Effect of Using Multiple Standardized Patients to Rate Interpersonal and Communication Skills on Intercase Reliability

VU, Nu Viet


DOI: 10.1080/10401339409539642

Available at:
http://archive-ouverte.unige.ch/unige:26623

Disclaimer: layout of this document may differ from the published version.
The Effect of Using Multiple Standardized Patients
to Rate Interpersonal and Communication Skills
on Intercase Reliability

Jerry A. Colliver, Michelle L. Marcy, Nu Viet Vu,
David E. Steward, and Randall S. Robbs

Southern Illinois University School of Medicine
Springfield, Illinois
ABSTRACT

Performance-based examinations that use standardized-patient (SP) cases have been used to assess the clinical competence of senior medical students in seven classes (1986 - 1992) at Southern Illinois University School of Medicine. Because of the length of the examinations, two or more (multiple) SPs have been used to simulate about half of the cases in each examination. An earlier study found little or no effect of using multiple SPs on the intercase reliability of checklist scores, written scores, and total scores. The present study was conducted to assess the effect of multiple SPs on the intercase reliability of SP ratings of interpersonal and communication skills. Although the ratings were not used to determine total scores or to make pass-fail decisions, it seems reasonable to expect that multiple SPs might have a greater impact on ratings on non-cognitive dimensions than on ostensibly more objective checklist and written scores. In general, the results showed that generalizability coefficients free of measurement error due to multiple SPs were larger than those reflecting measurement error due to multiple SPs by an average of about 5 percentage points, suggesting that the use of multiple SP raters on the same case reduced reliability by about 5 percentage points. Compared to earlier findings, the effect of multiple SPs on the reliability of ratings was about 4 or 5 times greater than their effect on checklist, written, or total scores. Nevertheless, the magnitude of the effect on ratings was not excessive.
Performance-based examinations that use standardized-patient (SP) cases are used widely to assess the clinical competence of medical students, residents, and physicians (Hart, in press). The appeal of the examinations is that they assess the very skills and competencies that are expected of physicians in the practice of clinical medicine. At Southern Illinois University School of Medicine, a Post-Clerkship Examination that uses SP cases has been administered to seven classes (1986-1992) of senior medical students (Vu et al., 1992). Students are expected to pass the examination to fulfill a part of their graduation requirements. The examinations have consisted typically of about 18 SP cases, requiring three days of testing per student and three weeks for an entire class of about 70 students. Because of the time required to administer the examination and in anticipation of scheduling and other logistic problems (e.g., bad weather or illness), two or more (multiple) SPs have been trained to simulate each case; and for each examination administered to date, about half of the cases have, in fact, required simulation by multiple SPs.

Naturally, questions have been raised about the consistency of the case portrayal and checklist completion by two or more different simulators/recorders on the same case, that is, specifically, about the impact of using multiple SPs on reliability. Fortunately, however, studies have shown little or no effect of multiple SPs on reliability (van der Vleuten & Swanson, 1990). In our three studies of this question, for example, analyses were performed on total examination scores, checklist scores, and postencounter written scores for the five classes (1986-1990) that had been tested up to that time (Colliver et al., 1990). In Study 1, for each class, generalizability analyses were performed to determine the intercase reliabilities for cases using a single SP and for cases using multiple SPs. Study 2 was conducted to compare the intercase reliabilities of those few cases that were used in two different classes, simulated by single SPs in one class and by multiple SPs in the other. Study 3 was undertaken to determine directly by statistical analysis of the multiple-SP cases only, the contributions of multiple SPs to measurement error and, hence, reliability. The results of all three studies showed little or no effect of multiple SPs on reliabilities of total, checklist or written scores.
Starting with class of 1988, in addition to completing checklists on history taking and physical examination, the SPs have rated students on each case on interpersonal and communication skills. The five dimensions on which the SPs rated the students were (1) clarity of communication, (2) thoroughness of explanation, (3) professional manner, (4) personal manner, and (5) overall patient satisfaction. The ratings were not used to determine total examination scores, nor were they used to make pass-fail decisions; however, students did receive feedback on the ratings and may have received suggestions for improvement if their ratings were consistently low across cases.

