Interrater Reliability of Standardized Actors Versus Nonactors in a Simulation Based Assessment of Interprofessional Collaboration

David N. Dickter, PhD;

Sorrel Stielstra, PhD;

Matthew Lineberry, PhD

Introduction: There is a need for reliable and practical interprofessional simulations that measure collaborative practice in outpatient/community scenarios where most health care takes place. The authors applied generalizability theory to examine reliability in an ambulatory care scenario using the following 2 trained observer groups: standardized patient (SP, actor) raters and those who received rater training alone (non-SPs).

Methods: Twenty-one graduate health professions students participated as health care providers in an interprofessional care simulation involving an SP, caregiver, and clinicians. Six observers in each group received frame-of-reference training and rated aspects of collaborative care using a behavioral observation checklist. The authors examined sources of measurement variance using generalizability theory and extended this technique to statistically compare the rater types and compute reliability for subsets of raters.

Results: Standardized patient ratings were significantly more reliable than non-SPs' despite both groups receiving extensive rater training. A single SP was predicted to generate scores with a reliability of 0.74, whereas a single non-SP rater's scores were predicted at a reliability of 0.40. Removing each rater one by one from the full 6-member SP sample reduced reliability similarly for all raters (reliability, 0.86–0.89). However, removing individual raters from the full 6-member non-SP sample led to more variable reductions in reliability (0.58–0.72).

Conclusions: Ongoing experience rating performance from within a particular simulation-based assessment may be a valuable rater characteristic and more effective than rater training alone. The extensions of reliability estimation introduced here can also be used to support more insightful reliability research and subsequent improvement of rater training and assessment protocols.

(*Sim Healthcare* 10:249–255, 2015)

Key Words: Medical simulation, Reliability, Assessment, Interprofessional education.

The need for interprofessional collaboration is clear. Health care providers must learn to work on cross-professional teams to coordinate patient care and reduce the propensity for medical errors.[1,2] To achieve high-quality interprofessional practice and collaborative care, health providers need skills-focused education and training. Interprofessional simulations are an excellent modality for individual practice and assessment in collaborative care skills and to support program evaluation and research in interprofessional education (IPE) more broadly. Although IPE is required by accreditation bodies for nursing, pharmacy, physical therapy, physician assistant, public health, and other professions[3] and has become mandatory for MD-granting education programs,[4] researchers have noted the dearth of published evaluation tools needed for directly observing and measuring collaboration competencies.[5] Collecting observation data in simulated teamwork encounters is necessarily a resource-intensive endeavor, requiring well-trained raters to make multiple complex evaluative judgments. Therefore, an enduring priority in assessment research is to both develop theoretical understanding of how various aspects of assessment systems relate to measurement reliability and to build a practical understanding of how to optimize such systems.

One factor to consider when using rater-based assessments of simulated IPE encounters is what requisite characteristics raters should possess and why. As one possibility, the standardized patients (SPs) in a simulated encounter can be asked to assess learners' performance immediately after the encounter. Given their training and experience in portraying their role, SPs may have more elaborate memory structures for potential learner behaviors and therefore be more likely to quickly recognize relevant learner behaviors and accurately evaluate them. Studies have found that SPs

From the Department of Education (D.N.D., S.S.), Western University of Health Sciences, Pomona, CA; and Department of Medical Education (M.L.), University of Illinois at Chicago, Chicago, IL.

Reprints: David N. Dickter, PhD, Interprofessional Education Research and Strategic Assessment, Department of Education, Western University of Health Sciences, 309 E Second St, Pomona, CA 91766 (e-mail: ddickter@westernu.edu).

Support by the Health Resources and Services Administration Cooperative Agreement No. 6UB4HP19202 (to D.N.D.).

The authors declare no conflict of interest.

Presented at the 14th Annual International Meeting on Simulation in Healthcare in San Francisco in January 2014.

Contents of this article are solely the responsibility of the authors and do not necessarily represent the official view of Health Resources and Services Administration.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.simulationinhealthcare.com).

