

Long-term Stability of Tutor Performance

Diana H. J. M. Dolmans, PhD, Ineke H. A. P. Wolfhagen, PhD, and Cees P. M. van der Vleuten, PhD

ABSTRACT

Purpose. The aim of this study was to investigate to what extent ratings of tutor performance remain stable in the long term. At many schools, teaching performance is assessed and these evaluations are consulted as part of the decision-making process for promotion, tenure, and salary. Since this information may have summative value, it is crucial that the reliability of the data be assessed. A previous study had shown that a single evaluation of a tutor is reliable when the responses of six students are used (interrater reliability). The present study focused on the stability of tutor evaluations over repeated occasions of evaluation.

Method. A generalizability study was conducted to estimate the number of occasions required to demonstrate stability. The study took place during three academic years (1992–93, 1993–94, and 1994–95) at the problem-based medical school of the University of Limburg (now Maastricht University). A total of 291 ratings were

analyzed (97 tutors rated during three sequential tutoring occasions). Two types of scores were used: an aggregate score calculated from ratings of 13 items and an overall judgment.

Results. The results indicate that when the scores are used to interpret the precision of individual scores, two evaluation occasions should be available for the overall judgment and four occasions for the aggregate score. If the tutor scores are consulted only to determine whether performances are above or below a cutoff score, a reliable decision can be made after only a single occasion of evaluation.

Conclusion. The results demonstrate that data collected over an extended period of time can be reliably used as part of the decision-making process for promotion, salary, and tenure.

Acad. Med. 1996;71:1344–1347.

Within the framework of improving educational quality, optimizing teachers' effectiveness has become a topical subject at many schools. Although teaching and research are both basic missions of a university, salary, promotion, and prestige are based almost exclusively on research productivity and not on teaching performance.¹ One of the incentives to improve teaching behavior is to include teaching performance in salary, promotion, and tenure decision.^{2,3} This

requires that teaching performance be assessed and be a part of the reward system, in which recurring poor performance has negative consequences and good performance may conversely have positive consequences.

Since teaching is a multifaceted activity, a fair and balanced appraisal of the performance of a teacher is obtained by using multiple sources of information. The information collected can be consulted as part of the decision-making process for promotion and tenure. If a teacher's performance is below standard on a single evaluation, this information should be interpreted cautiously and should not have negative consequences, or perhaps not immediately. External circumstances or extraneous factors might prevent a teacher from receiving good scores during one period of teaching.

Consistently low scores on repeated occasions over extended periods of time, on the other hand, should have negative consequences in the decision-making process for promotion and tenure. This leads to the question of how many occasions should be available for the scores to be taken seriously. In other words, to what extent are ratings of teaching performance stable over extended periods of time or in the long term?

At the problem-based medical school of the University of Limburg (now Maastricht University), all teachers are evaluated by students at the ends of their periods of tutoring. Several studies indicate that student ratings provide reliable and valid information about indicators of effective teaching.^{4,5} Since tutors' scores at this medical school are used for promotion, tenure, and salary decision-

Dr. Dolmans is educational psychologist, Dr. Wolfhagen is educational psychologist, and Dr. van der Vleuten is a professor of education; all with the Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands.

Correspondence should be addressed to Dr. Dolmans, Maastricht University, Department of Educational Development and Research, PO Box 616, 6200 MD Maastricht, The Netherlands. e-mail: <dolmans@educ.unimaas.nl>. Reprints are not available.

making purposes, it is crucial that the reliability of the data be assessed. Research on the reliability of the data can be considered in terms of the reproducibility of a single evaluation (interrater reliability) or the reproducibility, or stability, over repeated occasions (test-retest reliability). A previous study showed that a tutor score over one course period is reliable when the responses of six students are used, a number that is practicable with the current group size of nine to ten students.⁶ No study has, however, been conducted to determine how many occasions are needed over an extended period of time to make serious and fair academic decisions about an individual teacher's tutoring performance. The main aim of the present study was to investigate how many occasions are needed to obtain a fair appraisal of a teacher's tutoring performance in the decision-making process for promotion and tenure.

