Uluslararası Sağlık Bilimlerinde Disiplinler Arası Etkileşim Dergisi
Cilt 2, Sayı 2, 24 - 10, 2023

International Journal of Interdisciplinary Interaction In Health Sciences
Issue 2, Sayı 2, 24 - 10, 2023

Comparison of Different Raters Scores in Simulated Patient Training Program

Simüle Hasta Eğitim Programında Farklı Değerlendirici Puanlarının Karşılaştırılması

Giray KOLCU [1,2,3*], **Mukadder İnci BAŞER KOLCU** [1,2]

[1] Suleyman Demirel University, Faculty of Medicine, Department of Medical Education and Informatics, Isparta, Türkiye
[2] Girne American University, Faculty of Medicine, Department of Medical Education and Informatics, Kyrenia, Turkish Republic of Northern Cyprus
[3] Suleyman Demirel University, Health Science Instution, Isparta-Türkiye

ABSTRACT

Introduction: Simulated patient applications play a pivotal role in modern medical education, providing a safe space for students to practice and receive feedback. This study delves into the assessment of these applications using various raters and explores the inter-rater reliability, focusing on self-assessment, peer assessment, simulated patient assessment and trainer assessment methods.

Methods: This study used a comparative quantitative research design in which 34 students participated in simulated patient interviews and were evaluated by four groups of evaluators. The study analyzed the scores using ANOVA tests and calculated inter-rater reliability using Intraclass Correlation Coefficient (ICC), Cohen's Kappa, Fleiss' Kappa, and Light's Kappa statistics.

Results: The study revealed distinct assesment patterns: self-assessment and peer assessment were generous, providing diverse feedback. Simulated patient and trainer assessment yielded consistent results within a model and after standardized training. ANOVA analysis confirmed significant differences between the groups ($p<0.001$). ICC and Kappa values varied, emphasizing the importance of clarity in training and evaluation criteria.

Discussion: Despite limitations due to the small sample size, the study emphasized the need for multifaceted assessments in simulated patient practices. Compared with the evaluators' scoring, it was evaluated that the training received by the simulated patients before the assessment and the discussions on scoring were appropriate for standardization. The results of our study revealed the necessity of clear instructions and standardization trainings to increase the reliability and validity of assessment and evaluation in the process of evaluating simulated patient interviews with the developed rubric, and the feasibility of using various perspectives such as self-assessment and peer assessment. Future studies can further develop assessment methodologies to ensure comprehensive and objective evaluation of simulated patient practices.

Keywords: Medical Education, Simulated Patient Applications, Peer Assessment, Self-Assessment

ARAŞTIRMA MAKALESİ

*Corresponding author: giraykolcu@sdu.edu.tr

ÖZET

Giriş: Simüle hasta uygulamaları, modern tıp eğitiminde önemli bir rol oynamakta olup öğrencilere pratik yapma ve geri bildirim imkânı sunar. Bu çalışma, bu uygulamaların değerlendirilmesini ele almakta ve öz-değerlendirme, akran değerlendirmesi, simüle hasta değerlendirmesi ve eğitmen değerlendirmesi yöntemlerini içeren çeşitli değerlendirme yöntemlerinin arasındaki değerlendirme güvenilirliğini incelemektedir.

Yöntem: Bu çalışma karşılaştırmalı bir nicel araştırma tasarımı kullanılmış, 34 öğrenci simüle hasta görüşmelerine katılmış ve dört grup değerlendirici tarafından değerlendirilmiştir. Çalışma, ANOVA testlerini kullanarak skorları analiz etmiş ve Intraclass Correlation Coefficient (ICC), Cohen's Kappa, Fleiss' Kappa ve Light's Kappa istatistiklerini kullanarak değerlendirme güvenilirliğini hesaplamıştır.

