Using Standardized Patients to Assess the Interpersonal Skills of Physicians

JOHN R. BOULET, MIRIAM FRIEDMAN BEN-DAVID, AMITAI ZIV,
WILLIAM P. BURDICK, MICHAEL CURTIS, STEVE PEITZMAN and NANCY E. GARY

The Educational Commission for Foreign Medical Graduates (ECFMG) has developed a high-fidelity performance assessment that will be used to measure the clinical skills of graduates of foreign medical schools (FMGs). The assessment uses standardized patients (SPs) who are trained to document the data-gathering techniques and evaluate the communication skills of physicians. The use of SPs to assess professional qualities is not new, and this method is currently being used to make licensure decisions.[1] The specific purpose of the ECFMG clinical skills assessment (CSA) is to evaluate the readiness of FMGs to enter postgraduate training programs in the United States. To obtain ECFMG certification, graduates of foreign medical schools must have their diplomas verified, pass the United States Medical Licensure Examination (USMLE) Steps 1 and 2, pass an English proficiency test, and then pass the CSA. The CSA measures proficiencies in a number of skills, including history taking, physical examination, doctor–patient communication, and written communication. Since pass/fail decisions will be made based on the CSA scores, it is important to establish the psychometric adequacy of all CSA component scores that are generated.

Interpersonal and communication skills are essential for effective clinical work. Both patient compliance and patient satisfaction can be adversely affected by physician deficiencies in this domain. Furthermore, a physician's willingness and/or ability to address patients' concerns may affect medical diagnosis and treatment.[2] Therefore, in addition to medical knowledge and technical skills, the teaching and evaluation of communication and interpersonal skills have become priority tasks of medical educators. The literature suggests that standardized patients can be used to assess the noncognitive aspects of the doctor–patient relationship.[3] Although other methods of assessment exist for measuring doctor–patient communication skills, they may not provide both the verisimilitude and degree of standardization that are associated with SP-based examinations. Nevertheless, while the authenticity within the assessment process is important, it remains imperative that the generated scores be reliable and valid.

Historically, a number of methods have been used to evaluate the interpersonal skills of physicians, including written simulations, direct observations and questionnaires,[4] and SP evaluations. In SP examinations, or objective structured clinical examinations (OSCEs), the SPs or the physician examiners are generally asked either to document specific behaviors associated with interpersonal skills via checklists or to make holistic ratings based on patterns of defined behaviors and actions. Regardless of the type of scoring, the observation of students or physicians as they perform clinical tasks provides a high-fidelity milieu in which to assess doctor–patient communication skills. Such a milieu is likely to enhance both the content validity and the generalizability of the scores.

Traditionally, history-taking and physical-examination skills have been measured on SP examinations through the use of case-specific checklists. The SP is required to record whether the action described by each checklist item was performed. For interpersonal skills, which are less technical and less likely to be tied specifically to the content of the clinical encounter, the generation of case-specific checklists seems less meaningful. Recently, a number of researchers have compared the psychometric properties of checklists and rating scales used to assess communication skills.[5–7] For the most part, the more "objectified" checklists were not found to produce more reliable scores than their holistic, or global, counterparts.

The purpose of this study was to provide some preliminary evidence to support the validity and reliability of the scores derived from the ECFMG rating process for interpersonal skills. If SPs are going to provide judgments of interpersonal attributes that may have high-stakes consequences, it is essential that these ratings be valid and reproducible.

Methods

Description of the CSA. The ECFMG's CSA consists of ten scoreable SP encounters. Unlike some SP-based assessments that have specific stations for evaluating the ability of the physician to interact with the patient, information about interpersonal skills is collected in each of the cases on the CSA. As in most OSCE formats, the SPs document the history-taking and physical-examination skills of each examinee. For the ECFMG's CSA the history-taking and physical-examination checklists are combined to form a data-gathering score. The SPs also evaluate the interpersonal skills of the examinees based on global, or holistic, ratings for four dimensions (described below). A single rating of spoken-English-language proficiency of each examinee is also solicited from the SP in each station. While the SP is documenting and evaluating the data-gathering, interpersonal, and spoken-English skills, the examinee summarizes and interprets the data gathered in the clinical encounter via a written patient note. Trained health care professionals score the patient note.

