A study on the standard setting, validity and reliability of the standardized patient performance rating scale – student version

Ipek Gonullu
Ankara University School of Medicine https://orcid.org/0000-0001-6333-6087
Celal Deha Dogan
Ankara University Faculty of Education
Sengul Erden
Ankara University Faculty of Medicine
Derya Gokmen (dgokmen2001@yahoo.com)
Ankara University Faculty of Medicine
Derya Gokmen (dgokmen@yahoo.com)
Ankara University Faculty ofMedicine

Research article

Keywords: Performance rating scale, Standard setting, Standardized patient, Student ratings, Validity

Posted Date: March 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-89934/v2

Version of Record: A version of this preprint was published at Annals of Medicine on January 30th, 2023.
See the published version at https://doi.org/10.1080/07853890.2023.2168744.

Abstract

Introduction: The quality of standardized patients' performance plays a significant role in the effectiveness of clinical skills education. Therefore, providing standardized patients with constant feedback is pivotal, and as the students interact one-on-one with standardized patients, it is essential to learn the students' perspectives immediately after the encounters. There is not a single study where a standard scale with a cut-off score is specifically developed for students to assess the quality of standardized patients' performance immediately after the encounters. The three main goals of this study include: 1) Developing a scale to measure standardized patients' performance by students; 2) Examining its psychometric properties; and 3) Doing a standard-setting to determine a cut-off score for the scale.

Materials and methods: In the scale developing process; for the pilot testing, and validation process, 702 medical students participated, whereas for the standard-setting process, seven educators took part in this study. Exploratory and confirmatory factor analyses were performed. For the standard-setting study, the extended Angoff method was utilized.

Results: As a result of the exploratory factor analysis, the scale had a single-factor structure as confirmed by confirmatory factor analysis. Cronbach's alpha internal consistency coefficient was calculated as 0.91. The scale consisted of nine items. The score of a standardized patient at the borderline was 24.11 out of 45.

Conclusions: "Standardized Patient Performance Rating Scale – Student Version" is a valid and reliable scale for assessing the performance of standardized patients by the students immediately. It will guide standardized patient trainers during their training and will enable standardized patients to assess and develop their weaknesses on an individual performance level.

Introduction

Standardized patient (SP), is an integral part of teaching "communication and clinical skills" in today's medical education although Howard Barrows was the first to use SPs in the 1960s [1,2,3]. SP programs that are flexible to implement different learning styles, provide medical students standardized learning opportunities and a learning environment in line with adult learning principles. SP encounters, which are exceptionally well suited for teaching and assessing student performance in a safe environment, allow students to overcome the fear of harming a patient before they are required to encounter actual patients [4,5,6].

It is important to monitor SPs' performance to ensure consistency as well as to increase the effectiveness of education. In order to maintain the quality of SPs' performance, a common quality assurance method is the post-assessment analysis of actual interactions recorded during the encounters [7]. Besides, the faculty is always encouraged to complete brief written evaluations of SP performances, but feedback is often omitted [8].

Faculty working with SPs [8], SP trainers, SPs [7,9] and students are the most relevant stakeholders who can assess SPs' performance. In a study, both students and faculty have evaluated SPs' performance and the results showed that the ratings were similar which suggesting that students were able to recognize the quality of constructive feedback from SPs. [10]. As students interact one-on-one with the SPs, it is essential to get feedback from the students about SPs' performance immediately after the encounters. Because more than one student assesses each performance of an SP at different times, it is possible to get comprehensive information and monitor the progress of SPs' performance through time. In addition, SPs can receive systematic feedback on their strengths and weaknesses in achieving more effective and consistent performances for student-centered learning.

Like SPs' performance, psychological constructs are complex and difficult to measure. For this reason, having several tools that measure same constructs contribute to the literature. In this way, users will be able to compare different measurement tools and choose the most suitable for them. There are some scales to evaluate SPs' performance in the literature. The literature analysis reveals that while some of them mostly focus on accurate portrayals of case specifics [11,12], some are not [9,13,14]. Some scales can be used by all stakeholders [9,10,13,14]. To the best of the authors' knowledge, there are no scales in order to be used only by students and not case specifics.

Moreover, scales that have already been developed to assess SPs' performance include 21 to 28 items [13,14] and none of them is short enough to be used repeatedly during encounters. In cases where the number of SPs is high and evaluations should be done quickly after the encounter, the use of scales with fewer items will be more effective. For this reason, a new standard scale with fewer items can be more practical.

It is crucial to set cut-off scores when developing standard scales. Standard-setting is the methodology for defining achievement and proficiency levels as well as for identifying cut-off scores corresponding to those levels [15]. If the cut-off scores are not appropriately set, the results of the assessment could be questionable. For this reason, standard-setting is a critical component of the test development process [16]. A standard scale with a cut-off score facilitates decision-making about SPs' performance.