METHOD

Subjects. This study was conducted in a problem-based medical school. Tutor ratings of 140 tutors were available, amassed over a period of three academic years (1992-93, 1993-94, and 1994-95). The numbers of occasions per teacher varied between three and 12, with an average of 5.5.

Instrument. The tutor rating scale is completed by students. It contains 13 items: six items related to the tutor's task to guide students through the learning process, four items about the tutor's content-knowledge input, and three items about the tutor's commitment to the group's learning.⁷ Students are asked to rate their tutor's demonstration of the behavior described in each item as (1) insufficient, (2) neutral, or (3) sufficient. The feedback sent to individual tutors contains their students' ratings on individual items, expressed as the percentages of students who rated the behavior described in each statement as insufficient, neutral,

and sufficient. This is to facilitate interpretation. These percentages per item are averaged across the 13 items, resulting in total average percentages of insufficient, neutral, and sufficient for each tutor. Subsequently, the average percentage sufficient minus insufficient is computed for each tutor. This *aggregated score* varies between -100% and 100%. In addition, students are asked to give an *overall judgment* (ranging from 1 to 10, 6 being the pass score) of the performance of the tutor.

Procedure. At the medical school under investigation, teachers generally guide two tutorial groups during one course period. Consequently, two dependent ratings are available over the same time period. Since the purpose of this study was to detect the stability of the performance of a tutor over an extended period of time, only one occasion in each of several course periods was included in the analysis. To ensure reliable data per occasion, average tutor ratings of groups consisting of at least six students were used.⁶ To achieve a fully balanced design convenient for statistical analysis, a sample of three sequential occasions per teacher collected within a time period of three academic years was selected. The time periods between the three selected occasions varied between teachers from several months to one year. Tutors rated fewer than three times were excluded from the analysis. In the data set used for final analysis, 97 (69%) of the 140 tutors met the inclusion criteria. A one-way analysis of variance, conducted to compare the group of tutors included with those excluded from the analysis, showed that the groups did not differ based on the aggregated scores ($F_{(1,479)} = 0.00, p = .95$) and overall judgments ($F_{(1,479)} = 0.05, p = .82$). For this study, 291 ratings (97 tutors \times 3 occasions) were available. The average aggregated score was 61.4% (SD = 28.0, range of -51 to 99). The average overall judgment (scale of 1-10) was 7.6 (SD = 0.9, range of 4 to 9).

Statistical analysis. Generalizability studies were conducted to estimate the reliabilities of the aggregated score and the overall judgment.^{8,9} One of the advantages of generalizability theory over classical test theory is that it recognizes multiple sources of error, such as differences among tutors and differences in occasions, instead of only a single undifferentiated error component.⁹ Generalizability theory is based upon analysis of variance.⁸ As all tutors were judged on three occasions, an all-random tutor-crossed-with-occasion design was used, with tutors as the universe of generalization, or object of measurement. This design allows variance-component estimation of three sources: (1) differences between tutors (object of measurement); (2) differences between occasions; and (3) tutor-by-occasion interactions and unidentified sources of variance in error.¹⁰

RESULTS

Table 1 summarizes the sources of variability and the corresponding estimated variance components. The percentages of variance associated with tutors for the aggregated score and the overall judgment are 46.5% and 39.7%, respectively. Approximately 47% of all variance in the aggregated score and approximately 40% of all variance in the overall judgment can be attributed to variation between tutors. This percentage is the true variance, or the variance of interest, and indicates that the instrument is well able to discriminate between tutors across occasions. The largest effects, however, are the tutor-by-occasion interaction effect and the error effect, 53.5% and 59.7%, respectively. The interaction effect indicates that the relative standings of tutors change from occasion to occasion, which is also considered to be error. In studies about testing, this error component is usually a larger source of variance.

The estimated variance components were used to estimate reliability indices as functions of the number of occasions.

