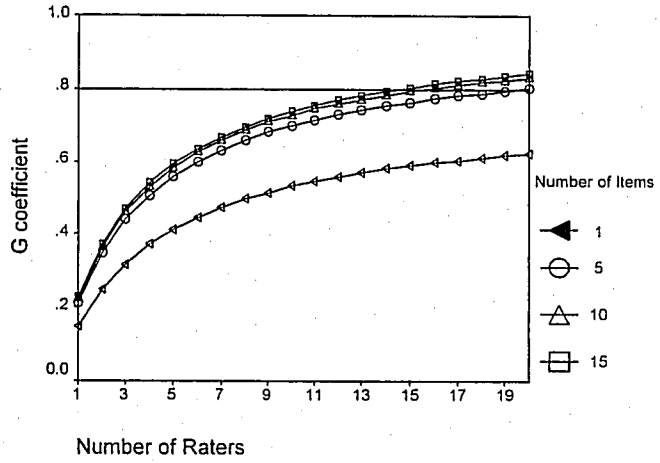

Figure 1. Estimated generalizability coefficients based on estimated variance components derived from ratings of 254 preceptors, each rated by three raters on nine items. The G coefficients are estimated for one to 15 items, and one to 20 raters.

Confidence intervals were estimated for various numbers of raters and items. Holding the number of items constant at ten, estimates of the width of 95% confidence intervals (for relative decisions) increased from .46 to .62 to .84 as the number of raters decreased from 20 to 10 to 5.

Discussion

While it is relatively simple to collect students' ratings of preceptors' teaching skills, these results suggest the mean rating for a given preceptor depends to a large extent on the particular set of students who provided ratings, unless a relatively large number of ratings are available. For the set of nine items studied here, ratings from 15 to 20 students would be required to obtain reasonably dependable estimates of preceptor scores. For the 1997–98 year, the mean number of ratings per preceptor was 2.6; the mode was 1. If ratings from only one student are available, then the width of the confidence interval for that preceptor's mean score is 1.86. Given that the scale ranges from 1 to 4, and that 84% of all preceptors' mean scores were higher than 3, a confidence interval this wide means differences between preceptors are essentially undetectable based on ratings from one student per preceptor. This improves when three raters are used, but still the resulting confidence interval is 1.08 units wide. This means that for a mean score of 3.4, the 95% confidence interval would be 2.9 to 3.9.

One of the important advantages of the generalizability approach is that it allows for assessment of multiple sources of error, compared with such methods as Cronbach's alpha, which estimate only one source of error at a time. Use of alpha is based on the assumption that the item scores are experimentally independent, an assumption that is not tenable when the same rater rates a given examinee on multiple items. The present data set provides a clear example of how violating this assumption may lead to an inflated estimate of reliability. The three-rater, 254-preceptor data set used here yields an alpha of .94, which in turn yields a 95% confidence interval of width .52. Thus, a preceptor with a score of 3.4 would have a

confidence interval of 3.1 to 3.7. This method of calculating confidence intervals overestimates the precision of the score.

It is also important to note that the findings reported here are consistent with previous research, such as the study by Kreiter et al.[3] While Kreiter et al. were investigating preceptors' ratings of students, as opposed to students' ratings of preceptors, the design of the G-study was the same, and the relative contributions of the various effects are very close to those reported here. Kreiter et al. also found that the number of raters had a much greater impact on generalizability than did the number of items. The percentages associated with the corresponding facets were very similar to those found here; for instance, they report 48% of the variance was associated with the rater:person effect, and 4% with the item main effect, compared with 44% and 5%, respectively, for the same components in this study.

Implications

If ratings of preceptors are to be used in decision making, it is essential to assess the dependability of the ratings. When reports contain numbers and graphs, readers may assume a degree of precision and reliability that is not in fact justified. Confidence intervals based on generalizability theory should remind readers of the error present in a set of scores, and encourage appropriate interpretation and use of results. One advantage of G-theory is that confidence intervals can be adjusted based on the number of raters or items used. This is important because in practice the number of ratings available may vary considerably.

Preceptors' scores based on only a few student raters must be interpreted with caution, as the actual scores may vary considerably depending on the particular set of students providing ratings. While students' ratings may be useful for red-flagging preceptors whose ratings are substantially below the mean, decisions about individual preceptors based on such ratings may be difficult to defend, unless large numbers of ratings (for instance, 15–20) are available. With respect to the number of items used to assess preceptors, adding more items will probably do little to improve the dependability of the measurement.

Correspondence: Kathleen Mazor, Office of Medical Education, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, MA 01655; e-mail: ⟨kathleen.mazor@umassmed.edu⟩.

References

1. Brennan RL. Elements of Generalizability Theory. Iowa City, IA: ACT Publications, 1992.
2. Shavelson RJ, Webb NM. Generalizability Theory: A Primer. Newbury Park, CA: Sage, 1991.
3. Woolliscroft JO, Howell JD, Patel BP, Swanson DB. Resident–patient interactions: the humanistic qualities of internal medicine residents assessed by patients, attending physicians, program supervisors, and nurses. Acad Med. 1994;69:216–24.
4. Kreiter CD, Ferguson K, Lee W-C, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. Acad Med. 1998;73:1294–8.
5. Crick JE, Brennan RL. Manual for GENOVA: A Generalized Analysis of Variance System. Iowa City, IA, 1993.