Published in final edited form as:

Med Sci Educ. 2017 September ; 27(3): 497–502. doi:10.1007/s40670-017-0418-0.

Detection of Biased Rating of Medical Students by Standardized Patients: Opportunity for Improvement

Marian Manciu^{*,1}, Roszella Trevino¹, Zuber D. Mulla², Claudia Cortez², and Sanja Kupesic Plavsic²

¹Physics Department, University of Texas at El Paso

²Department of Obstetrics and Gynecology, Center for Advanced Teaching and Assessment in Clinical Simulation, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center El Paso

Abstract

Objectives—To assess the inter-rater reliability of standardized patients (SPs) as they assess the clinical skills of medical students and to detect possible rating bias in SPs.

Methods—The ratings received by 6 students examined in 4 clinical stations by 13 SPs were examined. Each SP contributed at least 3 and at most 10 pairwise ratings, with an average of approximately 5 ratings per SP. The standard Cohen's kappa statistic was calculated and the distribution of scores among SPs was compared via both ANOVA the Kruskal-Wallis H test (one-way ANOVA by ranks). Furthermore, the number of discrepancies between pairwise raters (showing either “positive” or “negative” bias in the rating) were analyzed using ANOVA and a χ^2 goodness-of-fit test.

Results—The conventional method, which compared the statistics of kappa scores of the raters (including the prevalence-adjusted-bias adjusted kappa scores) did not reject the null hypothesis, that the raters (SPs) are similar. However, the analysis of the distribution of the discrepancies among the raters revealed that the differences between raters cannot be attributed to chance, particularly when a distinction was made between their overall “positive” and “negative” bias. A strong ($p < 0.001$) “negative” bias was detected, and the SPs responsible for this bias have been identified.

Conclusions—The statistical method suggested here, which takes into account explicitly the “positive” and the “negative” bias of the raters, is more sensitive than the conventional method (Cohen's kappa). Since the outliers (the biased SPs) affect the fairness of the grading of the medical students, it is important to detect any statistically-significant bias in the rating and to adjust correspondingly the SP's assessment.

Introduction

Standardized patients (SP) are individuals instructed to portray real patients in a uniform fashion in specific scenarios for teaching, assessment or research purposes. SPs are widely

*Corresponding author: mmanciu@utep.edu, (915) 747 7531.

employed in undergraduate medical education in the United States, and became a standard method of providing formative and summative assessment of medical students' clinical skills. SPs also play an integral role in the United States Medical Licensing Examination (USMLE)-Step 2 Clinical Skills (CS) examination [1]. The goal of this multi-station encounter exam is to evaluate the ability of medical students in history taking, physical examination, communication skills, counseling, clinical reasoning and writing patient notes after a patient encounter in the allotted time. The individual examinee's communication and clinical performance skills from 12 SP encounters are recorded by SPs using cases specific dichotomous checklist [2]. Communication and interpersonal skills and spoken English proficiency are also evaluated by SPs using generic rating scales, while patient notes are assessed by physician raters using holistic methods. Examinee must successfully complete the three components of the CS exam: Integrated Clinical Encounter (ICE compiling outcomes from history taking, physical exam and the patient note), Communication and Interpersonal Skills (CIS) and Spoken English Proficiency (SEP). The Annual Report from the National Board of Medical Examiners (NBME) shows that 34,794 USMLE Step 2 CS examinations were given in 2014–2015, and since its inception, the Clinical Skills Evaluation Collaboration, a partnership between the Educational Commission for Foreign Medical Graduates and the NBME, has administered nearly 400,000 examinations, comprising more than 4,600,000 SP encounters [3]. Although some studies reported that the use of SPs allow for the valid assessment of basic communication and clinical skills [4,5], there are no data on the reliability and interrater variability of such assessments on high stakes exams.

When SPs are used in the assessment of learner performance [6], a limited number of SPs will often evaluate a large number of learners which ensures that each learner receives equitable, uniform treatment by being exposed to the same degree of individual variability. Effective SP programs support the delivery of the medical curriculum and recruit professional SPs who have been trained in realistic and consistent case portrayal and feedback skills by tutors and discipline specific physicians. Program leadership ensures enrollment of the most appropriate category of SPs based on demographic characteristics, such as age and gender, in order to meet the needs and minimize the variability of the learning and assessment activities. Objective structured clinical examinations (OSCEs) consist of a series of “timed stations” which focus on different tasks. To minimize any possible “bias” in the SP rating students are typically rated by different SP or faculty observer who use a checklist of specific behaviors or a global rating form to objectively evaluate the student's performance.