Consequently, developing a valid and reliable scale specifically developed for students to evaluate SPs' performance in an educational setting with fewer items and implementing a standard-setting study for this scale will make a significant contribution to the literature. In this regard, the study has three main goals:

1. Developing a scale to evaluate SPs' performance by students.

2. Examine the psychometric properties of this scale.

3. Doing a standard-setting study to define a cut-off score for this scale.

Methods

This study has three phases: Phase 1 consisted of the development of the scale, Phase 2 involved the validation of the scale, Phase 3 involved standard-setting of the scale.

Participants

Two groups of participants were involved in this study: Sample 1 was used in Phase 1 and Phase 2, Sample 2 was used in Phase 3.

Sample 1

The medical curriculum in Ankara University School of Medicine (AUSM) runs a 6-year programme comprises 3 years of preclinical work followed by 3 years of clinical work (2 years of clerkships and one year's internship). Communication training with SPs in pre-clinical years is a mandatory part of the curriculum in AUSM, and SP encounters are conducted during the second and third years. For this reason, we included second and third-year medical students at 2016-2017 academic year. The criteria for participation in the study were having at least one previous SP encounter experience among volunteered students.

702 students participated to the Phase 1 study. While determining the sample size, the requirements of the multivariate data analysis methods [Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA)] used in this study, were considered. As these are the multivariate statistics, they require large sample sizes. According to Comrey and Lee, a sample size of 200 is fair, and a sample size of 300 is suitable for EFA [17]. Moreover, at least 300 cases are needed with low commonalities, a small number of factors, and just three or four indicators for each factor [18].

As EFA and CFA should be conducted with two different groups selected from the same population, we distributed the participants to each process. In the study, second year students performed the SPs' performance evaluation process earlier than third year students. Since EFA was performed earlier than CFA, EFA was done using the data of the second year students (n=307) and CFA was done using the data of the third year students (n=395).

Sample 2

The standard-setting study of the scale was carried out with a test-centered approach, following which expert opinions were collected. Experts with at least five years of experience in using and training SPs were selected by the purposive sampling method. Purposive sampling is a type of nonprobability sampling in which the researcher consciously selects specific elements or subjects for inclusion in a study to ensure that the elements will have certain characteristics relevant to the study. In addition, while selecting the experts, it was taken into account that they work in different departments of the medical school, as they may have different perspectives. For this purpose, two SP trainers and faculty from the Department of Medical Education and five faculty from the Departments of "Infectious Diseases", "Child

Health and Diseases", "Psychiatry", "Radiology" and "Forensic Medicine" participated to this phase of the study.

Development of Data Collection Tools

Phase 1

The scale development process comprised seven steps, which included a literature review, conducting interviews, synthesis of the literature review and interviews, developing items, consulting expert validation, preliminary application, and pilot testing [19].

Literature Review: At this stage, the keywords "standardized/simulated patient performance" and "standardized/simulated patient scale" were used on Web of Science, Google Scholar, and ProQuest search engines to research the relevant literature. During this stage, the two research studies focusing on the development of the measurement tools for SPs' performance evaluation were investigated [13,14] and eventually assessed.

The domains of SPs' performance were defined as follows: the ability to portray a patient, to observe the medical student's behavior, to recall the encounter, and to give feedback [20]. SPs must give accurate medical history and realistically depict the patient's educational level, psychological state, as well as emotional condition while observing the student's performance. After the interview, the SPs must recall the details of the student's behavior and give thoughtful, beneficial, and effective feedback from the standpoint of the patient SP was portraying. These conceptual definitions of the domain were decided to be measured for gauging the SPs' performance.

Conducting Interviews: In addition to the literature review, the interviews were conducted with 9 faculty and 50 students (different from the participants in Phase 1 and Phase 2), and two field experts who participated in SP training. Individual oral interviews of 15-30 minutes were made from faculty among 45 faculty who have been involved in SP selection or working with SPs for at least seven years, especially in communication skills training. They were asked what they considered to be the main attributes of the good and poor performance of an SP. They focused on the different performance characteristics in this role, such as persuasion, successful portrayal, respecting the scenario, and giving effective feedback. When the expected answers started to repeat after nine tutors, the interview was stopped. Both written and verbal answers were collected from these interviews.

Synthesis of the Literature Review and Interviews: The data from the literature review and the interviews were together evaluated with the domains of SPs' performance. As a result, the scope and content of the measurement tool intended to be developed were determined and nine indicators of performance were identified (Table 1).

Developing Items: An item pool consisting of 18 items was created, and two items were assigned to each indicator (Appendix 1) in order to prevent the narrowing of the scope of the scale in a situation where an

item was removed as a result of expert opinion or item analysis.

All developed items were positively worded. A five-point Likert-type scale was determined in consultation with the experts (three medical educators and one measurement-evaluation specialist, excluding the experts who participated in *Consulting Expert Validation*)

The response anchors of these items were defined as "poor (1)", "fair (2)", "good (3)", "very good (4)", and "excellent (5)". After taking these steps, a draft version of the scale was formed.

