Assessment of Uniqueness of Information Provided by Postencounter Written Scores on Standardized-Patient Examinations

Colliver, Jerry A.; Travis, Terry A.; Robbs, Randall S.; Vu, Nu Viet; Marcy, Michelle L.; Barrows, Howard S.

Assessment of Uniqueness of Information Provided by Postencounter Written Scores on Standardized-Patient Examinations

Jerry A. Colliver, Terry A. Travis, Randall S. Robbs, Nu V. Vu, Michelle L. Marcy and Howard S. Barrows

Eval Health Prof 1992 15: 465
DOI: 10.1177/016327879201500408

ASSESSMENT OF UNIQUENESS OF INFORMATION PROVIDED BY POSTENCOUNTER WRITTEN SCORES ON STANDARDIZED-PATIENT EXAMINATIONS

JERRY A. COLLIVER
TERRY A. TRAVIS
RANDALL S. ROBBS
NU V. VU
MICHELLE L. MARCY
HOWARD S. BARROWS
Southern Illinois University
School of Medicine

The complete clinical encounter station uses two testing methods in the assessment of clinical competence: checklists completed by standardized patients (SPs) who record actions students performed on history and physical examination and written responses by students to a series of written questions designed to elicit findings they see as pertinent to the problem, their working hypotheses, their plans for laboratory findings, and their diagnoses and management plans. Given that the critical part of any SP station is the standardized-patient encounter and that the postencounter session needed for the student to answer the written questions increases the time required to administer an already lengthy examination, questions have been raised about the need for written questions in addition to SP checklists in the clinical encounter station. Thus it is important to determine if the written scores are providing additional, nonredundant information over and above that provided by the checklist scores. The results of this study showed only moderate overlap between the checklist and written scores. The mean checklist-written correlation across 83 SP cases was .13, and the mean correlation between the average checklist and written scores across six classes of medical students was .32. The mean of the correlations across the six classes, disattenuated for measurement error, was only .57. The results provide strong evidence that the written scores provide unique information, in addition to that provided by checklist scores.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 15 No. 4, December 1992 465-474

Standardized patients (SPs) are being used widely for assessing the clinical competence of medical students and residents (Hart, Harden, & Des Marchais, 1992), and SP cases are being developed for use in national licensure and certification examinations in the United States (Melnick, 1992) and Canada (Reznick, 1992). A number of formats have been used for the presentation of SP stations, with the various formats differing along a number of dimensions, including the length of the SP station, the type and nature of the skills/competencies tested, and the methods used to score examinee performance (van der Vleuten & Swanson, 1990). One such format is the clinical encounter station that requires the examinee to interact with the SP in a complete clinical encounter that simulates the actual clinical situation the examinee will face in medical practice (Barrows, 1987). Whereas the typical OSCE (Objective Structured Clinical Examination; Harden & Gleeson, 1979) only "biopsies" the clinical encounter by looking at the performance of specific tasks or technical skills, such as "Obtain a history from this patient who gives a history of breathlessness" or "Carry out a neurological examination of the lower limbs, excluding sensation and coordination," the clinical encounter station requires the examinee to interact with the SP in a complete clinical encounter in which the examinee must integrate a variety of skills and abilities as well as inquire, generate hypotheses, and communicate with the patient.

One of the distinguishing features of the complete clinical encounter station is that two testing methods are used in the assessment of clinical competence: checklists completed by the SPs and written responses by students. Immediately following the student-SP encounter, SPs record on checklists the actions students performed on the history and physical examination, and, independently, students respond to a series of written questions designed to elicit findings they see as pertinent to the problem, their working hypotheses, their plans for laboratory findings, and their diagnoses and management plans. Examples of items on the SP checklist that the SP checks if the student elicits the history information or performs the physical examination action are as follows. "I first noticed this problem a year ago," "I do not have any numbness or tingling," "Undressed patient and looked at shoulder muscles, front and back," and "Checked strength in

shoulders and upper arms (patient asked to lift or bend arms)." Examples of written questions that students answer following the SP encounter are as follows. "Using the list below, place an 'x' in front of the findings you obtained from history and physical examination. Points will be taken off for incorrect answers." This is followed by a list of eight possible findings including Petechiae, Left Arm Drift, and Depressed Tendon Reflexes on the Left Side. "From the laboratory tests and diagnostic procedures below, choose 3 that you feel are necessary by placing an 'x' in front of the tests or procedures. Choose only those tests needed to make a more accurate diagnosis or plan appropriate management." This is followed by a list of tests and procedures including Antinucleus Antibody, CAT Scan, and EKG. Thus, on each SP case, students receive two separate scores, a checklist score and a written score.