There is some evidence in the literature that some SPs might be biased when evaluating medical students; for example, it was reported that when assessing the communication style of physicians, the reliability score of real patients was 0.78, but only 0.57 for SPs [7]. The same study found that the correlation between the mean real patient and SP scores was not strong (Spearman rank correlation=0.28) [7]. In another study, when both SPs and faculty were used to assess the communication skills and professionalism of resident physicians, the inter-rater agreement between faculty was substantial (Fleiss kappa=0.71), but the measured agreement between faculty and SPs was only fair (Fleiss kappa=0.36) [8].

Since a biased rating provided by an SP might affect the fairness of the grades received by medical students, it is imperative to detect such a bias during examinations, especially in environments which employ a relatively small number of SPs and students. Notable differences between SP programs are a consequence of different local needs (e.g., type of the facility, educational activities, curriculum, funding model, regulatory issues, etc.). Due to contextual differences among SP programs, there is a need to establish an evidence based methodology for the effective assessment of SP performance. Our study assessed the inter-rater reliability of SPs using conventional and innovative statistical methods.

Method

Our study protocol was deemed exempt from formal review by the Texas Tech University Health Sciences Center El Paso Institutional Review Board for the Protection for Human Subjects (approval number E16104).

Study sample

A sample from The Center for Advanced Teaching and Assessment in Clinical Simulation (ATACS) at the Paul L. Foster School of Medicine was analyzed. The data contained the rating of 6 students on 4 different medical conditions by 13 trained SPs. The signs and/or symptoms that the SPs were complaining of/displaying are: joint pain, fatigue, stomach pain and cough. All SPs were trained to portray these 4 case scenarios and all of them successfully completed feedback and communication training workshops in a consistent and standardized manner. At least two SPs (sometimes three) evaluated each of the 6 students for a given simulated medical condition using a checklist with 20 questions. Each of the 20 questions had a dichotomous response: “yes” or “no” and each SP was rating his/her own encounter with the student. To consistently apply our method, the five student ratings involving three SP raters have been considered as 15 independent pairwise ratings; overall, a total of 33 (15+18) pairwise ratings were available for analysis with each SP participating in at least 3 and at most 10 ratings with an average of about 5 ratings per SP.

Statistical analysis

The inter-rater agreement between two raters offering only categorical (e.g., yes/ no) ratings is commonly measured using the Cohen’s kappa coefficient, which corrects for the potential occurrence of agreement due to chance [9]. Since most of the students had been rated by 2 SPs and a few by 3 SPs, the pairwise scores were analyzed, as described above. In the situations when a definite “yes” or “no” answer was missing in the checklist (e.g., due to a malfunctioning video/audio system), that question was ignored for the purpose of statistical analysis. Although there are several classifications for the interpretation of a kappa value, it is commonly accepted that negative values indicate poor agreement, and values between 0–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1 almost perfect agreement [9].

Each SP was assigned a kappa coefficient for each pair of ratings in which she or he was involved. The hypothesis that a bias among SPs existed was investigated using the Kruskal-Wallis one-way analysis of variance (ANOVA) by ranks. This nonparametric technique was

used since kappa may not be normally distributed [10]. For the sake of completeness, we also analyzed our data using a traditional parametric one-way ANOVA with Tukey's HSD correction for multiple comparisons.

Since kappa can be influenced by the presence of a biased observer (e.g., an SP who is consistently a harsh evaluator) we attempted to identify positive or negative bias between ratings made by the SPs [11]. For this we employed two methods, the first being the one proposed by Byrt et al.[11], which is an extension of the Cohen' kappa analysis that takes into account the prevalence and the bias among raters.

In the second method, the number of discrepancies ("positive" or negative") were counted for each SP and each pairwise rating. Positive discrepancies represent a "yes" rating to a question for which the other rater said "no" ("negative" discrepancies, the opposite). The discrepancy ratings for the SPs were compared with ANOVA using Tukey's HSD correction and with the Kruskal-Wallis (KW) H test. Additionally, the expected frequencies of the total, positive and negative discrepancies were compared with the observed frequencies and a χ^2 goodness-of-fit test (df=12) was employed to see if chance was a likely explanation for the differences between the observed and expected discrepancies or if there was a bias in the ratings of the 13 SPs.

Results

Figure 1a shows that the average Cohen kappa scores of the SP raters ranged between 0.6 and 1, which indicates substantial or almost perfect agreement, for all the raters.

Both the ANOVA and Kruskal-Wallis rank tests showed that there were no statistically-significant differences between the average Cohen kappa scores of the raters in the population ($p=0.65$ and $p=0.23$, respectively, see the Table 1). The Prevalence-Adjusted Bias-Adjusted Kappa (PABAK) scores [11], plotted in Figure 1b, show no statistical significant differences between the SP raters ($p=0.29$ and $p=0.24$ via ANOVA and KW, respectively). The Bias Index [11] (Figure 1c) also is not statistically significant between raters ($p=0.14$, ANOVA; $p=0.16$, KW). Therefore, for our data, the conventional analysis of the kappa scores of the raters, including the prevalence-adjusted-bias-adjusted Cohen' kappa test (PABAK), show no statistically significant differences.