Consulting Expert Validation: To obtain an opinion on the 18-item draft scale, seven experts working in the field, four volunteer faculty experienced in using and training SPs, two linguists, and one of the authors, who is a measurement-evaluation specialist, were consulted. These experts examined the scale items in the context of content, scope, language, comprehensibility, measurement, and evaluation principles using an evaluation form. On the form, the experts stated their opinions on each item as "applicable," "not applicable," "needs revision" and subsequently included their recommendations for these items. Based on the recommendations from the experts, six items were excluded, and one item was revised; thus, a 12-item pilot version of the scale was created (Appendix 1).

Preliminary Application: At this stage, the scale was applied to a group of 81 students (different from the participants in Phase 1 and Phase 2), in order to determine the approximate duration of implementation, to correct any incomprehensible items, and to make changes, where necessary. As a result of the preliminary application, no item was misunderstood, none of the items were left unanswered, the instructions were comprehensible, and the evaluation of an SP took three to five minutes.

Pilot Testing: The 12-item pilot form was applied to a large group of medical students (N=702), following which the validity and reliability studies comprising *phase 2* of the study were performed. After the completion of these analyses, the scale was finalized. Relevant findings are presented in the "Data Analysis" section.

Phase 3

In this study, the extended Angoff method was used to determine the cut-off score for the scale. In this method, the experts estimated the number of scale points that they believed borderline examinees would obtain from each response item [15]. In this context, experts determined the level of performance of the SPs at the borderline by using the "Standard-Setting Form" for each item in the scale. In the "Standard-Setting Form", two sections specify the level of performance of the SP, which is at the borderline for each item. After carrying out discussions between the two sessions, the experts completed one of these sections in the first evaluation session and the other in the second evaluation session.

Data Collection

Phase 1

At AUSM the students practice interviews with SPs who are trained to act as patients with conflicts, as well as a defined medical and life history. They have regular script training sessions for learning new roles or refreshing established roles and practicing the giving of feedback. Before entering the university SP pool, all SPs signed the “SP Commitment Form” that includes confirmation that materials related to them could be used for educational or research purposes. During SP encounters, SPs give verbal feedback to the students from the patient’s perspective.

The Ethics Committee of AUSM approved the study. During the communication skills program, the students were informed about the study, and participation was voluntary. Before the interview with the SPs, two authors performed a rater training for the volunteer students in 20 minutes. The training consisted of information about the standards of SPs’ performance, how to assess SPs immediately and how to fill the scale. Consent was obtained before the interviews. Twenty-five SPs (6 male - 19 female, aged between 32 and 65 years) were utilized for the study.

The students were asked to respond to the scale immediately after the encounters. Each student evaluated the SP he/she interviewed one time during the communication skills program only. Care was taken for the students to complete the scale alone without any interaction with their peers or others.

Phase 3

First, the experts were trained by one of the authors who is a specialist in the field of measurement and evaluation. During this training, the aim and methods used in the standard-setting were explained, and information was given about the procedures to be performed on the standard-setting method. In the next step, the experts discussed the characteristics they considered should be present in the SP and agreed on the level of competence that an SP at the borderline should have. After the first evaluation session began, the experts were asked to give the scores (min: 1, max: 5) for each item by an SP at the borderline. This process was carried out individually. They were subsequently asked to share their evaluations with the group and justify them. Then, the experts stated their opinions about each other's evaluations, which were followed by a discussion. In the second session, experts were asked once again to provide the scores for the SP at the borderline. The duration of the first session, the discussion, and the second session was approximately one hour. The reason for using two rounds in the Angoff method is to reduce the deviations in the evaluations and to obtain results that are more reliable. The participants discussed their evaluations after the first round. The participants are given the opportunity to make changes in their evaluations in the second round because of these discussions if they deem it necessary.

Data Analysis

Phase 2

The Validity of the Scale

Construct Validity: Before EFA, Kaiser-Meyer-Olkin (KMO) and Bartlett test of sphericity were examined and the data was tested for its appropriateness for the factor analysis. Bartlett's test of sphericity was

applied to determine if the correlation matrix is different from the identity matrix. The statistical significance of the calculated chi-square value in Bartlett test can be interpreted as the data is appropriate for factor analysis [21].

The principal components method was used in the factor selection process. In the determination of the number of factors, scree plot and parallel analysis methods were used. Since the scale had a single factor structure, no rotation method was used.

In the CFA process, as the data did not satisfy the multivariate normality assumption, the analysis was carried out based on the weighted least squares method, and the standardized coefficients, corresponding t values, and some fit indices were evaluated. While the ratio of chi-square value to the degree of freedom of below 2.5 indicates the perfect fit, the corresponding values for Non-Normed Fit Index (NNFI), Comparative Fit Index (CFI), Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI) are above 0.95. For The Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR), values below 0.05 shows perfect fit [22].