Sonuçlar: Çalışma, farklı değerlendirme kalıplarını ortaya koymuştur: öz-değerlendirme ve akran değerlendirmesi cömert olup çeşitli geri bildirimler sunmuştur. Simüle hasta ve eğitmen değerlendirmesi ise bir model kapsamında ve standardize bir eğitim sonrasında uyumlu sonuçlar vermiştir. ANOVA analizi gruplar arasında önemli farklılıkları doğrulamıştır ($p<0,001$). ICC ve Kappa değerleri farklılık göstermiş, değerlendirme kriterlerindeki açıklığın önemini vurgulamıştır.

Tartışma: Küçük örneklem büyüklüğünden kaynaklanan sınırlamalara rağmen, çalışma simüle hasta uygulamalarında çok yönlü değerlendirmelerin gerekliliğini vurgulamıştır. Değerlendiricilerin puanlamaları ile karşılaştırıldığında ölçme ve değerlendirme öncesinde simüle hastaların aldıkları eğitim ve puanlama üzerine yürütülen tartışmaların standardizasyon için uygun olduğu değerlendirilmiştir. Çalışmamızın sonuçları; simüle hasta görüşmelerinin hazırlanan rubrik ile değerlendirilmesinde ölçme ve değerlendirme güvenilirliğini ve geçerliğini artırmak için açık talimatların ve standardizasyon eğitimlerinin gerekliliğini, öz-değerlendirme ve akran değerlendirmesi gibi çeşitli bakış açılarının kullanılabilirliğini ortaya koymuştur. Gelecekteki çalışmalar, simüle hasta uygulamalarının kapsamlı ve objektif değerlendirmesini sağlamak için değerlendirme metodolojilerini daha da geliştirebilir.

Anahtar Kelimeler: Tıp Eğitimi, Simüle Hasta Uygulamaları, Akran Değerlendirmesi, Öz-Değerlendirme

1. Introduction

Simulated patient applications are interactive and realistic educational tools used for the development of practical skills for medical students and healthcare professionals (1). These applications are specifically designed to provide users with experience before performing professional skills on/with real patients (1,2). Simulated patient applications include virtual patients imitating various clinical scenarios and conditions (3). These applications offer various opportunity for students and experts like; to practice clinical examination skills, diagnostic abilities, intervention skills for emergencies or routines, and critical skills such as communication, manangement and clinical thinking skills (4). Simulated patient applications hold great importance in terms of "patient safety" (5,6). They reduce the risk of conducting experiments on real patients while providing the chance to make mistakes and receive feedback during the learning process (1). These applications enhance patient safety by providing medical students and

healthcare workers with a safe environment to practice. Additionally, simulated patient applications contribute healthcare professionals to be more competent and reliable in patient care (4,7). While providing a realistic experience, these applications also improve the learning process and enhance the quality of healthcare services (7,8). These applications help future healthcare professionals make better decisions in patient care, effectively manage emergency situations, and improve patient communication skills (1,6–8). Using simulated patients in health and medical education create an important tool for medical students and healthcare professionals to enhance practical skills and increase patient safety (4,5,9). These applications offer students to have an opportunity to practice in a realistic/authantic environment with also having an opportunity to make mistakes and allowing them to learn from their mistakes and develop by receiving feedback (6–8). The widespread use of simulated patient applications contributes to improving the quality of healthcare services and providing better patient care.

Süleyman Demirel University Faculty of Medicine is one of Turkey's institutions for medical education. In 2019, the 'Simulated Patient Laboratory' was established within the Faculty of Medicine at Süleyman Demirel University. However, after the establishment of the laboratory, the Covid-19 pandemic occurred and formal education was moved to virtual education. Following the COVID-19 pandemic, in the academic year 2022-2023, the implementation of simulated patient training program was initiated to enhance the communication skills of 3rd-year students during patient interactions. A comprehensive approach has been adopted for the assessment of students' interactions within these applications. In these assessments, the feedback process has been enriched through self-assessment, peer assessment, assessment by simulated patients, and assessment by medical educators.