DOI: 10.1097/SIH.0000000000000094

can be effective in rating the simulation participants' performance[6,7] and can do so at least as effectively even as faculty or clinical supervisors when the target behaviors are communication skills or objective patient history checklist items.[8,9] However, such benefits may not accrue until SPs have reached a certain level of experience, and the cognitive demands of portraying the role can inhibit their ability to observe and evaluate learner behaviors.[10]

Videotaping encounters for later observation allows such SPs to role-play and assess in series rather than in parallel, likely addressing the issue of cognitive overload. Such videotaping also greatly increases the pool of potential raters, because videos could be transmitted electronically and reviewed by virtually anyone at their convenience. Taking this possibility to its extreme, researchers have even outsourced video rating tasks to inexpensive "microtask" clearing houses such as Amazon's Mechanical Turk service, hiring surgery-naive individuals over the Internet to observe and rate videos of learners performing surgical suturing.[11] Stepping outside the field of medical education, research in simulation-based assessments of managerial skill has found that senior managers are actually less accurate than nonmanager psychologists at discriminating among different dimensions of performance within examinees (eg, assessing "initiative" vs. "analysis").[12] At the very least, then, raters with relatively little experience in the task domain of interest may be considered as potential raters, although the validity and reliability of their scoring certainly cannot be assumed. In our study, we wished to investigate whether this "crowd-sourcing" approach might indeed be practical and economical, when an intensive training had been conducted with layperson raters who did not act in the simulations (non-SPs). Because it was impractical to provide faculty with the intensive training that we intended for our study, raters included in the study were in these 2 groups (SPs and non-SPs). We designed the simulation and assessment to be nonclinical in nature, to assess the objective competencies that would be appropriate across a range of professions, and to do so without the requirement of expertise in the health care domain.

To investigate the suitability of SPs and non-SPs for inclusion in an assessment system for IPE, we evaluated the reliability of scores on videotaped simulation-based ambulatory teamwork encounters[13] from raters belonging to each group. We collected data to test the hypothesis that SPs generate scores with greater interrater reliability than non-SPs. To do this, we introduce a technique developed by Zhou et al[14] to statistically compare reliability coefficients, which, to our knowledge, has not previously been used in medical education research. In addition, we used an extension of generalizability theory[15,16] to explore rater-specific factors that might explain any differences in measurement precision observed between the 2 groups.

METHODS

We obtained approval from the institutional review board at Western University of Health Sciences (WesternU) to administer, score, and analyze the data presented here.

Participants

Twenty-one students (learners) enrolled in their second or third year of graduate study at WesternU participated in the study (mean age, 29 years; male, 52%). Learners were studying osteopathic medicine (7), dental medicine (5), podiatric medicine (5), pharmacy (3), and physician assistant studies (1). All learners had completed courses in IPE designed to teach competencies including teamwork, collaboration techniques, and communication.[17] Learners were motivated to perform their best because they were informed that their participation would be recorded and they would receive feedback on their performance.

Simulation Scenario and Procedure

This study used the Ambulatory Team Objective Structured Clinical Examination (ATOSCE),[13] which was developed to broaden simulation scenarios from the predominantly emergent and hospital settings found in the extant literature, instead of incorporating outpatient scenarios. The scenario was developed as part of a grant to assess interprofessional collaboration in a geriatric setting among students and practicing health care providers. The scenario was both interprofessional and a realistic depiction of ambulatory care in that learners interacted with providers from multiple professions from a distance (telephone consultations) to form or modify treatment plans requiring collaborative care (eg, referring the patient to other providers, discussing concerns with other providers about medications or treatment plans).[13] Specific content for the scenario was drawn from experiences of clinicians on an advisory panel at the university. The scenario was designed to allow learners to practice and demonstrate continuity of care, conflict management, safety, and patient advocacy. The panel of clinicians and health care educators met to determine the corresponding behaviors that would be assessed in the simulation and to design the scenario to allow opportunity for observing each behavior from each learner, regardless of profession. The scenario and procedures were the same for all learners, including formal instructions before the simulation. In addition, actors portraying the patient, caregiver, and clinicians (collectively referred here using the simulation term "standardized patients" or SPs) were provided with detailed character profiles, wardrobes, props, scripts, and contingencies for responding to learner behaviors. Learners participated individually, with all other roles including other health care providers portrayed by SPs. The simulated encounter involved a learner interacting with an SP, portrayed as a 73-year-old woman, and caregiver (the patient's adult son) in a medical office (see Text Document, Supplemental Digital Content 1, http://links.lww.com/SIH/A217, which contains the case information, instructions, and scripts for SPs portraying the roles). The specialties of clinicians portrayed by SPs were determined by the individual learner's choice of whom to contact/refer to during the encounter. Conversations with these standardized health care providers occurred over the telephone and were recorded. All standardized roles followed scripted dialogue. Feedback has indicated that learners participating in this simulation find it to be realistic and conducive to using their collaboration skills.[13] Learners