The use of multiple SPs, again, raises questions about consistency, but here the questions are about the consistency of the ratings assigned by the two or more different SPs simulating the same case. Although the ratings were not used to determine total scores or make pass-fail decisions, it is of interest to know if there was an impact of multiple SPs on the reliability of these interpersonal and communication skills ratings. It seems reasonable to think the multiple SPs might have had a greater impact on ratings on non-cognitive dimensions, like patient satisfaction, than on the ostensibly more objective checklist and written scores. Hence, the present study was conducted to assess the effect of multiple SPs on the reliability of the SP ratings of interpersonal and communication skills, using data from the five classes (1988-1992) for which the ratings have been obtained. The analyses in this study, like those in Study 3 mentioned above for total, checklist, and written scores, were based on data obtained for each class from just those cases that were simulated by multiple SPs and were performed to determine statistically the impact of the multiple SPs on measurement error and, hence, reliability.

Methods

Detailed discussions of the SIU Post-Clerkship Examination, including thorough coverage of its development, administration, and scoring, are presented elsewhere and should be consulted for complete information about the examination (Vu et al., 1992). In brief, the examination is a performance-based examination that uses about 18 40-minute SP cases (20 minutes for the student-SP encounter and another 20 minutes immediately following the
encounter for students to answer written questions about the case). The examination is objectively scored and is administered to all senior medical students (about 70 per class) upon completion of their clinical clerkship rotations. Cases for the examination are chosen by the faculty Post-Clerkship Examination committee and represent the most frequently encountered patient problems as well as the most important patient problems that students are expected to evaluate and manage competently. Each case selected is developed by a faculty physician, who determines the competencies to be assessed on the case and, with the faculty committee, develops instruments for collecting the necessary data about student performance. These data consist of checklists completed by SPs who record actions performed by students on history and physical, and written responses by students following the patient encounters to questions concerning findings, tentative diagnostic conclusions, and plans for treatment and management. Starting with the class of 1988, students have been rated on the five dimensions described above, each on a five-point scale from excellent (5) to poor (1).

Statistical Methods. Analyses were performed on SP ratings on each of the five dimensions, for the five classes (1988-1992) for which ratings were obtained. Analyses were performed only on data obtained from those cases that were simulated by multiple SPs, to determine directly the measurement error resulting from the use of multiple SPs and the effect of that measurement error on reliability. For each of the five rating scales for each of the five classes, data from the multiple-SP cases were analyzed with the maximum-likelihood approach provided by the BMDP 3V statistical package program in Section 3V.5 Nesting (Dixon, 1981). Using this approach, four variance components were estimated for each rating scale for each class, namely, students, cases, patients-within-cases, and Students x Patients. (The latter term is the sum of Students x Cases plus Students x Patients-Within-Cases.) The patients-within-cases variance component provides a direct measure of the measurement error due to the use of multiple SPs to simulate the same case.

Using these variance components obtained with the BMDP 3V maximum-likelihood program, two generalizability coefficients were computed for each rating scale for each class,
one that is free of measurement error due to multiple SPs and the other that reflects this source of measurement error. The coefficient that is free of measurement error due to multiple SPs was computed by dividing the students variance component by the sum of the students plus the Students x Patients variance components. The coefficient reflecting measurement error due to multiple SPs was computed by dividing students by the sum of students plus Students x Patients plus patients-within-cases. A Spearman-Brown correction was applied to both coefficients to adjust to the number of cases (k) on which the coefficients were based. For the five classes (1988-1992), the number of multiple SP cases on which the coefficients were based were 9, 10, 8, 5, and 10, respectively.

In addition, the variance components (and standard errors) for each rating scale were pooled across the five classes, and overall generalizability coefficients based on the pooled variance components were computed. The pooled generalizability coefficients were weighted means of the original variance components, where the weights were the number of cases (k) and the number of students (n) on which the original variance components were based (Sirotnik, 1974). Using these variance components, two pooled generalizability coefficients were computed for each rating scale, one that is free of measurement error due to multiple SPs and the other that reflects this source of measurement error. A Spearman-Brown-type correction was applied to both coefficients to adjust the coefficients to a common number of k = 18 cases. The analytic approach used here is the same as used in Study 3 in our study of the effects of multiple SPs on the reliability of total, checklist, and written scores (Colliver, et al., 1990) and is similar to that described by Swanson and Norcini (1989) except that all variance components were estimated here in a single analysis rather than being obtained by combining results from two separate analyses.