Item Discrimination: In order to assess the item discrimination, the significance of the differences between the scores of the participants in the upper and lower 27% groups for each item was compared using the Mann -Whitney U-test. Since the scores given for each item were within the ranking level, parametric tests (t-test for independent groups, etc.) were not used in this comparison.

Reliability of the Scale

To assess the reliability of the developed scale, Cronbach's alpha [23], internal consistency coefficient, and split-half reliability coefficient [24] were calculated. The test-retest reliability coefficient could not be calculated because it was practically impossible to reach the participants twice. Since a large number of medical students evaluated SPs in this study, inter-rater reliability was not calculated because it would not be practical [25].

Phase 3: Standard Setting of the Scale

The methods commonly used in standard settings can be classified as test-centered and exam-centered. Angoff, which is a test-centered standard-setting approach, is a widely used and practical method. With this method, a cut-off score can be determined before the test is administered. In examinee-centered methods (e.g., borderline regression method, constricting groups method), the cut-off score is determined after the test is applied. In order to use this method, the experts involved in the standard-setting process need to be familiar with all the SPs because they have to classify those people as successful and unsuccessful. Since the experts in this study did not know SPs well enough, a decision was taken to use the Angoff method, a test-centered approach. In addition, test-centered and examinee-centered standard-setting methods give similar results, if applied correctly.

An adaptation of the Angoff method for items with more than two possible scores is called the extended Angoff method [26]. Candidates at the borderline are those at the sufficient-insufficient border and those who are considered barely sufficient. Using the extended Angoff method, the experts decided the scores for each item of the SP at the borderline and recorded these estimates. The mean of the estimates given by the experts was calculated for each item. The sum of means gave the cut-off score.

During the data analysis process, SPSS 21.0, Lisrel 8.7, Excel 2016, and Monte Carlo PCA for Parallel Analysis packages were used.

Results

Phase 1: Development of the Scale

Development of the scale was described in detail based on the seven-step process presented by the Association of Medical Education in Europe Guide at the “Development of Data Collection Tools”. Upon completion of the preliminary application, which is the sixth step of the above mentioned process, 12-item pilot form was created (Table 2).

Phase 2: Validation of the Scale

The Validity of the Scale

Construct Validity

Exploratory Factor Analysis: EFA was carried out on 307 participants. The KMO value for this study was calculated as 0.92. The results of the Bartlett test indicate that a chi-square value of less than 0.05 is significant, which, in turn, shows that the data is appropriate for factor analysis [27].

As a result of EFA, the scale had three factors with an eigenvalue greater than 1. When more than 200 samples are reached, it is recommended to examine the scree plot in accordance with the eigenvalues for determining the number of important factors [28]. There is a significant deceleration in the Scree plot after the first factor, and the rate of deceleration decreases and follows a horizontal course after the second factor. Moreover, the eigenvalue of the first factor (5.432) is approximately five times greater than the eigenvalue of the second factor (1.082) before the rotation. The first factor alone, yielding a high variance (45%), was interpreted as the scale having a single-factor structure. The parallel analysis for the factor number also supports the single-factor structure. The Scree Plot is presented in Fig 1.

Fig. 1 Line chart of eigenvalues

After establishing that the scale had a single-factor structure, the analysis was performed once again. The single-factor structure explained 59% of the total variance. Since the scale had a single-factor structure, no rotation was performed. In the analysis process, three items with a factor loading value below 0.40 (item 4, 0.19; item 5, 0.18, and item 7, 0.23), were excluded starting from the item with the

smallest factor loading (Appendix 2). The nine items and the factor loading values for the rest of the scale were presented in Table 3.

According to the result, the scale had a single factor structure and consisted of nine items. Its name is "Standardized Patient Performance Rating Scale – Student Version (SPS-S)"

Confirmatory Factor Analysis: CFA was performed for the verification of the single-factor structure resulting from the EFA of SPS-S developed within the scope of the study. In CFA analysis, the latent variable is the SPs' performance. This latent variable is abbreviated as "PERFORMAN." Observed variables are items of SPS-S abbreviated as S1 to S9. All variables entered into the model are displayed in Appendix 3.

As a result of CFA on the single-factor structure of SPS-S, the t values of the latent variables related to observed variables were greater than the critical value (2.58) and statistically significant at the 0.01 level. Appendix 3 shows the standardized coefficients and t values for the relationships in the model, respectively.

In the analysis, the software suggested that the errors related to the items s1 and s2 should be associated and that this association could result in a decrease of 26.91 in the chi-square value. This theoretically reasonable modification was accepted, and errors of items s1 and s2 were associated. After that, the ratio of chi-square value to the degree of freedom was calculated as 1.81. This value can be considered as an indicator of a perfect fit [18]. Other calculated indices are as follows: NNFI (Non-Normed Fit Index) 0.97, CFI (Comparative Fit Index) 0.98, GFI (Goodness of Fit Index) 0.97, AGFI (Adjusted Goodness of Fit Index) 0.96, RMSEA (The Root Mean Square Error of Approximation) 0.04, and SRMR (Standardized Root Mean Square Residual) 0.04. When the results are examined, the values of all fit indices are indicative of a perfect fit [22].