were provided with a written overview of the purpose of the simulation (see Text Document, Supplemental Digital Content 2, http://links.lww.com/SIH/A218, brief overview of the ATOSCE methods), and a simulation staff member provided general verbal instructions (see Text Document, Supplemental Digital Content 3, http://links.lww.com/SIH/A219, procedures and time durations for the simulation). Learners did not know the particulars of the case or assessment instrument until beginning the simulation, when they picked up the information hanging on the doorway, consisting of a written description of the SPs who would be entering the room (see Text Document, Supplemental Digital Content 4, http://links.lww.com/SIH/A220, setting, presenting situation and patient's hospital discharge summary). Each learner spent 12 minutes with the patient and caregiver, followed by up to 8 minutes in their "office" where they could make follow-up calls or referrals to other health care providers.

In the scenario, the patient was a geriatric diabetic suffering from the effects of a recent stroke. The case lent itself toward interprofessional teamwork, with issues for multiple health professions to address including chronic disease (diabetes), recent complications (stroke), and concerns about treatment (drug dosages, interactions, medication noncompliance). Other germane physical and psychosocial risks associated with elder care were also simulated, including concerns and frustrations from the caregiver. Learners who called a standardized health care provider to speak about adjustments to the treatment regimen or other interventions for the patient (eg, changes in dosage) encountered scripted resistance, requiring the learner to manage conflict with the provider over differences of opinion about what was best for the patient. The outpatient scenario challenged learners to place the patient at the center of the health care team, to include the caregiver on that team, and to work collaboratively with other health care providers while employing conflict management skills.[13] After the simulation, the SP and standardized family member debriefed the learner, and learners then met together for a group debriefing.

Measures

Based on video recordings of the encounters, raters assessed learner performance using a behavioral observation checklist (Table 1). Development of the content followed from the expert panel who advised on the development of the ATOSCE scenario. Based on the panel's guidance, competencies were addressed in the checklist that derived from the TeamSTEPPS program[18] (eg, team structure, leadership, communication) and the Partnership for Health in Aging[19] (eg, coordination across the care spectrum). For example, learners were rated on their use of the 2-challenge rule, a TeamSTEPPS communication principle (checklist item: "held ground when health care provider challenged their concern") whether they indicated to the patient and caregiver what the next step would be, and which other health care providers would be part of the care team and why (Partnership for Health in Aging–related behaviors). Twenty-one items were developed to assess communication, collaboration, continuity of care, safety, patient advocacy, or conflict management. Items were constructed to be as objective as possible, using dichotomous "performed/not performed" ratings of observed behaviors (eg, "called other health care

TABLE 1. The ATOSCE Assessment Tool

Student…
1a. Spoke directly with patient about their concerns
1b. Spoke directly with caregiver about their concerns
2a. Mentioned ≥1 physical safety hazards (ill-fitting walker, flip flops, lamp cord, throw rug)
2b. Corrected physical safety hazard or suggested a correction
3. Posed a question about medication prescription or compliance
4a. Identified a safety concern regarding medications (not regarding compliance)
4b. Discussed plan to rectify the medication problem (not regarding compliance)
5a. Spoke to patient alone
5b. Made referral (appointment) for patient or telephone contact with HCP or authority about possible elder abuse
6a. Noticed missing information (international normalized ratio blood test results or last page of discharge papers)
6b. Obtained or clarified missing information
7. Offered caregiver (son) a solution or resource to help with care of mother
8. Solicited input when making decisions about patient's treatment plan (ie, included them in discussion, giving them time to speak, not just lecturing them)
a. Solicited input from patient
b. Solicited input from caregiver
9a. Closed the session by telling the patient or son what the next step would be
9b. Confirmed patient or son understood
10a. Mentioned referrals to other HCPs
10b. Described the reasons(s) for referral(s)
11. Called other HCP(s) to refer patient
12a. Expressed concern to another HCP about potential safety issue in treatment plan (eg, medication, delay in appointment, blood pressure, not addressing patient's symptoms)
12b. Held ground when HCP challenged their concern

