Simulation-based Assessment of Paramedics and Performance in Real Clinical Contexts

Walter Tavares, ACP, BSc, Vicki R. LeBlanc, PhD, Justin Mausz, ACP, Victor Sun, ACP, BSc, Kevin W Eva, PhD

Abstract

Objective. The objective of this study was to seek validity evidence for simulation-based assessments (SBA) of paramedics by asking to what extent the measurements obtained in SBA of clinical competence are associated with measurements obtained in actual paramedic contexts, with real patients. **Methods.** This prospective observational study involved analyzing the assessment of paramedic trainees at the entry-to-practice level in both simulation- and workplace-based settings. The SBA followed an OSCE structure involving full clinical cases from initial patient contact to transport or transfer of care. The workplace-based assessment (WBA) involved rating samples of clinical performance during real clinical encounters while assigned to an emergency medical service. For each candidate, both assessments were completed during a 3-week period at the end of their training. Raters in the SBA and WBA settings used the same paramedic-specific seven-dimension global rating scale. Reliability was calculated and decision studies were completed using generalizability theory. Associations between settings (overall and by dimension) were calculated using Pearson's correlation. **Results.** A total of 49 paramedic trainees were assessed using both a SBA and WBA. The mean score in the SBA and WBA settings were 4.88 (SD = 0.68) and 5.39 (SD = 0.48), respectively, out of a possible 7. Reliability for the SBA and WBA settings reached 0.55 and 0.49, respectively. A decision study revealed 10 and 13 cases would be needed to reach a reliability of 0.7 for the SBA and WBA settings. Pearson correlation reached 0.37 ($p = 0.01$) between settings, which rose to 0.73 when controlling for imperfect reliability; five of seven dimensions (situation awareness, history gathering, patient assessment, decision making, and communication) reaching significance. Two dimensions (resource utilization and procedural skills) did not reach significance. **Conclusion.** For five of the seven dimensions believed to represent the construct of paramedic clinical performance, scores obtained in the SBA were associated with scores obtained in real clinical contexts with real patients. As SBAs are often used to infer clinical competence and predict future clinical performance, this study contributes validity evidence to support these claims as long as the importance of sampling performance broadly and extensively is appreciated and implemented. **Key words:** assessment; clinical competence; paramedic; simulation

PREHOSPITAL EMERGENCY CARE 2013;Early Online:1–7

Introduction

Assessments designed to make decisions regarding a candidate's clinical ability and readiness for unsupervised clinical practice or licensure are some of the most important and complex decisions educators, raters and, more broadly, licensing bodies have to make. Not only is the candidate's academic success or professional advancement contingent upon this complex task, but the downstream effects of these decisions could be serious as well, particularly with respect to patient safety. Competence is defined as "the degree to which an individual can use the knowledge skills and judgment associated with the profession to perform effectively in a domain of possible encounters defining the scope of professional practice."[1] Like other constructs, competence cannot be measured directly or comprehensively and, therefore, must be inferred through the observation of behaviors in response to clinical challenges.

One such form of observation occurs through the use of simulation-based assessments (SBA). One of the goals of SBA is to extrapolate the observations collected in a simulated environment (i.e., assessment context) to enable inferences to be drawn about future performance in real clinical contexts with real patients. In SBA, assessors attempt to optimize the balance between ecological validity, standardization, and control over content.[2–4] Cases can be carefully designed to challenge and hopefully reveal a candidate's ability to integrate multiple competencies, use judgment, and integrate problem solving and other clinically relevant features in realistic contexts. However, while

Received March 17, 2013, from Centennial College Paramedic Program (WT, JM, VS), Toronto, Ontario, Canada; University of Toronto, Wilson Centre for Medical Education (WT), Toronto, Canada; and Centre for Health Education Scholarship, Department of Medicine (KWE), University of British Columbia. Revision received May 21, 2013; accepted for publication May 24, 2013.