The second portion of our analysis focused on identifying SPs who exhibited a positive or negative bias, by counting the discrepancies (positive, negative and combined) for all the SPs. The average number of positive discrepancies per pairwise rating ($p=0.71$ via ANOVA; $p=0.49$ via KW, see Table 1) as well as overall discrepancies ($p=0.22$ via ANOVA; $p=0.16$ via KW) are not significantly different among the raters, however the number of negative discrepancies cannot be attributed to chance at 0.05 level ($p=0.015$ via ANOVA; $p=0.05$ via KW).

The objective of the subsequent investigation was to determine whether the occurrence of rating discrepancies could have been attributed to chance, and the total number of expected and observed discrepancies (positive, negative and combined) were compared using a χ^2 goodness-of-fit test (Figures 2a–c). The null hypothesis of a random distribution of total

discrepancies, as well as of positive discrepancies cannot be rejected ($p=0.18$ and $p=0.37$, respectively).

However, the situation is different for the negative discrepancies; in this case, all tests shows that the distribution of negative discrepancies is most likely not due to chance (ANOVA: $p=0.015$; KW: $p=0.05$; χ^2 goodness-of-fit test: $p<0.001$), the latest test being the most sensitive.

From Figure 2c, it is clear that SPs “D” and “H” tended to give more negative ratings to the students than the average SP.

Discussion

Adequate clinical skills, verbal and written communication skills are critical to effective clinical care and patient safety. The inclusion of case studies across clinical disciplines and multiple SP encounters is required to ensure objective and reliable assessment of the individual examinee’s clinical skills at the time of the USMLE Step 2 CS exam. In a typical medical school environment, because of time and resource constraints, there are a limited number of SPs who assess the clinical skills of medical students. Therefore the presence of even a few biased SPs can greatly affect the fairness of the grading. It is important that any such bias should be detected and the grade of the students should be adjusted accordingly. The analysis of our data shows that whereas the conventional comparisons of kappa scores might not detect discrepancies between raters, the method proposed here, which accounts for the overall distribution of positive, negative and combined discrepancies, found a clear “negative” bias that can be attributed to two SPs.

One can argue that the method suggested here, being more powerful in detecting the “outlier” biased SPs is prone to a larger false discovery rate. This is, however, not a critical issue since the analysis can be performed in real time (once the ratings are digitized) it is fairly simple to add additional SP encounters for the medical students examined by the “biased” SP. Even if the analysis is done “post-hoc” and such examination duplications are not feasible, the grade of the affected students can be adjusted accordingly by rescaling the number of positive (or negative) ratings of the biased raters.

Detection of biased SPs is equally important in the monitoring of the various stations during the training of SPs. The objective is to identify whether there are particular stations and/or case scenarios for which the training can be improved.

One of the major advantages of our proposed method is that it does not require a particular format for the data and hence is readily adaptable to the available clinical data. It also does not require that all the SPs should have the same number of ratings. For example, our study included 5 multiple SPs cases (3 raters per student) and 18 dual SP cases (2 raters per student); the SPs participated in 3 to 10 pairwise ratings.

To the best of our knowledge, this is the first report of the detection of biased ratings made by SPs during the evaluation of medical students. Furthermore, it was even suggested that *“the use of multiple SPs to simulate the same case have little, if any, effect on intercase*

reliability of SP examinations” [12]. This is clearly not valid for the “outliers” biased SPs, for which either an additional reexamination of the student or the adjustment of the grade is necessary. Despite a small number of ratings analyzed, it is hoped that the present study will provide useful information for more objective assessment of SP interrater variability in other programs.

A systematic review conducted by Setyonugroho and colleagues found a worldwide lack of consensus in the checklist forms that are used to assess the communication skills of undergraduate medical students [13]. Further studies are needed to assess the potential effect of biased SP ratings on the outcome of the USMLE - Step 2 CS exam and similar high stakes decision making exams in other countries. Similarly, detailed investigations of the OSCE checklists that are used to assess the skills of medical students are required.

Conclusion

When the conventional method of analyzing the kappa scores was applied to our study statistically-significant differences between SP raters were not detected. The average inter-rater agreement was at least substantial (no average kappa score was less than 0.61, and the majority of them were larger than 0.8). However, our novel approach, which takes into account the overall positive, negative and combined discrepancies between the raters, showed a statistically-significant difference among raters and was able to identify two SPs who exhibited a negative bias. The method proposed herein can easily be implemented and is more powerful in detecting bias among raters than the conventional approach that employs kappa statistics. The suggested method of detecting the “biased” outliers does not require a stringent format for data analysis and can be easily extended to cases in which multiple raters participate in an arbitrary number of pairwise ratings.