Checklist format: raters mark yes or no for each item.

provider(s) to refer the patient," "obtained or clarified missing information," "expressed concern to another health care provider about potential safety issues in the treatment plan"). The simulation and checklist items were piloted and revised for a 2-year period before data collection for this study, in a process that included review of the case information and simulated patient data, live observation of the simulation in progress, and ratings on the checklist. Some items were modified to remove ambiguity. A separate faculty panel also reviewed the simulation and checklist and concluded that they were highly relevant to interprofessional care situations that might be experienced with geriatric patients.

Raters

Six SPs and 6 non-SPs independently observed the videotaped simulations. Rating data were fully crossed (ie, all raters evaluated all learners on all behaviors). Observations were performed independently on separate computers in a campus computer laboratory dedicated for that purpose for a 3-day period. Raters wore headphones and were instructed not to discuss the videos. A staff member was present while raters watched the videos and scored performance.

The non-SPs were recruited from local graduate programs. Most had received or were in the process of receiving advanced degrees in fields outside of medicine (eg, psychology, public health) and 1 had previously facilitated an IPE course at the university. The non-SPs completed a full-day, 8-hour training session that included an item-by-item description of the instrument, behavioral observations and group discussion of videotaped examples, and frame-of-reference training[20] of performance exemplars to reach calibration on ratings. This extensive training was conducted to provide raters with a thorough familiarity with the assessment checklist and to allow them to discuss and calibrate with each other on their observations and ratings. Training included the following activities: (1) a review of each assessment item and discussion about the behaviors that it was intended to cover; (2) group observation and rating of video vignettes displaying examples of performance and nonperformance of the target behavior; (3) group discussion (researcher led) about how rater trainees would rate a learner's performance given hypothetical situations with various learner behaviors; (4) group observation and rating of video vignettes using ambiguous learner performance examples of the target behaviors to foster discussion and further calibration; and (5) practice ratings of all assessment behaviors using 6 full-length learner videos (rated independently, followed by group discussion to calibrate).

Standardized patients were drawn from the pool of actors in the clinical skills laboratory at WesternU, trained with the same method and experienced at performing in this specific simulation scenario and in evaluating learners on the performance rubric (collectively, SPs had completed more than 150 ATOSCEs and discussed their ratings afterward, serving as a form of ratings calibration). For approximately 22% of ratings, some SPs were viewing an encounter in which they had performed. However, all encounters had taken place 6 to 12 months before rating, making any unique recall about the encounter unlikely. All raters had sufficient time to view the videos and were able to review and repeat segments of the digital recording as needed.

Statistical Analysis

We hypothesized that SPs would be more reliable than non-SPs because of their extensive ongoing practice with the simulation and familiarity with the assessment tool. Ratings were analyzed using generalizability theory,[15,16] an extension of analysis of variance, which partitions measurement error into its constituent parts. The analysis simultaneously estimates multiple sources of error so that the particular sources that contribute most to overall error can be identified and minimized.[15,16,21,22] Sources of variance associated with learner differences, rater differences, item differences, and interactions thereof were calculated separately for the 2 rater types (SPs vs. non-SPs) using the SPSS VARCOMP procedure with restricted maximum likelihood estimation. Generalizability coefficients (ie, reliability coefficients for relative comparisons among learners) for the 2 groups were computed according to formulas for mixed multi-facet generalizability theory studies detailed in Brennan.[16] Given that the individual items were not sampled randomly from a population of possible items but rather exhaustively covered the items of interest, the "items" facet was fixed in the computation of coefficients (ie, learner-by-item variance was treated as true score variance); in generalizability theory notation, this is given as a $p \times i \times r$ design with items fixed. We computed 95% confidence intervals (CIs) for the generalizability coefficients using the formulas detailed in a report by Zhou et al.[14] Akin to "reliability if item removed" statistics in classical test theory, we also computed coefficients with individual raters removed to determine whether specific raters contributed disproportionately to measurement error.