The authors would like to thank York Region Emergency Medical Services and Peel Regional Paramedic Service for their support. The authors would also like to thank all the paramedic trainees and paramedics who agreed to participate in this study, as well as Centennial College (Toronto, Ontario, Canada) Health Studies Department for their generous support.

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Address correspondence to Walter Tavares, School of Community and Health Studies, Centennial College, P.O. Box 631 Station A, Toronto, ON, Canada, M1K 5E9. e-mail: wtavares@centennial college.ca

doi: 10.3109/10903127.2013.818178

SBAs are associated with numerous advantages (e.g., eliminating patient safety concerns, ensuring representativeness), there are limitations to consider. For instance, some medical conditions (i.e., certain physical exam findings) may not be adequately simulated and some meaningful contextual forces (e.g., perceived consequences) may be difficult to replicate. Simulations are argued to only ever be surrogates of reality and therefore limited when used to understand or predict how individuals will function in real clinical contexts. Finally, validity arguments are often weak, especially in paramedic settings where a paucity of research exists. Still, SBA continues to be widely used and relied on for many high stakes decisions in paramedic training and practice. As the role of paramedicine in the health-care system continues to grow and the scope of practice (in depth and breadth) grows along with it, the utility of high stakes SBA depends crucially on its psychometric characteristics.

According to the Standards for Educational and Psychological Testing, validity refers to the appropriateness, meaningfulness, and usefulness of specific inferences made from test scores.[5] To support the interpretation of tests scores, or put differently, to justify drawing of generalizations about an individual's competence based on scores gathered in a simulation context, supporting evidence must be collected.[5] One such form of evidence is to determine whether performance in a simulation environment is associated with performance in actual clinical environments. Attempts to establish evidence of this kind are often challenging, however, due to challenges inherent in professional activities (e.g., the nonrepresentativeness that can occur when observing performance in real world settings). Measures of performance often differ between settings and/or across numerous confounding variables, such as time pressures, maturation, and number of clinical exposures, and thereby threaten the interpretation of results.

The objective of this study was to seek validity evidence for SBA by asking to what extent the measurements obtained in paramedic SBA of clinical competence are associated with measurements obtained in actual paramedic contexts, with real patients. To do so, this study takes advantage of an innovative program of assessment included in the Paramedic Program offered jointly through Centennial College and the University of Toronto. Candidates in this program are assessed using both an OSCE (SBA) and a workplace-based assessment (WBA) referred to as paramedic clinical sampling (PCS) during the same time period using the same rating tool.

Methods

Overview

This prospective observational study involved analyzing work-based and simulation-based assessment of paramedic trainees at the entry to independent practice level (i.e., candidates who had completed all training requirements and were eligible for graduation from a paramedic program, but who had not practiced as independent clinicians). The SBA followed an OSCE structure in which paramedic candidates rotated through a series of standardized stations, each involving full clinical cases consistent with paramedic practice (i.e., from initial patient contact to point of transport or transfer of care). The WBA followed a paramedic clinical sampling (PCS) approach, which involves sampling clinical performance across a number of clinical encounters, each assessed by a different rater on a different day in a different context. Unlike the SBA, PCS does not include any control over content and is not standardized, but the encounters involve real and unpredictable clinical cases while responding to 9-1-1 calls in an emergency medical service. Ethical approval for this study was provided by Centennial College in Toronto, Ontario, Canada (REB #086).

Participants

Participants were paramedic trainees in their final year of a two-year paramedic program offered jointly through Centennial College and the University of Toronto (Ontario, Canada) in 2011 and 2012 who participated in both the simulation- and workplace-based assessments as part of their program's summative evaluation process. The scope of practice for this group of trainees involves both basic life support (BLS) and advanced life support (ALS).[6,7] Patient care standards as defined by the Ontario Ministry of Health Emergency Health Services Branch (the provincial regulatory body).