Self-assessment refers to participants evaluating their own performance, identifying strengths and weaknesses, and determining opportunities for future improvement (10). In simulated patient trainings, this assessment process enables students and healthcare professionals to critically examine their own skills, levels of knowledge, and professional behaviors. Self-assessment encourages students and healthcare professionals to realistically evaluate their own performance, while also promoting critical thinking and the development of reflective practice skills (10,11). With the scope of watching the recorded videos of their performance of simulated patient interview, students were able to self assessed themselves. After training with simulated patient, participants have the opportunity to observe and analyze themselves. This enables them to recognize their strengths, identify areas for improvement, and proceed with a more conscious approach towards setting personal learning goals. The self-assessment process allows participants to provide themselves with effective and instructive feedback. The ability to assess one's own performance helps students and healthcare professionals effectively manage their learning process. Moreover, the self-assessment process enhances motivation among participants to reach learning objectives and consistently improve their professional development (10,11). By promoting personal and professional growth, the self-assessment process supports students and healthcare professionals in delivering patient care more competently. In simulated patient applications, self-assessment becomes a crucial tool that enriches the learning experience and facilitates participants' continuous progression towards self-improvement (11).

Peer assessment involves students and healthcare professionals providing feedback to each other by observing and supporting each other in a collaborative learning environment (12). There are many studies related to peer education in medical education (12,13). Peer assessment in simulated patient practices provides participants with the opportunity to collaborate, conduct skill analyses, and work on improvement together (12). Peer assessment plays a significant role in simulated patient practices because it allows participants to objectively evaluate each other's performances with different perspectives and experiences and also learn from others performances (13). This assessment process assists students and healthcare professionals in identifying their strengths and areas for improvement. Additionally, Additionally, receiving peer feedback before practicing with or on real patients enables participants to learn from mistakes and enhance their skills in a safe environment (10,11). Peer assessment encourages development in critical areas such as communication skills, collaboration abilities, and professional conduct (12,13). Peer reviews conducted in simulated patient practices help

healthcare professionals recognize their strengths and weaknesses and continually improve themselves (12–14).

With the scope of Simulated Patients Training programs in our curriculum, all simulated patients enrolled a certificated educational program. In the context of this program, they also took a course about observing and assessing the students performance and giving feedback through the observation. The Calgary-Cambridge Consultation Model (Guide) was followed as an educational resource. The educators from Department of Medical Education and Informatics created a new rubric based on this resource. Both students and simulated patients there informed about this rubric and the principles of using it. In addition, simulated patient interviews are also scored by medical educators in our faculty and another feedback from a different perspective is also given in accordance with this scoring.

In this process, a need for analysis has emerged regarding the evaluation of raters for the overall evaluation of training effectiveness. “Inter-rater reliability” refers to the consistency between different raters in an evaluation or scoring process. This evaluates the agreement and similarity between the results of different raters scoring the same samples using the same evaluation criteria. Inter-rater reliability is important for the objectivity and reliability of the evaluation process. High inter-rater reliability means that different raters give similar or consistent results when evaluating the same samples (15). This indicates that the evaluation process is reliable and there is agreement between different raters. Inter-rater reliability ensures that the evaluation process is free from subjective judgment and that the evaluation is made objectively and fairly. Additionally, inter-rater reliability ensures the consistency and validity of the evaluation process, making the results more reliable and stable.

One of the most common methods used to calculate inter-rater consistency in the literature is the "Intraclass Correlation Coefficient" (ICC) (15–17). ICC measures the consistency and agreement between raters when different raters evaluate the same samples (16). ICC calculates inter-rater reliability by taking into account different variance components (17,18).

The Kappa statistic is also used to calculate inter-rater consistency. (19). Kappa statistics is a statistical method used to measure the agreement between two or more raters when evaluating data on a nominal or ordinal scale (19,20). The kappa statistic aims to calculate true agreement beyond random agreement between raters (21). The kappa statistic shows how rater-rated agreement compares to random agreement (20,22). A positive kappa value may indicate that raters were better at agreeing, while a negative kappa value may indicate that agreement was worse than random. The kappa statistic is calculated by dividing the observed agreement rate by the expected random agreement rate. This statistic is considered corrected for random agreement between raters. Kappa value takes a value between 0 and 1, 0 indicating no agreement and 1 indicating complete agreement. In Kappa statistics, values less than 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability (20,22).