RESULTS

Table 2 shows the variance components associated with each facet for the 2 samples, and Figure 1 depicts the overall

TABLE 2. Estimated Variance Components for the p x i x r Design, by Rater Type

	SPs		Non-SPs	
	Estimated Variance Component	Total Variance, %	Estimated Variance Component	Total Variance, %
Learner	0.010	4.0	0.003	1.3
Rater	0.000	0	0.003	1.3
Item	0.047	19.0	0.036	15.0
Learner × rater	0.002	0.8	0.007	2.9
Rater × item	0.010	4.0	0.018	7.5
Learner × item	0.112	45.2	0.089	37.1
Learner × item × rater and residual error	0.067	27.0	0.084	35.0

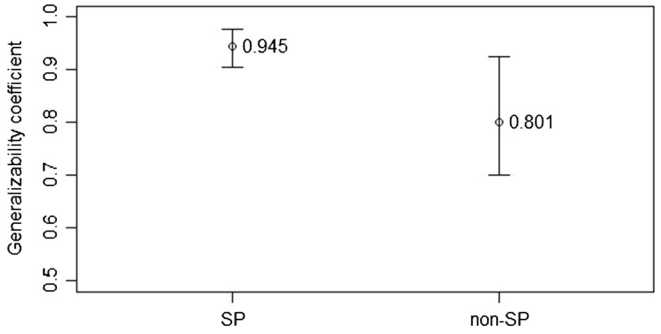

FIGURE 1. Generalizability coefficients for 6-rater panels, by rater type. For the coefficients, 95% CIs are shown.

generalizability coefficients and 95% CIs for the full SP and non-SP samples. Note that ratings for 2 of the 21 items (12a and 12b in Fig. 1) were missing for approximately 10% of examinees because of contingencies linking those items (ie, learners did not perform 12a and raters left 12b blank, or it was unclear whether the standardized clinician sufficiently disagreed with the learner and the rater left 12b blank). The pattern of missing data could not be deemed random, so to avoid biasing subsequent analyses, we omitted these 2 items.

For both SP and non-SP samples, learner-by-item variance was the largest component of variance, followed by the 3-way interaction of learners, items, and raters plus residual error. The SP sample achieved a reliability coefficient ($E\rho^2_{SP}$) of 0.95 [95% CI, (0.91–0.98)], which was significantly greater than that of the non-SP sample [$E\rho^2_{non\text{-}SP} = 0.80$; 95% CI, (0.70–0.92); $P < 0.05$ for the difference]. Using data from Table 2, we also estimated the phi coefficient[23] (using 6 raters and 19 assessment fixed-facet items), a reliability index often used to estimate an examinee's absolute level of skill against some standard; the estimates were 0.94 for SPs and 0.75 for non-SPs.

To evaluate the reliability expected for varying numbers of raters, we conducted separate D-studies for the SP and non-SP rater groups, shown in Figure 2. For SPs, 1 rater is predicted to yield a reliability ($E\rho^2$) of 0.74, whereas a single non-SP rater is predicted to be quite unreliable ($E\rho^2 = 0.40$). To reach even a minimum reliability of 0.70 would require taking the average of 4 non-SP raters.

To determine whether any particular raters were less well calibrated with the others, we systematically re-estimated the G coefficient by extracting each of the study's raters from the

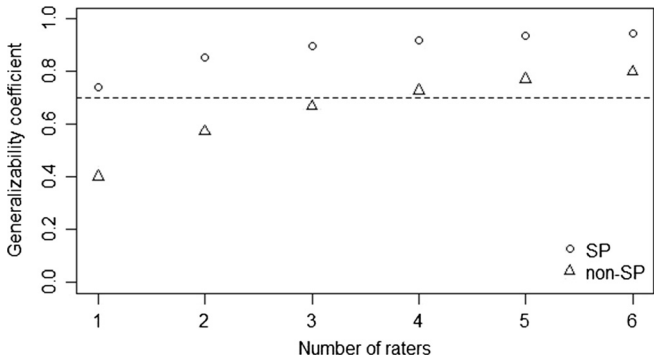

FIGURE 2. D-study reliability coefficients, by number of raters and rater type.