Raters/Observers

Raters for both the SBA and WBA were active paramedics who had experience rating clinical performance in their respective settings but who varied in the number of times they had used the assigned rating scale. All raters in the SBA had received rater training (i.e., frame of reference and performance dimension training)[8] and were provided with details about the case they would assess along with explicit performance expectations in advance of the SBA. In contrast, raters in the WBA setting were not provided with rater training, nor could case details or performance expectations be provided given the unpredictability of the environment. Instead, raters used profession-specific clinical standards (i.e., existing medical directives for the service area and provincial patient care standards),[6,7] as well as their own clinical experience to generate their ratings.

Measure

For both the SBA and WBA, each paramedic trainee was assessed using a seven-dimension (situation awareness, history gathering, patient assessment, decision making, resource utilization, communication, and procedural skills) global rating scale (GRS).[9] Each dimension was scored using a seven-point adjectival scale from 1 = unsafe: "unsuitable for supervised practice or progression" to 7 = exceptional: "highly recommended for independent practice or progression, an example for others." This GRS has demonstrated evidence of content and discriminant validity, as well as high interrater reliability (0.75–0.94) and intrarater reliability (0.94) when used to assess paramedic trainees.[9]

Procedures

The SBA followed an OSCE structure[10] in which trainees rotated through five independent stations that required them to interact with full clinical cases relevant to paramedic practice (i.e., from initial patient contact to point of transport or transfer of care) rather than performing isolated tasks. Cases were drawn from a curriculum blueprint and developed based on actual clinical cases. Broadly, the cases involved (1) an adult multisystem trauma victim, (2) an adult medical patient (e.g., cardiac patient), (3) a medical patient with distracting injuries, (4) an obstetrical emergency involving neonatal resuscitation, and (5) a pediatric medical or trauma patient. Trainees were evaluated by one rater in each station, using the seven-dimension GRS described above.

The WBA involved rating samples of clinical performance in an emergency medical service with real contexts and real patients. The clinical setting involved two independent but similar emergency medical services (EMS) in southern Ontario, Canada. The sampling involved having trainees scheduled with a different rater on a different ambulance/paramedic response unit, in a different response station/geographic area (one per day), until five assessments were completed. In OSCE parlance, each day with a new rater could be considered equivalent to a "station." Due to logistical constraints and the real-world nature of the assessment, raters and cases varied for each trainee (i.e., trainee A and trainee B may have been evaluated by a completely different set of raters on a completely different set of cases).

While assigned to each rater in the workplace setting, trainees responded to emergency calls and were required to demonstrate the technical and nontechnical skills expected of an entry-level paramedic, from point of contact to transfer of care, for any patient interaction that they happened to be presented with. The assigned rater (a paramedic) supervised the interaction to ensure patient safety, assessed the candidate's performance, then completed the seven-dimension GRS described above. In circumstances where the assigned rater was required to intervene due to patient safety concerns, raters were instructed to score the candidates' performance to the point where responsibility for patient care was transferred and to consider the need to intervene in their assessment. As this was intended to be summative, raters were instructed to avoid providing formative feedback (similar to the OSCE) and to not share the GRS scores they assigned. For the purposes of this study, the first call/assessment of the day was used if a candidate was assessed on more than one patient interaction in a given day. This allowed us to evaluate reliability and validity based on a limited number of interactions (promoting external validity), to ensure a random selection of cases, to limit bias, and to ensure independence between observations.

In both settings, raters were instructed to evaluate the individual, functioning as a leader within the context of a team (which may have varied from case to case within and between settings). Finally, to promote consistency in application of the rating tool, raters were instructed to review the definition for each of the seven dimensions and the criteria for selecting a score prior to beginning the rating process (see rating definitions table on GRS).

Data Collection

Data were collected using a paper-based GRS in both the SBA and WBA settings. For the SBA setting, completed GRS forms were collected immediately following the assessment, transferred to an excel database for use by the educational program, and then archived. For the WBA setting, raters were asked to complete the GRS and return it to the paramedic program in a sealed envelope that was signed across the seal. Returned GRS forms were transferred into an Excel database for use by the educational program, and then archived. Once trainees had left the paramedic program (i.e., were no longer affiliated with the educational institution in any way) the data were retrieved for analysis.