Item Discrimination: For each item included in the scale, the mean ranks were higher in favor of the group in the upper 27%, and these differences were significant ($p < 0.01$).

Reliability of the Scale

Cronbach's alpha internal consistency coefficient was calculated as 0.91, and the split-half reliability coefficient as 0.87. These findings show that the internal consistency coefficient of the scale is at the desired level. The Cronbach Alpha coefficient of 0.80 and above indicates that the test has a high level of internal consistency [29].

Phase 3: Standard Setting of SPS-S

Seven experts were consulted in order to determine the cut-off score of SPS-S (nine items. The experts were asked to use the extended Angoff method and give a score between 1 and 5, taking into account an SP at the borderline for each item. This scoring was undertaken in two rounds (R), as shown in Table 4.

When the cut-off scores for each item were examined after the second round, the experts allocated the lowest cut-off point to item 4 (2.857) and the highest cut-off point to item 9 (4.14). Also, the experts stated that an SP at the borderline should have an average of 3.44 points from each item. The standard deviation values were examined to determine the variability between the experts' scores, and the variability between expert opinions was less in round 2 (0.53) than in round 1 (0.59).

The mean of the total points of each expert for each item was taken to calculate the cut-off score. Then, these means were summed to obtain the cut-off score.

Cut-off score =3.00+3.78+3.56+3.33+4.00+3.11+3.33

Cut-off score = 24.11

Therefore, according to the findings from the extended Angoff method, for an SP to be sufficient, he/she must obtain at least 24.11 out of 45 from SPS-S.

Discussion

The authenticity of role-playing and the quality of feedback provided by SPs is of high importance for the quality of learning during SP contact learning sessions [13,14]. In this sense, assessing the individual performance of SPs in an educational setting and assisting them in decreasing their weaknesses can improve the quality of the clinical skills training program. In addition, Objective Structured Clinical Examination (OSCE) uses SPs who are subject to many measurement errors [30]. As a result, it is important to consider the training of SPs and monitoring their performance and development. The performance of SPs may only be assessed from the recorded videos mostly by SP trainers and sometimes by faculty or students. On the other hand, assessing the performance of individual SPs from the videos systematically and giving feedback on their performance needs time. Furthermore, it is very difficult to assess all the records. We might be able to solve this problem by involving the students in the assessment process of the SPs' performance immediately after the encounters. Real-time feedback for SPs to improve their work can be more easy and useful.

In this study, we presented a unique evaluation tool filled out by medical students, immediately after their encounters with SPs. SPS-S was developed, and validity, reliability, and standardization studies were performed. As a result of EFA, it was determined that the scale had a single-factor structure confirmed by CFA. The internal consistency of the scale was shown to be at the desired level by the reliability analysis. The scale consists of nine items scored out of 5, implying that the lowest achievable score is 9 and the highest is 45. The cut-off score of an SP was 24.11 out of a total of 45 determined by the extended Angoff method. In this context, for someone to successfully qualify as an SP, he/she must obtain at least 24 points from SPS-S.

Literature supports that students can adequately assess the value of the education they receive [31] and they are critical stakeholders in medical education for which their engagement is a vital component

[32,33]. They offer a unique perspective that adds value to curricular issues and intangibles of the learning environment which may be opaque to educators [32,33]. Involving the students as key stakeholders in their education can have a profound impact on students and the institutions that serve them [32]. During the assessment of the SPs' performance, straightforward and intuitive perspectives could be valuable and essential to consider as students interact with SPs one-on-one, their instant. The assessment of the SPs' performance by students might have certain advantages compared to the assessment done by other stakeholders. For example, several students can assess SPs, whereas only a few faculties and even fewer SP trainers can. Furthermore, students can assess SPs at different times, so it is possible to monitor SPs' performance progress through time. Moreover, this scale can be used to identify SPs who need further training early on by picking out the ones who scored less than 24, which will bring efficiency to the SP training and development process.

In order to determine the performance of an SP is qualified or not, there should be at least more than one student evaluation. Although it is not known exactly how many interviews will be evaluated, as many assessments as possible should be done which would be better before a judgment is made related to SP qualifications. This number could be clarified in future studies.

The strength of this study is that the SPS-S is not a case-specific scale and can be used for various scenarios. Due to the low cut-off score, the weak and strong performance of SPs can be defined easily and immediately. We recommend that if it is possible, SPS-S must be used by students immediately following an encounter for time-saving purposes because it is short. However, if it is not convenient to assess immediately it can be conducted at a later time point as well. It can easily be completed by students in three to five minutes. SPS-S was personalized for students; it contains items relating to their interactions with SPs and it has a cut-off score, which is a critical component of the test development process.