Table 1

Sources of Variability and Estimated Variance Components for Student Ratings of Tutors over Three Occasions, University of Limburg Medical School*				
Source of Variability	Estimated Variance Component	Degrees of Freedom	Standard Error	Percentage of Total Variance
For aggregated score				
Differences between tutors	366.75	96	73.89	46.5
Differences between occasions	0.00	2	1.47	0.0
Tutor-occasion interaction and unidentified sources of variance in error	422.16	194	42.86	53.5
For overall judgment				
Differences between tutors	0.3348	96	0.0738	39.7
Differences between occasions	0.0055	2	0.0076	0.6
Tutor-occasion interaction and unidentified sources of variance in error	0.5031	192	0.0511	59.7

*During three academic years (1992-93, 1993-94, and 1994-95), a total of 291 ratings were analyzed (97 tutors rated during three sequential tutoring occasions). For each occasion evaluated, each tutor received two average scores: an aggregate score (scale of -100 to +100) calculated from ratings of 13 items and an overall judgment (scale of 1 to 10). See text for details.

Three indices were estimated. First, a reliability coefficient of the scores was estimated by using an absolute interpretation (a domain-referenced approach). This *dependability coefficient* includes mean differences as well as inconsistencies of rank-ordering across occasions in the error term. Second, the standard error of measurement (SEM) was estimated. The SEM is the standard deviation of the error scores on the original score scale and can be used to interpret the precision of individual scores by estimating a confidence interval around an observed score. Finally, Brennan has proposed a "reliability-like" index for the reproducibility of the decision being taken, given a particular cutoff score, called the adjusted phi coefficient.^{11, pages 108-9} In this mastery-referenced perspective, the interest lies not in the score but in whether the score has passed or failed a criterion. Below-standard performance is considered a score below zero on the aggregated score (scale of -100 to 100) and below 6 on the overall judgment (scale of 1-10, 6 being the pass score). These were rather arbitrarily chosen as cutoff scores.

Table 2 shows the dependability coefficients, the corresponding SEMs, and the adjusted phi coefficients as functions of the number of occasions for both the aggregated score and the overall judgment. The differences in reliability between both kinds of scores are only marginal but are slightly better for the aggregated score. This is not unusual, since this score is based on multiple and specific items. Assuming that the goal of collecting data about a tutor's performance is to draw inferences about the absolute score, at least five occasions are required for the aggregated score and six occasions for the overall judgment to obtain a minimal dependability coefficient of .80. The SEM, however, is more informative. The SEM can be used to estimate confidence intervals for individual tutor scores. For example, the 95% confidence interval of a score can be estimated by multiplying the SEM by 1.96.¹² Taking the (arbitrary) standpoint that a difference of at least 40 points on the aggregated score (scale of -100 to 100) is required for it to be interpreted reliably, the SEM should be

lower than or equal to 10.20 (20/1.96) at the level of 95%. For the overall judgment, a difference of at least 2 points (scale of 1-10) should allow reliable interpretation. Therefore, the SEM should be lower than or equal to 0.51 (1/1.96) at the level of 95%. When these criteria are used, about four occasions are required to obtain a reliable aggregated score and two occasions to obtain a reliable overall judgment. The ultimate goal of assessing a tutor's performance is to draw inferences about whether his or her performance is below a standard. The reliability of this decision is reflected in the adjusted phi coefficient. Although no standard exists for the acceptability of adjusted phi coefficients, both scores seem to have acceptable "decision reliability" after only a single occasion of evaluation.

CONCLUSION

The aim of this study was to investigate the long-term stability of tutor ratings given by students. If these ratings per tutor were to vary considerably from occasion to occasion, decision makers should exercise caution in using the data for salary, promotion, and tenure decisions. Hence, a generalizability study was conducted to estimate the number of occasions that should be available to obtain a reliable estimation of a teacher's tutoring performance. The results showed that, if the aggregated score and overall judgment are used to interpret the precision of individual scores, four and two occasions, respectively, should be available. If the tutor scores are consulted only to determine whether performances are below or above cutoff scores of zero (for the aggregated score) and six (for the overall judgment), a reliable decision can be made after only a single occasion of evaluation. As a consequence, decision makers may have a high degree of confidence in the data.