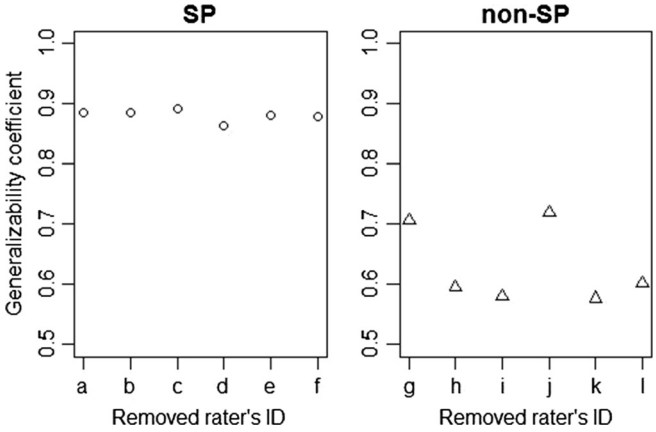

FIGURE 3. Generalizability coefficients for 5-member rating panels removing each rater once, by rater type.

sample 1 at a time, shown in Figure 3. For the SP sample, removal of particular raters has a fairly uniform and small negative effect on reliability (range, 0.863–0.892; span, 0.029). For the non-SP sample, removal of individual raters has a more variable negative effect on reliability (range, 0.577–0.719; span, 0.142), with removal of raters "g" and "j" in that sample having relatively little effect on reliability. Despite the non-SP sample range being nearly 5 times larger than in the SP sample, the difference in variances between the 2 samples is not statistically discernible, given the small sample sizes (modified robust Brown-Forsyth Levene-type test, 3.211; $P = 0.10$).

Mean ratings for SPs and non-SPs were similar as were average scores for most rubric items, suggesting that SPs were more consistent with one another than non-SPs without being more lenient or severe in their ratings. Comparisons of mean ratings by rater group (SP vs. non-SP) were run for the 19 items. Only three were significant (items 4a, 4b, and 11; Table 1), with non-SPs rating the learners higher on average (mean differences of 0.25, 0.36, and 0.19, respectively; $P < 0.05$).

To explore whether reliability was affected when SP raters viewed videos in which they had acted, we conducted additional analyses. Means (SDs) were computed for the overall group (6 raters) and subset (excluding SPs who acted with the ratee; on average, 1.5 of 6 raters are excluded) and were found to be the same [mean (SD), 10.9 (2.4)]. We also computed interrater reliability coefficients for both the full SP data set and a data set in which all instances of raters viewing their own acting were removed. The latter data set has an ill-structured measurement design (ie, it is neither fully crossed nor nested). As such, special analytic techniques as described by Putka et al[24] were used to compute a "$G(q,k)$" coefficient, where q is a correction factor for the imperfect crossing and k is the harmonic mean number of raters per learner in the ill-structured design. These coefficients were 0.93 in the full data set and 0.91 in the reduced data set, suggesting very comparable reliability regardless of whether SP raters rated cases in which they had acted or not.

DISCUSSION

This study provides 2 important findings, both with implications for measurement and evaluation in medical

education. First, even educated observers with extensive, in-depth training were less effective than SPs in the simulated patient encounters being assessed. A single SP rater is predicted to yield a reliability slightly greater than that of 4 non-SP raters.

Our findings suggest that ongoing calibration and cumulative experience may be superior for consistency in ratings to a 1-time training, even if intensive and using a simplified yes/no behavior checklist. Crowd-sourced rater recruiting strategies using content-naive raters may thus tend to yield unreliable scores in this context, and future rater training efforts might be enhanced by incorporating further practice and ongoing calibration.

Our results demonstrate that high reliability on the assessment of simulated patient encounters can be achieved without relying on additional outside observers for the evaluation of learners. This is helpful because if the SPs are able to both portray a patient role and evaluate learners, time and costs may be reduced, particularly if faculty are spared from this duty. Second, although a larger number of raters are generally expected to improve reliability, our results suggest that all raters may not be equally well calibrated. In the future, researchers and practitioners may find value in the use of these "rater-deleted" analyses, perhaps to offer extra training to specific individuals or to exclude raters who are less calibrated with the group.

Although SPs showed higher reliability than non-SPs in this study, some questions remain as to precisely why this effect was observed. We have raised the possibility that SPs' experience with the scenario and rubric and their ongoing calibration led to more elaborate memory structures for common behaviors learners display. However, other explanations are possible. For instance, it may be that the SP raters were more personally invested in the rating task and thus more careful with their observations. Further, this study does not separately estimate the benefits to reliability of role-playing practice versus ongoing rater calibration. It would be valuable to explore the reasons underlying the observed results to better understand options for improving reliability.