Sample Size

For this prospective observational study we collected data on all available trainees who had completed both the SBA and WBA in the same time period ($n = 57$). We estimated a correlation between SBA and WBA of $r = 0.3$ and calculated that a total sample size of 30 would be sufficient to achieve statistical significance.[11]

Data Analysis

For both settings, descriptive statistics were calculated and repeated measures ANOVA was used to compare means. To address our primary objective of seeking validity evidence for the SBA, we calculated the Pearson's correlation between scores assigned on the GRS in both settings (overall and by dimension). With any measurement there is almost always a degree of error (indicated by the reliability of that measurement) that may attenuate results and thereby yield an inaccurate impression of the correlation between variables.[12] Therefore, we estimated what the correlation between settings would be under improved reliability conditions (referred to as disattenuated correlations). This may provide a more accurate impression of the true association that exists between settings. All correlation analyses were performed using SPSS Ver. 19.

A fundamental aspect of an assessment's validity is its reliability (i.e., the extent to which the scores consistently differentiate between individual candidates), which can be calculated using generalizability theory.[13–15] This statistical approach uses ANOVA to separate the variance observed in the scores to determine what portion of the total variance can be attributed to the variable of interest (which in this case are the participants) and what portion can be attributed to various sources of measurement error. These proportions are then used to calculate reliability (or generalizability) coefficients and values range from 0 to 1 with higher values representing higher reliability. With sources of variance identified, the information can then be used to determine what strategies might best contribute to improved reliability (a process referred to as "decision studies" (or D-studies). D-studies provide researchers with information that might be helpful in informing future assessment practices (e.g., estimating reliability under different conditions, such as increased number of cases, number of raters, or number of items).

Based on the procedures described above, we used identical generalizability theory analyses/designs (i.e., case nested in trainee crossed with item) for both the SBA and WBA to make direct comparisons regarding the magnitude of measurement error/reliability. A D-study was then conducted to determine how gains in reliability might be best achieved. All generalizability analyses were conducted using G-String IV.

Results

Out of a possible 57 trainees, 49 were assessed over three weeks using both SBA (i.e., OSCE) and WBA (i.e., PCS) in April 2011 and 2012. The remaining 8 trainees completed the SBA, but did not complete the WBA as described above (i.e, fewer than 5 observations were collected). Raters involved in the SBA did not evaluate any of the trainees in the work-based setting. For the SBA, each candidate completed five stations, which varied in content depending on year and track. The overall mean score for trainee in the OSCE was 4.88 (SD = 0.68) out of a possible 7. The mean score by dimension ranged from 4.51 (SD = 0.83) (decision making) to 5.14 (SD = 0.62) (resource utilization). See Table 1 for details regarding dimension means (including standard deviations).

For the WBA, a total of 231 patient encounters (an average of 4.7 cases per candidate) were included in the WBA data set. The overall mean score for trainee was 5.39 (SD = 0.48) out of a possible 7. The mean score by dimension ranged from 5.31 (SD = 0.43) (resource utilization) to 5.51 (SD = 0.43) (procedural skill). See Table 1 for details regarding dimension means (including standard deviations). ANOVA comparing scores on each dimension revealed that trainees generally received higher scores from raters conducting their assessments within the context of WBA relative to those offering judgment in the context of SBA (see Table 1).