The kappa statistic is an important tool for measuring consistency between raters and the reliability of the evaluation process (22,23). However, it is important to implement and interpret it correctly, depending on the characteristics of the evaluation process, the type of data and its purpose (22,23).

Fleiss' Kappa: Fleiss' Kappa is a statistical method derived from Cohen's Kappa and used in the analysis of categorical data (24). Used when multiple raters evaluate nominal or ordinal data (24–26).

Light's Kappa is a variation of the Kappa statistic that calculates the average of all combinations of one-to-one Kappa values between raters (27). It is a measure of scoring agreement that takes into account multiple raters and their pairwise agreements. When calculating Light's Kappa, you first calculate the Kappa statistic between each pair of raters (28). Then, by averaging all these pairwise Kappa values, you get Light's Kappa (23). This method provides an overall measure of agreement that takes into

account each rater's agreement with all other raters. Light's Kappa is particularly useful when multiple raters are involved in the evaluation process (21,28). It provides an overall rating by taking into account the agreement of multiple raters. By considering all combinations of one-to-one Kappa values, Light's Kappa provides a more robust estimate of agreement between raters and captures an overall consensus or disagreement between raters. It helps evaluate the consistency and reliability of evaluation results in a study or evaluation process involving more than one rater.

Aim: This study aims to compare the scores of different raters within the scope of "inter-rater reliability" in simulated patient training program that conducting in Suleyman Demirel University, Faculty of Medicine in the third year of undergraduate medical program.

2. Material and Method

The study was designed comparatively in a quantitative research design. Approval was received for the study from the Süleyman Demirel University Clinical Research Ethics Committee (Date: 16.08.2023 No:168). Süleyman Demirel University Simulated Patient Laboratory data were used for the study. In the study, scenarios prepared for "improving patient- physician interviewing skills" were used for 34 students. Within the scope of these scenarios, students interviewed with simulated patients. The interviews were scored by 4 different groups of raters (Self-assessment scoring, Peer assessment scoring, Simulated patient assessment scoring, Trainers' assessment scoring). Means and standard deviations of the ratings were calculated. Scorings were shown graphically. Statistical difference between scores was analyzed with ANOVA test. ANOVA (Analysis of Variance) is used to examine differences between groups. ANOVA analyzes how a dependent variable varies depending on one or more independent variables. The p value obtained as a result of the ANOVA test helps determine whether the difference between groups is statistically significant. If the p value is less than a certain alpha level (usually 0.05), the difference between groups is considered significant. In our study, if the p value was less than a certain alpha level (usually 0.05), the difference between the groups was considered significant.

The common method used in inter-rater analyzes is interclass correlation coefficient (ICC) analysis. ICC is a widely used statistical method to evaluate the degree of consistency between different raters on a measure. With this analysis, the level of reliability is measured by calculating the degree of correlation between the scores given by the raters. A two-factor model ICC is recommended to evaluate the reliability of measurements made by different raters for the same students. In our study, Cohen's Kappa coefficient was calculated for correlation analysis. In addition to Cohen's Kappa coefficient, Fleiss' Kappa and Lights Kappa were also calculated.