A communication challenge is also included in most CSA cases. These challenges are deliberately provocative statements introduced by the SPs to encourage further communication between doctor and patient. For example, an SP with chest pain will tell the physician that he is now feeling fine and wishes to leave the emergency department. The physician is expected to respond to the challenge, thereby expanding the sample of communication behaviors that are evaluated by the SP.

Interpersonal skills rating. A behavior-anchored rating tool was developed by the ECFMG to allow SPs to systematically evaluate the interpersonal skills (IPSs) of physicians. Four dimensions of performance are evaluated: (1) skills in interviewing and collecting information—this includes references to clarity of questions, use of open versus closed questions, jargon, verification, summarization and transitions; (2) skills in counseling and delivering information—this includes references to giving information, counseling, closure of the encounter, summarization, and connection; (3) rapport (connection between a doctor and a patient)—attentiveness, body language, confidence, attitude, empathy, and support are assessed in the evaluation; and (4) personal manner—this includes facets such as hygiene, draping, mood, and introduction. The SPs evaluate each of these four dimensions using a four-point Likert scale ranging from unsatisfactory to excellent. The SPs are trained based on detailed matrices of behavioral descriptors that reflect interpersonal skills. For each dimension (e.g., rapport), and each pos-

sible score category (unsatisfactory, marginally satisfactory, good, excellent), a detailed list of behaviors is provided. For example, on the "skills in interviewing and collecting information" dimension, there are specific descriptions of behaviors for clarity of questions, open versus closed questions, etc., under each of the score categories. These matrices are used to train the SPs to identify behaviors that will substantiate specific ratings. Standardized-patient training for evaluating interpersonal skills takes approximately eight hours to complete. The instruction includes role play, practice evaluation using videotaped encounters, and sensitivity training to help eliminate potential biases.

Checklist items. As part of the data-gathering checklists there are also specific communication and counseling items (four or five per case) that the SP marks as being, or not being, done. These items (e.g., "examinee introduces self to patient," "examinee closes interaction in a respectful manner") are not used to generate the doctor–patient communication score. Instead, they are retained in the checklist (1) to provide some objective anchors for the SP's holistic evaluations of IPSs and (2) to furnish a score with which to gather some evidence of criterion-related validity for ratings of the IPS dimension. A score based on the percentage of communication checklist items achieved in each station is calculated.

Sample. Both FMGs and U.S. medical students recently participated in a large-scale norming study designed to test CSA case material and methods. While over 350 FMGs participated, the sample used for this report included only the 123 who had successfully completed the USMLE Steps 1 and 2 and the ECFMG English-comprehension test. Based on the ECFMG credentialing requirements, this subsample of physicians would be eligible to take the CSA. Participation in the study was voluntary and had no bearing on the future careers of the participants. Each FMG received an honorarium for participating in the CSA. The sample of 123 FMGs was exposed to five different CSA test forms, comprised of 38 unique SP encounters. The results for the U.S. medical students are not reported here.

The sample consisted of 68 men and 55 women; a majority identified themselves as Asian (35.8%), Middle Eastern (16.3%), or African (8.1%). Most (86.2%) were not American citizens, and 76% reported that the language most often spoken at home was not English. Approximately half of them had attended a medical school where the language of instruction was English. The mean age was 31.8 years (SD = 5.0) and, on average, 6.9 (SD = 5.2) years had elapsed since the medical degree had been received. The mean three-digit passing scores on the USMLE Steps 1 and 2 for this group were 193 and 189, respectively.

Analysis. Reliability analyses were conducted to investigate the consistency of IPS scores, both across the rated interpersonal dimensions and over the ten clinical encounters that an examinee was required to complete. Correlations among the IPS dimension scores (four dimensions), total scores, communication checklists, and other CSA components were calculated to provide evidence for the validity of interpretations of IPS scores. All analyses were performed using the SAS statistical software package.[8]

Results

The overall mean IPS score (averaged over the four dimensions) was 2.62 (SD = .39), somewhere between marginally satisfactory and good. The mean scores for the four IPS dimensions were 2.7, 2.5, 2.7, and 2.6. The reliability of the IPS dimension scores was also high ($\alpha = .81$), indicating that the profile of interpersonal attributes was consistent for a given examinee. This highlights the fact that a physician with poor skills in one interpersonal area (e.g., counseling) is unlikely to have excellent skills in another (e.g., interviewing).