One of the limitations of the study is that the criterion validity cannot be tested, because there is no other reliable, valid and short scale to assess the SPs' performance only by students. However, in similar studies, SPS-S can be used as a criterion scale to test validity. The other limitation is instead of having two separate groups consisting of mixed of second and third year medical students; we used the second year medical students in the EFA calculation process and the third year medical students in the CFA calculation process. This decision was based on the assumption that second and third year medical students have very similar characteristics. Besides, second year medical students performed the SPs' performance evaluation process earlier than third year medical students did.

Since all experts expressed very similar opinions during the content validity examination of the scale, the content validity index and ratio were not calculated. However, the inability to calculate this index can be considered as a limitation of the study.

Another limitation can be the presence of a recall bias amongst students during their evaluation of the SPs' performance. For this reason, the scale may not be used after the summative assessments. Over

half of the questions on the scale represent the SPs ability to give feedback. This limits the utility of the SPS-S to only formative role-playing sessions, and cannot be used in scenarios such as OSCEs in which students are not immediately given verbal feedback at the end of the session. Besides, in addition to the nine items of SPS-S, there was no space for narrative assessment. This could be more helpful for the SP to learn areas in need of improvement.

Conclusion

To ensure the educational quality of the program where SP encounters take place, evaluation of the SPs' performance is important. This study presents a unique addition to SP training by introducing a student evaluation tool of the SPs' performance immediately after the encounters. The use of SPS-S, which has been confirmed for validity, reliability, and standard-setting studies, will guide SP trainers during their SP training and continued education post-training. It will also help SPs to assess and develop their weaknesses on an individual performance level. For further studies using SPS-S, we recommend the researchers to re-assess validity and reliability using CFA and internal consistency coefficient. In addition, the scale can be modified for other stakeholders who will use it to assess the SPs' performance. In conclusion, students can use this scale for the evaluation of SPs in the field of health sciences.

Declarations

Acknowledgement : Not applicable

Declaration of interest: The authors declare that they have no competing interests

References

1. Barrows HS. Simulated (standardized) patients and other human simulations: A comprehensive guide to their training and use in teaching and evaluation. Chapel Hill, NC: Health Sciences Consortium;1987
2. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med. 1993;68:443-443.
3. Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE Guide No 42. Med Teach. 2009;31(6):477-486.
4. Talwalkar JS, Cyrus KD, Fortin AH. Twelve tips for running an effective session with standardized patients. Med Teach. 2020;42(6):622-627.
5. Becker KL, Rose LE, Berg JB, Park H, Shatzer JH. The teaching effectiveness of standardized patients. J Nurs Educ. 2006;45(4):103-111.
6. Bokken L, Rethans JJ, Jöbsis Q, Duvivier R, Scherpbier A, van der Vleuten C. Instructiveness of real patients and simulated patients in undergraduate medical education: a randomized experiment.

Acad Med. 2010;85(1):148-154.

7. Nestel D, Bearman M. (Eds.) Simulated Patient Methodology: Theory, Evidence and Practice. Chichester, UK: John Wiley & Sons, Ltd.; 2015.
8. Nestel D, Tabak D, Tierney T, Layat-Burn C, Robb A, Clark S, et al. Key challenges in simulated patient programs: An international comparative case study. BMC Med Educ. 2011;11:69.
9. Perera J, Perera J, Abdullah J, Lee N. Training simulated patients: evaluation of a training approach using self-assessment and peer/tutor feedback to improve performance. BMC Med Educ. 2009;9(1):37.
10. Himmelbauer M, Seitz T, Seidman C, Löffler-Stastka H. Standardized patients in psychiatry – the best way to learn clinical skills? BMC Med Educ. 2018;18:72.
11. Baig LA, Beran TN, Vallevand A, Baig ZA, Monroy-Cuadros M. Accuracy of portrayal by standardized patients: results from four OSCE stations conducted for high stakes examinations. BMC Med Educ. 2014;14(1):97.
12. Erby LA, Roter DL, Biesecker BB. Examination of standardized patient performance: accuracy and consistency of six standardized patients over time. Pat Educ Couns. 2011;85(2):194-200.
13. Wind LA, Van Dalen J, Muijtjens AM, Rethans JJ. Assessing simulated patients in an educational setting: the MaSP (Maastricht Assessment of Simulated Patients). Med Educ. 2004;38(1):39-44.
14. Bouter S, van Weel-Baumgarten E, Bolhuis S. Construction and validation of the Nijmegen evaluation of the simulated patient (NESP): assessing simulated patients' ability to role-play and provide feedback to students. Acad Med. 2013;88(2):253-259.
15. Cizek GJ, Bunch M. Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests. USA: Sage Publications; 2007.
16. Bejar II. Standard setting: What is it? Why is it important? R&D Connections, 2008;7:1-6.
17. Comrey A, Lee H. A first course in factor analysis. Hillsdale, NJ: Lawrence Erlbaum Associates; 1992.
18. Tabachnick BG, Fidell LS. Using multivariate statistics. Pearson Education Limited; 2013.
19. Artino Jr AR, La Rochelle JS, Dezee K J, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No. 87. Med Teach. 2014;36(6):463-474.
20. Wallace P. Coaching standardized patients: For use in the assessment of clinical competence. New York: Springer Publishing Company; 2006.
21. Çokluk Ö, Şekercioğlu G, Büyüköztürk Ş. Multivariate statistical SPSS and LISREL applications for social sciences. Ankara: Pegem Academy publications; 2010.(Original work published in Turkish).
22. Hooper D, Coughlan J, Mullen MR. Structural equation modelling: Guidelines for determining model fit. Electron J of Business Research Methods. 2008;6(1):53-60.
23. Pallant J. SPSS survival manual. UK: McGraw-Hill Education; 2016.
24. Cronbach LJ. Essentials of physiologicaltesting. New York: Harper & Row Publisher; 1984.
25. Ozarkan, HB, Doğan, CD. A Comparison of Two Standard-Setting Methods for Tests Consisting of Constructed-Response Items. Eurasian Journal of Educational Research 2020;90:121-138.