3. Results

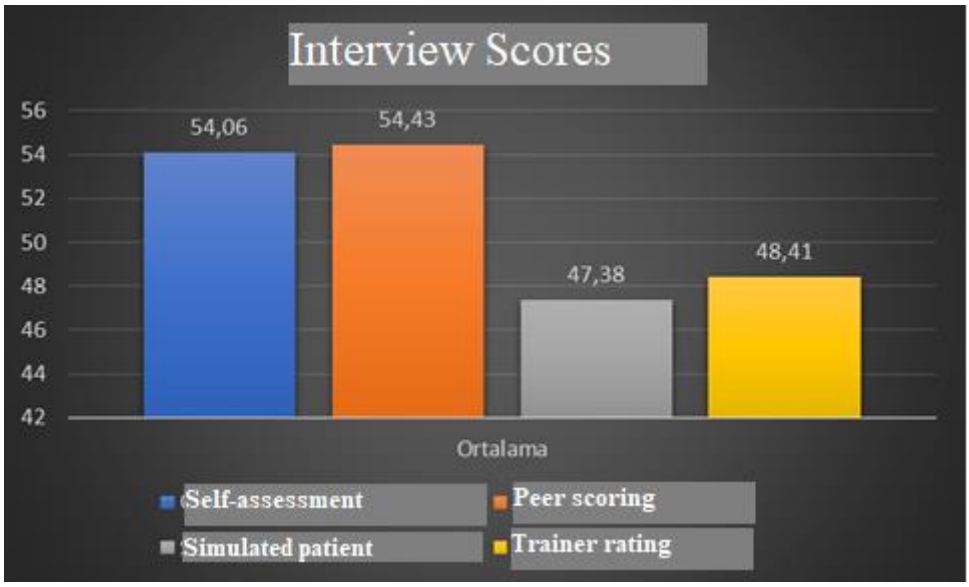
34 third-year medical school students participated in the study (n:34). 20 (59%) of the students were women and 14 (41%) were men. In this study, data of 34 students were evaluated. According to the self-assessment results, the average score of the participants was 54.06±5.10 (mean ± standard deviation), with the minimum score of 43, and the maximum score of 60. According to the peer scoring results, the average score of the participants was 54.43±5.69, with the minimum score of 37 and the maximum score of 60. According to the simulated patient scoring results, the average score of the participants was 47.38±7.83, with the minimum score of 34 and the maximum score of 60. According to the trainer scoring results, the average score of the participants was 48.41±6.70, with the minimum score of 36 and the maximum score of 60 (Table 1).

Table 1. Scoring of Students

	n	Mean	±	Standard Deviation	Min	Max
Self-assessment	34	54,06	±	5,10	43	60
Peer scoring	34	54,43	±	5,69	37	60
Simulated patient	34	47,38	±	7,83	34	60
Trainer rating	34	48,41	±	6,70	36	60

The graph of the interview scores shows the distribution of the Self-assessment scoring/Peer scoring and Simulated patient scoring/Trainer scoring groups (Graph 1).

Graph 1. Interview Scores



ANOVA analysis among the raters in the study showed that there was a statistically significant difference between the groups ($p<0.001$). Examination of the relationship between self-assessment and peer assessment showed that there is no statistically significant difference ($p = 0.684$). And also comparing the simulated patient assessment scores and trainers' assessment scores again no statistically significant difference was found ($p = 0.149$). There is a statistically significant difference between self-assessment scores comparing with simulated patient and trainers' assessment scores ($p<0.001$). Similarly, there is also a statistically significant difference between peer assessment scores comparing with simulated patient and trainers' assessment scores ($p<0.001$) (Table 2).

Table 2. ANOVA (Repeated Measures ANOVA (Non-parametric))

	Peer scoring	Simulated patient	Trainer rating
Self-assessment	0.684	< 0.001	< 0.001
Peer scoring		< 0.001	< 0.001
Simulated patient			0.149

Friedman		
χ^2	df	p
35.83	3	< .001

For all raters, ICC was calculated as 0.40 and Fleiss' Kappa was calculated as 0.724. Peer scoring-self-evaluation scoring ICC was calculated as 0.64 and Light's Kappa was calculated as 0.912. For simulated patient scoring-Trainer scoring, ICC was calculated as 0.96 and Light's Kappa was calculated as 0.140 (Table 3).