One limitation of this study is its use of a single scenario and associated performance assessment. Medical education assessment research has generally found that learners' performance varies depending on the particulars of any given scenario they are rated on, referred to as "content specificity" or "case specificity," such that providing multiple opportunities for learners to demonstrate performance is advantageous in rating the competencies of a learner, especially if one wishes to make relatively broad inferences about learners' knowledge or skills.[25] In the present case, however, we are most interested in comparing characteristics of raters. Although adding cases would be expected to increase the reliability of learners' scores as generated by both groups of raters, our point is that it is important to pay attention to whom the simulation developer selects as the raters.

As such, although content specificity is well established, it is not clear that our findings would change appreciably if a different scenario had been used. It would be valuable for future research to estimate how variable individual-case reliability is in OSCEs like this by computing "reliability if case deleted" coefficients, analogous to the "reliability if rater deleted" coefficients demonstrated here. The results of such research might speak to the generalizability of our findings.

In addition, the current study relied on only 2 kinds of raters (SPs and non-SPs). Had faculty observers (who were not available for a comparably lengthy rater training program) been included, for example, it is possible that with their extensive substantive knowledge of the professions and clinical practice, they would have been more accurate raters. On the other hand, the behaviors measured in the simulation were predominantly nonclinical in nature and did not call on medical expertise. In addition, although we intentionally sampled learners from a variety of health professions, we did not have a sufficiently large sample to explore the possible effect of profession on learner performance. It would be interesting in future research to explore whether learners in various health professions perform differently overall or on specific behaviors. In addition, although we found that the extensive training provided to non-SPs was not sufficient to achieve a desired level of reliability without using many raters, it would be useful to know how much the training "dosage" would relate to reliability. Furthermore, reliability is of course only 1 aspect of the broader concept of validity, and more reliable ratings may not necessarily be more valid. However, given that this study used ratings of observable behaviors, we suspect that the large improvements in reliability associated with rater type likely correspond to more valid rating as well.

CONCLUSIONS

Standardized patients made more reliable observational ratings of learners' behaviors in a simulation-based interprofessional collaboration scenario than trained raters without the additional practice and calibration associated with repeatedly practicing the scenario and rating process with other raters. Generalizability theory can provide practical guidance when making decisions about the relative merits of investing limited institutional resources in training and selecting raters, and rater-specific analyses may help identify particular raters in need of recalibration. Furthermore, we recommend using statistical hypothesis testing with generalizability theory as demonstrated in this research to better gauge the significance of comparative reliability findings.

ACKNOWLEDGMENTS

The authors would like to thank the following experts who were also involved in the development of the ATOSCE including, but not limited to: Elizabeth Andrews, DDS, MS; Sheree J. Aston, OD, PhD; Michelle Emmert, EdD; Vincent A. Finocchio; Sandra Garner, MEd; Patricia Greene, DMD; Janice Hoffman, PharmD; Elizabeth Mendoza; Mary Hudson-McKinney, DPT; Jordan Orzoff, PhD; Donna Redman-Bentley, PT, PhD; James Scott, PharmD, Sam Shimomura, PharmD, MEd; John Tegzes, MA, VMD; Valerie Wren, OD; and Jasmine Yumori, OD. We thank Anandi Law, BPharm, MS, PhD, Dr Redman-Bentley, the Simulation in Healthcare reviewers and editors for their feedback on this paper, and Dan J. Putka, PhD, for guidance on statistical software.