Questions have been raised about the need for written questions in addition to checklists completed by SPs in the clinical encounter station, even though the written questions are designed to measure something quite different than the checklists. Given that the critical part of any SP station regardless of format is the standardized-patient encounter with a checklist designed to evaluate that encounter, the written questions could be seen as superfluous or at least redundant. The postencounter session needed for the student to answer the written questions doubles the length of the station and hence doubles the time required to administer an already lengthy examination. Thus it is important to determine if the written scores on the clinical encounter stations are providing additional, nonredundant information over and above that provided by the checklist scores. Otherwise, there is no justification for their continued use in SP-based examinations.

The purpose of the present study was to assess the presence and extent of overlap between checklist and written scores using data from six classes of senior medical students on an examination composed of clinical encounter stations. For each of the 83 stations used with the six classes, for which both checklist and written scores were available, the correlation between the checklist and written scores was computed. For each class, the correlation was computed between checklist and written scores averaged across all stations used with that class. To provide an assessment of overlap not attenuated by measurement

error, correlations corrected for measurement error were also computed. The intercase reliabilities for checklist and written scores, used to disattenuate the correlations, are reported here for the first time.

METHODS

THE EXAMINATION

At the Southern Illinois University School of Medicine, a Post-Clerkship Examination consisting of clinical encounter stations has been used to assess the clinical competence of six classes (1986 to 1991) of senior medical students (about 70 students per class). Students were expected to pass the examination as a part of their graduation requirements. Detailed descriptions of the development, administration, and scoring of the clinical-competence examination are presented elsewhere and should be consulted for a thorough discussion of the examination (Barrows, Williams, & Moy, 1987; Williams et al. (1987). In brief, the examination was a performance-based examination that used about eighteen 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 was objectively scored and was administered to all senior medical students on completion of their clinical clerkship rotations. The examination cases, all based on actual patient cases, were selected from a list of chief complaints and diagnoses of patient cases that faculty expect graduating students to be able to evaluate and in some cases manage. The clinical competencies to be assessed were determined and instruments were developed for collecting necessary data about student clinical performance. Students proceeded through the examination by first seeing each SP, performing a focused history and physical, and providing the patient with tentative diagnostic conclusions and management plans. SPs recorded on checklists whether selected actions were performed by a student on the history and physical examination. Students then responded to a series of written questions designed to elicit findings pertinent to the problem, their working hypotheses, their plans for laboratory investigation,

their interpretation of laboratory findings, and their diagnosis and management plan. Thus, on each case, students received separate scores obtained from checklists completed by SPs and from written responses of students following the patient encounters. Both checklist and written scores were means of percentages of possible points on separate examination sections. Responses were evaluated using a preset answer key with predetermined weights assigned to each question. Standards of acceptable performance for each case were proposed by the case author and reviewed by a committee using current standards of quality care for the medical problem and the expected level of performance of a graduating student.

STATISTICAL METHODS

For each SP case, the Pearson correlation between the checklist and written scores was computed. For each class, an average checklist score and an average written score were computed for each student by averaging the student's checklist and written scores across all cases that had both checklist and written components. The Pearson correlation between these checklist and written score averages was computed for each of the six classes. Also, for each class, generalizability analyses were performed on checklist and written scores, to obtain an intercase reliability for each (Brennan, 1983; Colliver, Verhulst, Williams, & Norcini, 1989). These intercase reliabilities, then, were used to disattenuate the checklist-written correlations for measurement error (Lord & Novick, 1968). Intercase reliabilities for all classes combined were based on variance components pooled across classes. Correlations were averaged across cases and classes using Fisher's z transformation (Ferguson & Takane, 1989).