Table 2 illustrates the proportion of variance attributable to each variable included in the study for each of SBA and WBA and the combination of both

Table 1. Means and standard deviations for SBA and WBA overall and within dimension

	SBA		WBA		
Dimension	Mean	STD	Mean	STD	*p* value
Situation awareness	4.97	0.76	5.37	0.50	<0.001
History gathering	4.90	0.57	5.41	0.48	<0.001
Patient assessment	4.84	0.67	5.43	0.51	<0.001
Decision making	4.51	0.83	5.33	0.49	<0.001
Resource utilization	5.14	0.62	5.31	0.43	0.08
Communication	5.01	0.56	5.40	0.49	<0.001
Procedural skill	4.85	0.74	5.51	0.43	<0.001
Mean scores	4.88	0.68	5.39	0.48	<0.001

SBA, simulation-based assessment; WBA, workplace-based assessment; STD, standard deviation.

Table 2. Summary of facets, their estimated variance components with percentage of total variance, and *G*-coefficients for the SBA (i.e., OSCE), WBA, and when combining SBA with WBA

	Simulation-based assessment (OSCE)		Workplace-based assessment (PCS)		Combined SBA with WBA	
Effect	VC	% of total variance	VC	% of total variance	VC	% of total variance
t	0.19	11.8	0.09	12.2	0.11	9.0
c:t	0.68	42.0	0.46	62.2	0.66	54.0
i	0.04	2.5	0.00	0.0	0.01	1.0
ti	0.00	0.0	0.01	1.4	0.00	0.00
ci:t	0.67	41.4	0.18	24.3	0.44	36.0
Total var.	1.62		0.74		1.23	
Generalizability	*G* = 0.55		*G* = 0.49		*G* = 0.61	

t, trainee; c:t, case nested in trainee; i, items; ti, trainee and item interaction; ci:s, case and item interaction nested in trainee confounded with random error; VC, variance components; SBA, simulation-based assessment; WBA, workplace-based assessment; OSCE, objective structure clinical evaluation; PCS, paramedic clinical sampling; Var., variance.

Table 3. Pearson correlations (and dissattenated correlations) between SBA and WBA by dimension

Dimension	Pearson's correlation (disattenuated correlation)	p Value
Situation awareness	0.34 (0.68)	0.02
History gathering	0.35 (0.69)	0.01
Patient assessment	0.35 (0.69)	0.01
Decision making	0.29 (0.57)	0.04
Resource utilization	0.22 (0.43)	0.14
Communication	0.29 (0.57)	0.04
Procedural skill	0.23 (0.45)	0.11
Overall	0.37 (0.73)	0.01

assessments. Both settings revealed similar patterns, with the largest sources of error being case differences (WBA = 62% and SBA = 42%) and the residual error term of item × case (nested in trainee) (WBA = 24.3% and SBA = 41.4%), and revealed relatively little variance attributable to the students (12% in both WBA and SBA). This suggests a strong context specificity effect (i.e., that performance in one case is not a strong predictor of performance in another case) in both instances and speaks to the importance of collecting multiple samples of performance and the difficultly of meaningfully differentiating between trainees with few observations. Combining these variance components into generalizability (i.e., reliability) coefficients reveals that if one averages across 5 cases, the reliability of the resulting scores was 0.55 and 0.49 for SBA and WBA, respectively. Decision studies performed on these coefficients revealed that achieving a reliability of 0.7 would require 10 cases to be observed and rated in SBA and 13 cases to be observed and rated in WBA. When both sets of data were aggregated together, such that 10 cases were considered, the *G*-coefficient was found to be 0.61.

Using the scores assigned on the GRS in both settings, the Pearson's correlation between SBA and WBA was $r = 0.37$ ($p < 0.01$). These correlations ranged from 0.22 ($p = 0.14$) to 0.35 ($p < 0.01$). Five of the seven dimension-specific correlations (situation awareness, history gathering, patient assessment, decision making, and communication) were statistically significant. Disattenuating these correlations to account for the imperfect reliabilities of the tests given the use of only 5 cases revealed an association between SBA and WBA of $r = 0.73$ overall, which ranged in dimension from 0.43 (resource utilization) to 0.69 (history gathering and patient assessment) (see Table 3).