26. Hambleton RK, Plake BS. Using an extended Angoff procedure to set standards on complex performance assessments. Appl Meas Educ. 1995;8(1):41-55.
27. Field AP. Discovering statistics using SPSS. 3rd ed. London: Sage Publications; 2009.
28. Stevens JP. Applied multivariate statistics for the social sciences. 5th ed. NY: Routledge; 2009
29. McDonald RP. Test theory: A unified treatment. Lawrence Erlbaum Associates Publishers; 1999.
30. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective. Med Teach. 2013;35(9):e1437-e1446.
31. Nasser F, Hagtvet KA. Multilevel analysis of the effects of student and instructor/course characteristics on student ratings. Res High Educ. 2006;47:559–590.
32. Geraghty JR, Young AN, Berkel TDM, et al. Empowering medical students as agents of curricular change: a value-added approach to student engagement in medical education. Perspect Med Educ (2020) 9:60–65.
33. Burk-Rafel J, Jones RL, Farlow JL. Engaging learners to advance medical education. AcadMed. 2017;92(4):437–40.

Tables

Table 1: Nine indicators pointing the domains of SP performance

	INDICATORS	DOMAINS
1.	Persuasiveness of acting	To ability to portray a patient
2.	Portraying a patient	
3.	Acting according to the scenario	
4.	Observing the student performance	To observe medical student's behavior
5.	Recalling the encounter	To recall the encounter
6.	Using communication skills while giving feedback	To give feedback
7.	Competency in giving feedback	
8.	Efficacy of the feedback	
9.	Professional attitude while giving feedback	

Table 2: Pilot version of the scale

	Pilot version items
1	The standardized patient plays the role realistically.
2	The standardized patient’s role is understandable.
3	The standardized patient’s answers are appropriate to the questions
4	The appearance of the standardized patient fits his/her role.
5	The standardized patient uses appropriate language when giving feedback.
6	The standardized patient incentivizes me to ask questions during the feedback session.
7	The standardized patient listens to me while giving feedback.
8	During the feedback session, the standardized patient communicates how he/she felt as a patient during the interview.
9	During the feedback session, the standardized patient gives specific examples from the interview
10	During the feedback session, the standardized patient gives remedial feedback.
11	The standardized patient gives feedback in a kind manner.
12	The standardized patient’s feedback is relevant to my performance.

Table 3: SPS-S Items and Factor loading values

	SPS-S Items	Factor loading value
1	The standardized patient plays the role realistically.	0.729
2	The standardized patient’s role is understandable.	0.744
3	The standardized patient’s answers are appropriate to the questions	0.789
4	The standardized patient incentivizes me to ask questions during the feedback session.	0.720
5	During the feedback session, the standardized patient communicates how he/she felt as a patient during the interview.	0.773
6	During the feedback session, the standardized patient gives specific examples from the interview	0.723
7	During the feedback session, the standardized patient gives remedial feedback.	0.812
8	The standardized patient gives feedback in a kind manner.	0.815
9	The standardized patient’s feedback is relevant to my performance.	0.824