Acknowledgments

Research reported in this paper was supported by the National Institute of General Medical Sciences of the National Institutes of Health under linked Award Numbers RL5GM118969, TL4GM118971, and UL1GM118970. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Hawkins RE, Swanson DB, Dillon GF, Clauser BE, King AM, Scoles PV, Whelan GP, Burdick WP, Boulet JR, Homan AG. The introduction of clinical skills assessment into the United States medical licensing examination (USMLE): Description of USMLE step 2 clinical skills (CS). *J Med Licensure Discipline*. 2005; 91:21–5.
2. Dillon GF, Boulet JR, Hawkins RE, Swanson DB. Simulations in the United States Medical Licensing Examination™ (USMLE™). *Qual Saf Health Care*. 2004; 13(Suppl 1):141–145. DOI: 10.1136/qshc.2004.010025
3. 2015 National Board of Medical Examiners (NBCM) Annual Report (<http://www.nbme.org/publications/>).
4. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med*. 1990; 2:58–76.
5. Stillman P, Swanson D, Regan MB, et al. Assessment of clinical skills of residents utilizing standardized patients: a follow-up study and recommendations for application. *Ann Intern Med*. 1991; 114:393–401. [PubMed: 1992883]

6. Epstein RM. Assessment in medical education. *New England Journal of Medicine*. 2007; 356(4): 387–96. DOI: 10.1056/NEJMra054784 [PubMed: 17251535]
7. Fiscella K, Franks P, Srinivasan M, Kravitz RL, Epstein R. Ratings of physician communication by real and standardized patients. *The Annals of Family Medicine*. 2007; 5(2):151–8. DOI: 10.1370/afm.643 [PubMed: 17389540]
8. Dabrh AM, Murad MH, Newcomb RD, Buchta WG, Steffen MW, Wang Z, Lovett AK, Steinkraus LW. Proficiency in identifying, managing and communicating medical errors: feasibility and validity study assessing two core competencies. *BMC Medical Education*. 2016; 16(1):233.doi: 10.1186/s12909-016-0755-5 [PubMed: 27589949]
9. Szklo, M., Nieto, FJ. *Epidemiology beyond the basics*. Aspen Publishers, Inc; Gaithersburg, Maryland: 2000.
10. Vierkant, RA., editor. A SAS® macro for calculating bootstrapped confidence intervals about a kappa coefficient. Available at <http://www2.sas.com/proceedings/sugi22/STATS/PAPER295.PDF> Accessed October 27, 2016
11. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *Journal of clinical epidemiology*. 1993; 46(5):423–9. DOI: 10.1016/0895-4356(93)90018-V [PubMed: 8501467]
12. Colliver JA, Morrison LJ, Markwell SJ, Verhulst SJ, Steward DE, Dawson-Saunders E, Barrows HS. Three studies of the effect of multiple standardized patients on intercase reliability of five standardized-patient examinations. *Teaching and Learning in Medicine: An International Journal*. 1990 Jan 1; 2(4):237–45.
13. Setyonugroho W, Kennedy KM, Kropmans TJ. Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. *Patient Educ Couns*. 2015; 98:1482–1491.

Figure 1.

a. Distribution of kappa scores for SP raters. The central line (red) indicate the median and the bottom and top edges of the box indicate the 25-th and 75-th percentile, respectively. The whisker extend to the most extreme data points not considered outliers, and the outlier scores are marked with a red +.

b. Distribution of the Prevalence-Adjusted-Bias-Adjusted-Kappa (PABAK) scores of the SPs.

c. Distribution of the Bias Index of the SPs.

Figure 2.

- a.** Expected and observed total number of discrepancies for each SP.
- b.** Expected and observed positive discrepancies for each SP.
- c.** Expected and observed negative discrepancies for each SP.

TABLE 1
Summary of the results of the statistical analysis

Because kappa scores (having a maximum value of one) are typically not normally distributed, both ANOVA and Kruskal -Wallis tests have been employed

Variable	p from ANOVA	p from Kruskal-Wallis	χ^2	p from χ^2 test
Cohen' Kappa scores	0.6546	0.2342		
PABAK scores	0.2876	0.2403		
Bias Index	0.1398	0.1620		
All discrepancies	0.2247	0.1617	24.34	0.183
Positive discrepancies	0.7138	0.4887	12.92	0.3748
Negative discrepancies	0.015	0.0509	34.72	0.000518