Table 3. Inter-rater reliability

	Raters	ICC	Fleiss' Kappa	Lights Kappa	p
All raters	4	0.40	0,724		
Scoring your peer - scoring with self-evaluation	2	0,64		0,912	0,912
Simulated patient scoring – Trainer scoring	2	0,96		0,140	<0,001

4. Discussion and Conclusion

In medical education curriculums, for patient safety and students development training the students with simulated patients before practicing with or on real patients, have taken their place as an important element of modern medical education (1,8,29).

In this study, we aimed to evaluate the inter-rater reliability and compare the assessment scores of different raters enrolled in assessment process of Suleyman Demirel University, Faculty of Medicine's training program with simulated patients. Although the low number of participants of this pilot study is considered a limitation, it provides very valuable information in terms of the feedback it provides.

In literature in lots of study the researchers explored the effectiveness of simulated patient scoring and trainer scoring in medical education (29,30). Several studies have explored the impact of self-reflection, self-assessment, and peer-assessment on simulated patient counseling sessions (31,32). Although there is no consensus on this issue yet, it is common to think that it would be useful to evaluate students from different aspects (30–33). The study aimed to determine which method of scoring was more reliable and valid in assessing medical students' clinical skills.

The researchers hypothesized that simulated patient scoring would provide similar and objective evaluations compared to trainer scoring.In the study, it was shown that there were two main groups in scoring: peer scoring/self scoring and simulated patient scoring/trainer scoring. Although there was no statistically significant difference between self-scoring and peer scoring in the study, it was seen that these groups acted more "generously". Peer scoring and self-ratings are very valuable in that they provide evaluation and feedback from different perspectives. However, it does not provide sufficient evidence for conversion into notes in terms of objectivity.

This area can be developed with more comprehensive studies planned in the future. Although there was no statistically significant difference between the scores of the simulated patients and the scores of the medical educators, it was observed that they acted more "frugally". It is important that training, instructions, and evaluation criteria are clear and unambiguous to ensure high reliability between raters during the evaluation process. In the study, it was observed that the scores of simulated patients who enrolled a certifigated education program were compatible with the trainers. These ratings provide stronger evidence in terms of objectivity.

Evaluation from different aspects is very valuable in evaluating simulated patient applications. We believe that the scoring process can be improved with more comprehensive studies designed to score simulated patient applications.

Declaration of Ethical Code

In this study, we undertake that all the rules required to be followed within the scope of the "Higher Education Institutions Scientific Research and Publication Ethics Directive" are complied with and that none of the actions stated under the heading "Actions Against Scientific Research and Publication Ethics" is not carried out.

References

1. Bearman M, Nestel D. Simulated Patient Methodology: Theory, Evidence and Practice. In 2014.

2. Sayek İ. Hacettepe Üniversitesi Tıp Fakültesi eğitiminde son uygulamalar. Acta Medica Cordoba. 2004;35(2):63–4.

3. Sezgin A, Kınıklı Gİ, Kaşıkçı M. Hastane Öncesi Sağlık Personelinin Hizmet İçi Eğitiminde Simüle Hasta Uygulamasının Acil Olgu Yönetimindeki Etkinliğinin Değerlendirilmesi. Hastan Öncesi Derg. 2023;7(3):317–30.

4. Mercan N, Özcan CT, Aydın MS. Psikiyatride ve iletişim eğitiminde simüle hasta uygulamaları. Psikiyatr Güncel Yaklaşımlar. 2018;10(3):302–11.

5. GamzeSarıkoç MelihElçin CT. An Innovative Practice in Psychiatric Nursing Education: Standardized Patients. Journal. 2016;9(2):61–6.

6. Ağadayı E, Çetinkaya S, Karagöz N, Nemmezi Karaca S, Bozdoğan N. Simüle hasta uygulamasında öğrencilerin anamnez alma becerilerinin öğrenci, hasta ve öğretim üyesi gözünden değerlendirilmesi.

7. Yıldırım Sarı H, Doğan P. Öğrencilerin Görme Engelli Simüle Hasta ile İletişim Becerilerinin Değerlendirilmesi: Pilot Çalışma. J Infant, Child Adolesc Heal. 2022;2(1):1–10.