Table 4: Expert Ratings Obtained from the Extended Angoff Method

Expert ID		Item Number									
		1	2	3	4	5	6	7	8	9	Row Means(SD)
1	Round 1	3	4	4	3	3	2	3	4	4	3.33 (0.70)
	Round 2	3	3	2	2	3	3	3	4	4	3.00 (0.70)
2	Round 1	4	3	4	3	4	3	4	3	4	3.56 (0.52)
	Round 2	4	4	3	3	4	4	4	4	4	3.78 (0.44)
3	Round 1	4	4	4	2	3	3	4	3	4	3.44 (0.72)
	Round 2	4	4	4	3	3	3	4	3	4	3.56 (0.52)
4	Round 1	3	4	4	3	4	3	3	3	4	3.44 (0.52)
	Round 2	4	4	3	3	3	3	3	3	4	3.33 (0.50)
5	Round 1	4	4	4	4	4	5	4	4	5	4.22 (0.44)
	Round 2	4	4	3	4	4	4	4	4	5	4.00 (0.50)
6	Round 1	3	3	4	2	2	2	3	2	3	2.67 (0.70)
	Round 2	3	3	4	2	3	3	3	3	4	3.11 (0.60)
7	Round 1	3	3	3	4	4	3	4	3	4	3.44 (0.52)
	Round 2	3	3	3	3	4	3	4	3	4	3.33 (0.50)
Round 1 Means		3.43	3.57	3.86	3.00	3.43	3.00	3.57	3.14	4.00	3.44 (0.59)
Round 1 SD		0.53	0.53	0.38	0.82	0.79	1.00	0.53	0.69	0.58	
Round 2 Means		3.57	3.57	3.14	2.86	3.43	3.29	3.57	3.43	4.14	3.44 (0.53)
Round 2 SD		0.53	0.53	0.69	0.69	0.53	0.49	0.53	0.53	0.38	

andard Deviations , **Round 1 Means**: Column means for round 1, **Round 1 SD** : Column standard deviations for round 1, **Round 2** s: Column means for round 2, **Round 2 SD** : Column standard deviations for round 2

Figures

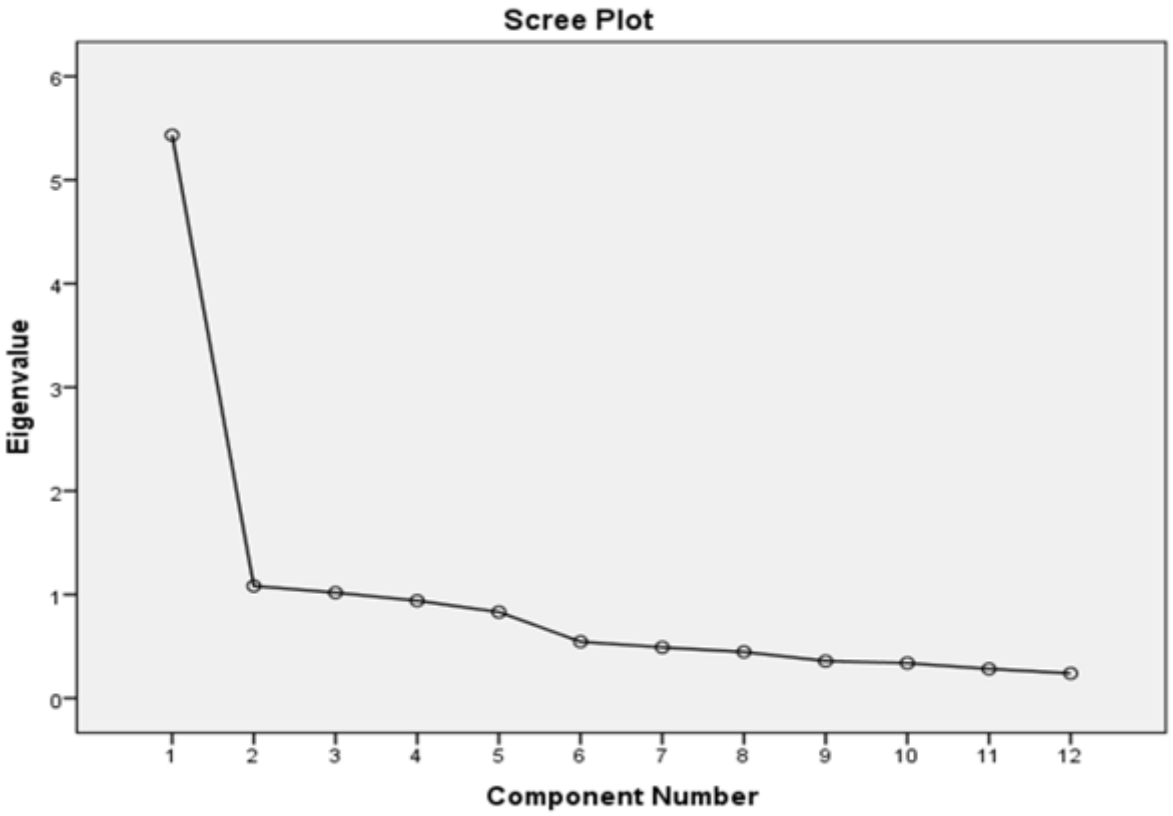

Figure 1

Line chart of eigenvalues