Results

For ratings on "clarity of communication," the generalizability coefficient free of measurement error due to multiple SPs was .71 and the coefficient reflecting measurement error due to multiple SPs was .64, indicating a 7 percentage point decrease in reliability due to the use
of multiple SPs. (See Table 1.) For each of the five classes (1988 - 1992) considered separately, the differences between the two coefficients ranged from 4 to 13 percentage points. For ratings on "thoroughness of explanation," the coefficient reflecting measurement error due to multiple SPs was .76 and the coefficient reflecting that source of measurement error was .72, a difference of 4 percentage points. The differences between the two coefficients for the five separate classes ranged from 1 to 10 percentage points.

For ratings on "professional manner," the two coefficients were .78 and .75, respectively, differing by 3 percentage points. (See Table 1.) For the five classes, the differences ranged from 2 to 10 percentage points. For ratings on "personal manner," the two coefficients were .72 and .67, with a 5 percentage point difference. For the five classes, the differences ranged from 3 to 9 percentage points.

Finally, for ratings on "overall patient satisfaction," the two coefficients were .79 and .74, with a 5 percentage point difference. (See Table 1.) For the five classes, the differences ranged from 2 to 18 percentage points.

Discussion

In general, the results showed that generalizability coefficients free of measurement error due to multiple SPs were larger than coefficients reflecting measurement error due to multiple SPs by an average of about 5 percentage points, suggesting that the use of multiple SP raters on the same case decreased intercase reliability by about 5 percentage points. By contrast, multiple SPs had a much smaller, almost negligible, effect on total, checklist, and written scores, as reported in our earlier study (Colliver et al., 1990). For total scores, the generalizability coefficient free of measurement error due to multiple SPs was .65 and the coefficient reflecting that source of error was .64, a difference of 1 percentage point. For checklist scores, the two coefficients were .57 and .54, a difference of 3 percentage points; and for written scores, the coefficients were .62 and .61, a difference of 1 percentage point. (Note: the coefficients for total, checklist, and written scores reported in our earlier study were for an examination consisting of k = 10 cases (Colliver et al., 1990), whereas here they were adjusted to k = 18.
cases, for purposes of comparability with the coefficients for the ratings.) In brief, the differences in reliability of total, checklist, and written scores due to multiple SP were about 1 or 2 percentage points. Thus, the effect of multiple SPs on the reliability of ratings is about 4 or 5 times greater than the effect of multiple SPs on the more objective checklist and written (and hence total) measures. The explanation would seem to be related to the more subjective nature of ratings as compared to checklist and written scores. Nevertheless, the magnitude of the effect of multiple SPs on ratings is on average only about 5 percentage points, which is not insignificant, but not excessive either.

Similar results for checklist (data collection) scores were reported by Swanson and Norcini (1989), for a reanalysis of data from two data sets. In the studies reanalyzed by Swanson and Norcini, multiple physician observers, not multiple SPs, completed the data-collection checklists, and little or no effect on reliability was observed. The results showed that using two physician observers to complete checklists on the same case reduced overall variance of the checklist scores only about 1 percentage point. However, for ratings of communication skills, having two physician raters reduced variance by only about 1 or 2 percentage points, which is smaller than the reduction found for communication skills ratings in the present study. In the present study, the reduction was 7 percentage points for "clarity of communication" and 4 percentage points for "thoroughness of explanation". These differences between the two studies in the magnitude of the effect of multiple SPs on ratings of communication skills may be related to using multiple physician observers rather than multiple SPs, although the reason for a physician observer versus SP difference (if real) is not clear. It seems more likely that the differences are simply due to sampling error given the variability in the differences between the two generalizability coefficients across the five classes in the present study.

In conclusion, the results show that the use of multiple SPs reduced the reliability of ratings of interpersonal and communication skills by about 5 percentage points. Compared to earlier findings, the effect of multiple SPs on the reliability of ratings is about 4 or 5 times greater than their effect on total, checklist, and written scores. Nevertheless, the magnitude of
the effect is not excessive. However, if SP ratings are used to make important decisions concerning graduation, licensure, or certification, the use of multiple SPs for rating examinees on interpersonal and communication skills should be undertaken with caution.
References