RESULTS

CORRELATIONS

For all six classes, a total of 83 cases had both a checklist and a written score component. The case-by-case correlations between the

checklist and written scores ranged from –.20 to .73. The mean correlation across all 83 cases was .13.

The correlations between the average checklist and written scores for each of the six classes ranged from .16 for the class of 1988 to .50 for the class of 1986 (see Table 1). The mean of the correlations for the six classes was .32.

RELIABILITIES

For checklist scores, the generalizability coefficients for the six classes ranged from .16 for the class of 1990 to .61 for the class of 1988 (see Table 1). Using variance components pooled across all six classes, the generalizability coefficient for an examination consisting of 18 cases was .52.

For written scores, the generalizability coefficients ranged from .25 for the class of 1987 to .65 for the class of 1991 (see Table 1). Using variance components pooled across the six classes, the generalizability coefficient for an 18-case examination was .61.

DISATTENUATED CORRELATIONS

The correlations disattenuated for measurement error for each of the six classes ranged from .33 to 1.06 (see Table 1). The mean of these disattenuated correlations was .57.

DISCUSSION

The results show only moderate overlap between the checklist and written scores. The mean correlation across the 83 cases was .13, and the mean correlation between the average checklist and written scores across the six classes was .32. The case-by-case correlations were lower and more variable than those for the six classes, due to the lower reliability of individual case scores. The mean of the disattenuated correlations across the six classes was only .57. Given that the checklist and written scores had quite similar variances for each of the six classes, a common-factor interpretation of the attenuated correlation

TABLE 1

Correlations Between Average Checklist and Written Scores for Six Classes[a]

		Checklist				*Written*				*Correlations*	
Year	k	*Student*	*Case*	*Student by Case*	*Gen. Coef.*[b]	*Student*	*Case*	*Student by Case*	*Gen. Coef.*[b]	r	*Adjusted* r
1986	(11)	18.19	130.80	186.54	.52	25.33	113.47	163.10	.63	.50	.87
1987	(12)	12.13	46.06	141.83	.51	4.31	67.22	153.97	.25	.38	1.06
1988	(14)	23.01	195.45	207.19	.61	5.81	123.37	132.61	.38	.16	.33
1989	(13)	11.90	205.04	189.03	.45	8.92	112.71	112.99	.51	.22	.46
1990	(18)	3.26	164.35	304.77	.16	13.97	67.93	151.37	.62	.19	.60
1991	(15)	13.07	187.24	217.28	.47	16.77	60.16	137.73	.65	.43	.78
All classes	(18)	13.13	159.04	214.62	.52	12.52	88.83	141.91	.61	.32	.57

a. Correlations disattenuated for measurement error are also presented as are generalizability coefficients (with variance components) on which disattenuated correlations were based. *k* is number of cases with both checklist and written components, on which generalizability coefficients and correlations are based.
b. Generalizability coefficients.

would be that of about 30% of the checklist and written score variances are due to a factor shared in common by the checklist and written scores (i.e., common factor variance) (Harmon, 1967). Given that the reliabilities of the checklist and written scores were similar, the disattenuated correlation could be interpreted to mean that about 55% to 60% of the error-free variance in the checklist and written scores is due to this shared common factor. In other words, only 55% to 60% of the error-free variance in the checklist and written scores is due to overlap in what is measured by these two scores. Thus 40% to 45% of what is being measured by checklist scores and by written scores is unique (i.e., nonoverlapping or nonredundant). A regression interpretation of the attenuated correlation squared (i.e., $.32^2 = .10$) would be that only 10% of the variance in the written scores is predictable from checklist scores and hence that 90% of the written score variance is independent of checklist score variance. The disattenuated correlation squared (i.e., $.57^2 = .32$) could be interpreted to mean that only 32% of the error-free variance in written scores is predictable from the error-free variance in checklist scores and hence that 68% of the written score variance is independent of checklist variance.