The generalizability of the IPS ratings over all ten cases was high ($\rho^2 = .85$), suggesting that examinee performance was reasonably consistent across the sample of different cases (SPs). This coefficient can be interpreted as the expected correlation between mean ratings of independent groups of ten SPs. Nevertheless, the variance attributable to the particular case was not negligible, indicating that the various SPs differed somewhat in average stringency. The dependability coefficient ($\phi = .63$), which considers the particular case (or, equivalently, the particular SP) as a potential error source, was still moderately high. Divergence in the mean scores assigned by various SPs (interquartile range = 2.45 to 2.88) may have been a function of the abilities of the examinees who took particular cases. Although the sample of FMGs was exposed to 38 different cases, each examinee took only ten cases, and therefore the samples of examinees who took certain cases may have had different levels of interpersonal skills than those who took others. The generalizability coefficients for the four IPS dimensions for a ten-case assessment were .70, .81, .81, and .72, respectively.

The correlations among IPS scores and other CSA components are presented in Table 1. The intercorrelation patterns provide evidence that the interpersonal tool, as rated by the SPs, provides valid scores. First, as expected, correlations between interpersonal scores (both for the total and for the various dimensions) and English-language proficiency were high. Here, the strongest association was between English and the "interviewing and collecting information" dimension (r = .69). While the same SP evaluated both the IPSs and English skills, possibly leading to disattenuated coefficients, one would expect at least moderate amounts of shared variance between these components/dimensions. Second, the IPS scores were more highly correlated with data gathering than with the patient note. While the SPs are specifically trained to provide information only when certain questions are asked, physicians with strong interpersonal skills would be likely to be more proficient at

Table 1. Correlations among Various Component Scores of the Clinical Skills Assessment, Educational Commission for Foreign Medical Graduates (ECFMG) Norming and Validity Study, 1996–1997

	IPS*	IPS1	IPS2	IPS3	IPS4	COMCK	ENG	DG	PN
IPS (total score)*		.90	.90	.93	.90	.64	.64	.44	.37
IPS1 (interviewing)			.78	.79	.76	.54	.69	.53	.44
IPS2 (counseling)				.77	.71	.65	.54	.38	.34
IPS3 (rapport)					.81	.57	.56	.33	.28
IPS4 (personal manner)						.56	.55	.40	.34
Communication checklist (COMCK)							.42	.31	.26
English (ENG)								.29	.41
Data gathering (DG)									.63
Patient note (PN)									

*IPS = interpersonal skills.

soliciting relevant information from the patient. The patient note, which is not scored by the SP, measures an examinee's ability to document and interpret the information collected in the clinical encounter. This ability would be expected to be less related to one's interpersonal skills. Finally, as expected, the correlations between the communication checklist scores and the interpersonal ratings were reasonably high (r = .54–.65).

Discussion and Conclusions

The results of this investigation indicate that well-trained SPs can be used to provide reliable evaluations of physicians' interpersonal skills. The moderately high intercorrelations among the dimension ratings suggest that SPs form an overall impression of the physician's abilities and do not differentiate particularly well among the behaviors that these ratings are based on. Alternately, the interpersonal behaviors of physicians would be expected to be reasonably homogeneous (i.e., a physician with excellent skills in interviewing would also be expected to have excellent rapport). Nonetheless, the inter-station reliabilities (over ten cases) reported here would be acceptable not only for teaching and learning but also for high-stakes examinations. This finding contrasts with that of Hodges et al.[5] who found that the generalizability of communication skills between communication stations was low. Our results support the hypothesis that, at least for interpersonal skills, an examinee's performance is not highly dependent on the medical content of the case. This is encouraging in that, if interpersonal skills were highly case-specific, medical educators would be forced into the difficult and laborious task of linking training to specific types of clinical encounters. From a test-development perspective it would also suggest that case-based standards be set and that additional content constraints need to be placed on the sets of cases considered for CSA administrations. Given the present results, it is not necessary to balance CSA forms based on the expected levels of interpersonal skills required for various types of patient interactions.

The lack of content specificity of interpersonal skills in our data is most likely a function of the structure of the CSA. Consistent with the purpose of the CSA, only cases that reflect common clinical encounters were included. If cases that created major communication difficulties (e.g., sexual abuse, bipolar mood disorder) were entertained, this finding might not hold.