Valuing good teaching performances as positively as good research performances may probably contribute to im-

Table 2

No. of Occasions	Aggregated Score			Overall Judgment		
	Dependability Coefficient	Standard Error of Measurement	Adjusted Phi Coefficient	Dependability Coefficient	Standard Error of Measurement	Adjusted Phi Coefficient
1	.46	20.55	.91	.40	0.71	.85
2	.63	14.53	.95	.57	0.50	.92
3	.72	11.86	.97	.67	0.41	.95
4	.78	10.27	.98	.73	0.36	.96
5	.81	9.19	.98	.77	0.32	.97
6	.84	8.39	.98	.80	0.29	.97
7	.86	7.77	.99	.82	0.27	.98
8	.87	7.26	.99	.84	0.25	.98

*During three academic years (1992-93, 1993-94, and 1994-95), a total of 291 ratings were analyzed (97 tutors rated during three sequential tutoring occasions). For each occasion evaluated, each tutor received two average scores: an aggregate score (scale of -100 to +100) calculated from ratings of 13 items and an overall judgment (scale of 1 to 10). These scores were used to calculate estimated variance components (see Table 1), which in turn were used to estimate three reliability indices as functions of the number of occasions (shown above). See text for details.

provement of teaching behaviors. In the medical school of the University of Limburg, tutor scores are a part of the reward system. The overall judgment is used for tenure and promotion decisions. The central administration office keeps a record of these scores. In the reward system these scores have summative value. If a tutor has recurring insufficient scores, the chair of the department will be informed. A teacher who performs badly will be disqualified from the tutor role. This will cause problems within a department, because the educational load remains unchanged and other colleagues within the department are required to compensate for the disqualification. In addition, the teacher in question will be disqualified from particular coordinating educational roles, whereas good performance in these roles is one of the prerequisites for promotion to higher ranks. For example, to become an associate professor, a faculty member should have fulfilled several coordinating roles within the educational program.⁷

This study has demonstrated that data collected over an extended period of time can be reliably used as part of the

decision-making process for promotion and tenure. However, as mentioned in the introduction, teaching is a multifaceted activity and multiple sources of evaluation are needed for a balanced appraisal of a teacher's performances. At the medical school of the University of Limburg, the tutor role is only one of the various teaching roles that can be fulfilled in the problem-based curriculum. This implies that not only the tutor role but also other educational roles should be assessed. These evaluations should all be incorporated into a teaching portfolio, to be consulted when promotion or salary decisions are considered. If these decisions are indeed based upon measures of different roles, the numbers of occasions per role may actually be lower than the numbers reported in this article, since a composite score is probably more reliable than a score based on measures of a single role.

REFERENCES

1. Van der Vleuten C. Improving medical education. *BMJ*. 1993;306:284-5.
2. Braskamp LA, Ory JC. *Assessing Faculty Work: Enhancing Individual and Institutional Performance*. San Francisco, CA: Jossey-Bass, 1994.
3. Lovejoy FH, Clark MB. A promotion ladder for teachers at Harvard Medical School: experience and challenges. *Acad Med*. 1995;70:1079-86.
4. Marsh HW. Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. *J Educ Psychol*. 1984; 5:707-54.
5. Abrami PC, d'Appolonia S, Cohen PA. Validity of student ratings of instruction: what we know and what we do not. *J Educ Psychol*. 1990; 82:219-31.
6. Dolmans DHJM, Wolfhagen HAP, Schmidt HG, Vleuten van der CPM. A rating scale for tutor evaluation in a problem-based curriculum: validity and reliability. *Med Educ*. 1994;28:550-8.
7. Dolmans DHJM, Wolfhagen HAP, Snellen-Balendong HAM. Improving the effectiveness of tutors in problem-based learning. *Med Teacher*. 1994;4:359-67.
8. Crick JE, Brennan RL. *Manual for Genova: A Generalized Analysis of Variance System*. Iowa City, IA: American College Testing Program, 1983.
9. Brennan RL, Kane MT. Generalizability theory: a review. *New Directions for Testing and Measurement*. 1979;4:33-51.
10. Shavelson RJ, Webb NM. *Generalizability Theory: A Primer*. London: Sage, 1991.
11. Brennan RL. *Elements of Generalizability Theory*. Iowa City, IA: American College Testing Program, 1983.
12. Ferguson GA. *Statistical Analysis in Psychology and Education*. 5th ed. Auckland, New Zealand: McGraw-Hill, 1981.