The results provide strong evidence that the written scores are providing additional, nonredundant information over and above that provided by checklist scores. This seems reasonable given that checklists document actions students performed on history and physical examination and that written responses reflect students' reactions to and interpretations of their actions in terms of their perceptions of the pertinence of the findings, their hypotheses, and their plans for further inquiry and management. Of course, the results do not establish the validity of the written scores, or for that matter, of the checklist scores either. But it is of importance to demonstrate that the written scores provide unique information over and above that provided by the SP checklists, to justify their continued use in SP-based examinations. Ideally, it would be nice to do a definitive, criterion validity study by correlating the checklist and written scores with generally accepted gold-standard criteria to show that the checklist and written scores predict those aspects (and only those aspects) of clinical competence that they were designed to measure. But, such gold-standard criteria are not available, certainly not performance-based criteria, which

examinations like the Post-Clerkship Examination would demand. If anything, standardized-patient examinations probably come closer to constituting such gold-standard criteria of clinical competence than any type of measure available. Consequently, our demonstration that written scores provide unique information over and above that provided by checklists is of importance.

The results of several other studies that reported correlations between checklist and written scores have been summarized, and the correlations were disattenuated for measurement error for presentation in the summary (van der Vleuten & Swanson, 1990). The correlations for these other studies ranged from .22 to .69, with an unweighted mean of .46; the disattenuated correlations ranged from .32 to .91, with an unweighted mean of .73. These correlations are slightly larger than those reported here. However, these studies differed from the present study and each other in a number of ways, including the format of the postencounter written examination (e.g., short answer, multiple choice, and pattern recognition), the length of the written test (5 min to 20 min), and the number of examinees on which the correlations were based. The present study was based on data for six classes of over 400 senior medical students and thus allowed for stable estimates of the checklist-written correlations both attenuated and disattenuated. Nevertheless, the results of these studies and the present study are not that dissimilar, and all show that written scores provide unique information, in addition to that provided by the checklist.

REFERENCES

Barrows, H. S. (1987). Multiple stations. In I. R. Hart & R. M. Harden (Eds.), *Further developments in assessing clinical competence* (pp. 111-113). Montreal, Quebec: Can-Heal Publications.

Barrows, H. S., Williams, R. G., & Moy, R. H. (1987). A comprehensive performance-based assessment of fourth-year students clinical skills. *Journal of Medical Education, 62*, 802-809.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Publications.

Colliver, J. A., Verhulst, S. J., Williams, R. G., & Norcini, J. J. (1989). Reliability of performance on standardized-patient cases: A comparison of consistency measures based on generalizability theory. *Teaching and Learning in Medicine, 1*, 31-37.

Ferguson, G., & Takane, Y. (1989). *Statistical analysis in psychology and education.* (6th ed.). New York: McGraw-Hill.

Harden, R. M., & Gleeson, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education, 13*, 41-54.

Harmon, H. M. (1967). *Modern factor analysis* (2nd ed.). Chicago: University of Chicago Press.

Hart, I. R., Harden, R. M., & Des Marchais, J. (Eds). (1992). *Current developments in assessing clinical competence.* Montreal, Quebec: Can-Heal Publications.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Melnick, D. (1992). What is happening to licensing examinations? In I. R. Hart, R. M. Harden, & J. Des Marchais (Eds.), *Current developments in assessing clinical competence* (pp. 4-8). Montreal, Quebec: Can-Heal Publications.

Reznick, R. K. (1992). The Q5 project: A national clinical skills examination. In I. R. Hart, R. M. Harden, & J. Des Marchais (Eds.), *Current developments in assessing clinical competence* (pp. 24-26). Montreal, Quebec: Can-Heal Publications.

van der Vleuten, C.P.M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and Learning in Medicine, 2*, 58-76.

Williams, R. G., Barrows, H. S., Vu, N. V., Verhulst, S. J., Colliver, J. A., Marcy, M. L., & Steward, D. E. (1987). Direct, standardized assessment of clinical competence. *Medical Education, 21*, 482-489.