Based on the encouraging psychometeric properties of holistic ratings, and the moderate correlations between these ratings and short sets of specific checklist items, the use of standardized patient judgments is supported. While methods based on checklists may be more objective and therefore provide more reliable scores, they also trivialize the assessment of interpersonal skills by negating the patient's perception of the totality of the encounter.[9] This, in effect, will serve to diminish the validity of the resulting scores by constraining the assessment to only those specific behaviors (e.g., eye contact) that may be contained on the checklist. Furthermore, regardless of the scoring method, there is little evidence to suggest that any health care provider's outlook can be substituted for the patient's perception of the interaction.[4] Finally, any biases that might be introduced through the use of holistic judgments are likely to be no greater than those associated with non-patient ratings, and are effectively minimized through proper training and calibration and the use of multiple judgments.

We found that, across cases, there was some divergence in the mean scores assigned by various SPs. The sample of 123 FMGs was exposed to 38 different SP cases. Thus, based on the structure of the CSA (ten scoreable cases), it is impossible to ascertain to what degree the samples of examinees who took each of the cases were comparable in terms of interpersonal abilities. Nevertheless, the consistency of scores across the ten-case combinations is encouraging. Current studies are being directed at specifically investigating sources of error attributable to the SP and/or the particular case being simulated. Here, it is necessary to have several SPs per case and to have multiple ratings within a particular case. Even though the interpersonal rating provided by a particular SP is a subjective assessment, and many ratings are provided by ten separate SPs, one would expect to obtain reasonably high inter-rater reliabilities for a given encounter.

Like Hodges et al.,[5] we found that data-gathering scores were moderately correlated with interpersonal skills (r = .44). Whether poor interpersonal skills impede the collection of relevant data[2] or the SPs provide more information to examinees with good interpersonal behaviors remains to be fully understood. Furthermore, graduates of foreign medical schools often have English-language difficulties, which are likely to affect their ability to obtain necessary information from the SP, who in turn may not sense an adequate interaction. The ECGMG, while simultaneously enhancing SP training and developing quality-control procedures, is investigating these and other potential sources of error in the IPS ratings. It should be remembered, however, that the assessment of an examinee's interpersonal skills is based on the independent ratings of ten well-trained SPs, which effectively minimizes potential error sources and does provide for reliable IPS scores.

The quality of the patient–physician relationship is extremely important for an effective clinical encounter. The results from this investigation support the ECFMG's premise that a generalizable set of interpersonal-skills scores can be derived from an SP-based assessment of clinical skills.

Correspondence: John R. Boulet, PhD, Clinical Skills Assessment Program, Educational Commission for Foreign Medical Graduates, 3624 Market Street, Philadelphia, PA 19104-2685; e-mail: <jboulet@ecfmg.org>.

References

1. Reznick R, Blackmore D, Cohen R, et al. An objective structured clinical examination for the licentiate of the Medical Council of Canada: from research to reality. Acad Med. 1993;68(10 Suppl):S4–S6.
2. Evans BJ, Stanley RO, Mestrovic R, Rose L. Effects of communication skills training on students' diagnostic efficiency. Med Educ. 1991;25:517–26.
3. Norman GR, Barrows HS, Gliva G, Woodward C. Simulated Patients. In: Neufeld R, Norman GR (eds). Assessing Clinical Competence. New York: Springer Publishing Company, 1985.
4. 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;69:216–23.
5. Hodges B, Turnbull J, Cohen R, Bienenstock A, Norman G. Evaluating communication skills in the objective structured clinical examination format: reliability and generalizability. Med Educ. 1996;30:38–43.
6. Schnabl GK, Hassard TH, Kopelow ML. The assessment of interpersonal skills using standardized patients. Acad Med. 1991;66(9 Suppl):S34–S36.
7. Cohen DS, Colliver JA, Robbs RS, Swartz MH. A large-scale study of the reliabilities of checklist scores and ratings of interpersonal and communication skills evaluated on a standardized patient examination. Adv Health Sci Educ. 1997;1:209–13.
8. SAS/STAT User's Guide, 4.2, Volume I. Cary, NC: SAS Institute, Inc.; 1990.
9. Van der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Med Educ. 1991;25:110–18.