8. YeşimŞenol İbrahimBaşarıcı S. Student Opinions About Standardized Patient Practices: First-Year Results. Journal. 2014;13(41):19–26.

9. Şendir M. Kadın sağlığı hemşireliği eğitiminde simulasyon kullanımı. Florence Nightingale J Nurs. 2013;21(3):205–12.

10. Bulut İ, Sapancı A, Kara İH. Tıp eğitiminde kullanılan ölçme ve değerlendirme araçlarının geleneksel ve alternatif ölçme ve değerlendirme araçları olarak sınıflandırılması. J Med Educ Informatics. 2015;1(1):2–11.

11. Mıdık Ö. Tutumun ölçme ve değerlendirmesi. Tıp eğitiminde ölçme ve değerlendirme Ankara Türkiye Klin. 2018;

12. Özan S, Yurdabakan İ. Öz ve Akran Değerlendirmenin Temel İletişim Becerileri Başarısı Üzerindeki Etkileri. Tıp Eğitimi Dünyası. 2008;27(27):27–39.

13. Güllüdere HH, Yardım S, Sezik M, Şenol Y. Akran yardımı ile eğitimin tıp eğitiminde kullanımı. Tıp Eğitimi Dünyası. 2014;13(39):19–25.

14. Sarıkaya Ö, Uzuner A, Gülpınar MA, Keklik D, Kalaça S. İletişim becerileri eğitimi: İçerik ve değerlendirme. Tıp Eğitimi Dünyası. 2004;14(14).

15. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979 Mar;86(2):420–8.

16. Mcgraw K, Wong SP. Forming Inferences About Some Intraclass Correlation Coefficients. Psychol Methods. 1996 Mar 1;1:30–46.

17. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016 Jun;15(2):155–63.

18. Cicchetti D. Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instrument in Psychology. Psychol Assess. 1994 Dec 1;6:284–90.

19. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.

20. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam med. 2005;37(5):360–3.

21. Warrens MJ. Five ways to look at Cohen's kappa. J Psychol Psychother. 2015;5.

22. McHugh ML. Interrater reliability: the kappa statistic. Biochem medica. 2012;22(3):276–82.

23. Rae G. The Equivalence of Multiple Rater Kappa Statistics and Intraclass Correlation Coefficients. Educ Psychol Meas [Internet]. 1988 Apr 1;48(2):367–74. Available from: https://doi.org/10.1177/0013164488482009

24. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas. 1973;33(3):613–9.

25. Fleiss JL, Nee JC, Landis JR. Large sample variance of kappa in the case of different sets of raters. Psychol Bull. 1979;86(5):974.

26. Cao H, Sen PK, Peery AF, Dellon ES. Assessing agreement with multiple raters on correlated kappa statistics. Biometrical J. 2016;58(4):935–43.

27. Hubert L. Kappa revisited. Psychol Bull. 1977;84(2):289.

28. Warrens MJ. Inequalities between multi-rater kappas. Adv Data Anal Classif. 2010;4:271–86.

29. Taylor S, Haywood M, Shulruf B. Comparison of effect between simulated patient clinical skill training and student role play on objective structured clinical examination performance outcomes for medical students in Australia. J Educ Eval Health Prof. 2019;16:3.

30. 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 [Internet]. 2009;9(1):37. Available from: https://doi.org/10.1186/1472-6920-9-37

31. Ljungman AG, Silén C. Examination involving students as peer examiners. Assess Eval High Educ [Internet]. 2008 Jun 1;33(3):289–300. Available from: https://doi.org/10.1080/02602930701293306

32. Bartlett A, Pace J, Arora A, Penm J. Self-Reflection and Peer-Assessments Effect on Pharmacy Students’ Performance at Simulated Counselling Sessions. Vol. 11, Pharmacy. 2023.

33. Nakayama N, Ejiri H, Arakawa N, Makino T. Stress and Anxiety in Nursing Students Between Individual and Peer Simulations. Nurs Open. 2020;