REFERENCES

1. Kohn LT, Corrigan JM, Donaldson MS. *To Err Is Human: Building a Safer Health System*. Washington, DC: National Academy Press; 2000.
2. World Health Organization. *Framework for Action on Interprofessional Education and Collaborative Practice*. Geneva, Switzerland: World Health Organization; 2010.
3. Zorek J, Raehl C. Interprofessional education accreditation standards in the USA: a comparative analysis. *J Interprof Care* 2013;27:123–130.
4. Liaison Committee on Medical Education. New standard ED-19-A approved. Available at: http://www.lcme.org. Accessed August 1, 2014.
5. Reeves S. The rise and rise of interprofessional competence. *J Interprof Care* 2012;26:253–255.
6. Zhang X, Roberts WL. Investigation of standardized patient ratings of humanistic competence on a medical licensure examination using Many-Facet Rasch Measurement and generalizability theory. *Adv Health Sci Educ Theory Pract* 2013;18:929–944.
7. Shirazi M, Labaf A, Monjazebi F, et al. Assessing medical students' communication skills by the use of standardized patients: emphasizing standardized patients' quality assurance. *Acad Psychiatry* 2014;38:354–360.
8. Humphrey-Murto S, Smee S, Touchie C, Wood TJ, Blackmore DE. A comparison of physician examiners and trained assessors in a high-stakes OSCE setting. *Acad Med* 2005;80:S59–S62.
9. Baig LA, Violato C, Crutcher RA. Assessing clinical communication skills in physicians: are the skills context specific or generalizable. *BMC Med Educ* 2009;9:22.
10. Newlin-Canzone ET, Scerbo MW, Gliva-McConvey G, Wallace AM. The cognitive demands of standardized patients: understanding limitations in attention and working memory with the decoding of nonverbal behavior during improvisations. *Simul Healthc* 2013;8:207–214.
11. Chen C, White L, Kowalewski T, et al. Crowd-Sourced Assessment of Technical Skills: a novel method to evaluate surgical performance. *J Surg Res* 2014;187:65–71.
12. Sagie A, Magnezy R. Assessor type, number of distinguishable dimension categories, and assessment centre construct validity. *J Occup Organ Psychol* 1997;70:103–108.
13. Dickter DN, Stielstra S, Mackintosh S, Garner S, Finocchio VA, Aston SJ. Development of the Ambulatory Team Observed Structured Clinical Evaluation (ATOSCE). *Med Sci Educ* 2013;23:554–558.
14. Zhou H, Muellerleile P, Ingram D, Wong SP. Confidence intervals and *F* tests for intraclass correlation coefficients based on three-way mixed effects models. *J Educ Beh Stat* 2011;36:638–671.
15. Shavelson RJ, Webb NM. *Generalizability Theory: A Primer*. Newbury Park, CA: Sage; 1991.
16. Brennan RL. *Generalizability Theory*. New York: Springer; 2010.
17. Aston SJ, Rheault W, Arenson CF, et al. Interprofessional education: a review and analysis of programs from three academic health centers. *Acad Med* 2012;87:949–955.
18. King HB, Battles J, Baker DP, et al. TeamSTEPPS™: Team Strategies and Tools to Enhance Performance and Patient Safety. In: Henriksen K, Battles JB, Keyes MA, et al., eds. *Advances in Patient Safety: New Directions and Alternative Approaches*. Rockville, MD: Agency for Healthcare Research and Quality; 2008.
19. Partnership for Health in Aging. Multidisciplinary Competencies in the Care of Older Adults at the Completion of the Entry-Level Health Professions Degree. Available at: http://www.americangeriatrics.org/files/documents/health_care_pros/PHA_Multidisc_Competencies.pdf. Accessed August 1, 2014.
20. Roch SG, Woehr DJ, Mishra V, Kieszczynska U. Rater training revisited: an updated meta-analytic review of frame-of-reference training. *J Occup Organ Psychol* 2011;85:370–395.
21. Boulet JR. Generalizability theory: basics. In: Everitt BS, Howell D, eds. *Encyclopedia of Statistics in Behavioral Science*. Chichester: John Wiley & Sons, Ltd; 2005:704–711.
22. DeShon RP. Generalizability theory. In: Drasgow F, Schmitt N, eds. *Measuring and Analyzing Behavior in Organizations: Advances in Measurement and Data Analysis*. San Francisco, CA: Jossey-Bass; 2001.
23. Webb NM, Shavelson RJ, Haertel EH. Reliability coefficients and generalizability theory. In: Rao CR, Sinharay S, eds. *Handbook of Statistics, Volume 26: Psychometrics*. Amsterdam: Elsevier; 2006.
24. Putka DJ, Le H, McCloy RA, Diaz T. Ill-structured measurement designs in organizational research: implications for estimating interrater reliability. *J Appl Psych* 2008;93:959–981.
25. Norman G, Bordage G, Page G, Keane D. How specific is case specificity? *Med Educ* 2006;40:618–623.