Discussion

The need for scientific validation of high-stakes performance exams derives from the social requirement that assessment decisions be defensible.[16] The objective of this study was to test the validity of SBA by asking to what extent the measurements obtained in SBA are associated with measurements obtained by WBA involving actual paramedic contexts and real patients. The results of this study are the first to suggest that assessment of paramedic trainees completed in a simulation-based environment using a global rating scale and a multiple sampling strategy is significantly associated with performance in actual clinical contexts when using the same measure and a similar sampling strategy.

The construct of (paramedic) clinical competence is complex, abstract, and only adequately assessed when direct observations of behaviors are made in response to clinical challenges. Swanson et al. argue that assessing the clinical performance of health professionals requires testing complex knowledge and skills in real-world contexts, where they are actually used.[17] However, like many other health professions, the difficulties associated with WBA (e.g., patient safety, unpredictability, rater biases) have led the paramedic community to rely on SBA for high-stakes entry to practice performance examinations. The advantages associated with simulation are well documented.[18–20] However, validity evidence supporting the inferences drawn when used as an assessment modality have been less certain. This study supports the growing body of research that suggests simulation-based strategies translate to real clinical contexts[21] and provides evidence for and supports the continued use of SBA in this context.

It must be noted, of course, that validity is dependent on implementation. Consistent with this general rule, we found that the correlation between WBA and SBA increased with increased reliability (as illustrated through the disattenuated correlation coefficients). D-studies revealed that both assessment methods require performance scores to be collected on about 10 (in SBA) to 13 (in WBA) cases to achieve a reliability coefficient of at least 0.7. However, gains in reliability may also be achieved by improving the degree to which raters are able to differentiate between candidates. Whether that might occur through the use of rater training, different measurement tools, more deliberate case selection, or reducing the demands placed on raters during the rating task so that they can direct more attention to the candidate performance[22] continues to be an area of research. As raters may use different criteria and standards to judge clinical performances and may idiosyncratically vary in what they observe,[23,24] this study supports the assertion that multiple ratings using multiple contexts should be collected and combined for optimal appraisals of clinical performance.[25]

One such strategy toward optimizing appraisals of clinical performance that supports van der Vleuten's concept of a program of assessment[26] would be to consider combining the SBA with WBA. The reliability appraisals among the SBA and WBA were similar despite the lack of control over cases and contexts, lack of standardization, and lack of rater training (e.g., performance dimension, frame of reference, and GRS training) in the work-based setting. This suggests that, despite their theoretical differences in terms of strengths and weaknesses, WBAs may provide similarly trustworthy information relative to the more standardized and controlled SBA when sufficient observations are collected. At the same time, the imperfect correlation suggests it is possible that each type of assessment has a distinct role in informing the underlying construct and, therefore, should be considered together when making decisions regarding a trainee's clinical competence. Incorporating several competency elements and multiple sources of information on multiple occasions can be expected to strengthen the validity argument.[26] In this study, when scores from the WBA and SBA were combined reliability increased to 0.61 for the sum total of ten performance scores. Future research will need to determine how best to integrate SBA and WBA into a final comprehensive evaluation process.

Not all dimensions were significantly correlated between SBA and WBA and the scores were higher overall in the WBA when compared to the SBA. The two dimensions (resource utilization and procedural skill) that failed to reach significance may be particularly difficult to assess in one or both contexts, either due to limited opportunity or rater difficulties. Alternatively, for these two dimensions, performance in SBA may simply not be predictive of performance in real or different clinical contexts. We did not attempt to identify causes related to the difference in mean scores between SBA and WBA. It is possible that raters are more lenient in real-world environments or that the cases chosen for SBA were more difficult due to their being high-acuity and low-frequency conditions (e.g., obstetrical emergency requiring neonatal resuscitation).

Limitations

Some limitations associated with this study need to be acknowledged. First, only two emergency medical services participated. While they contained a mix of urban, suburban, and rural/remote response stations, the findings reported here may not generalize to all emergency medical services. Future studies in different emergency medical services will need to be conducted. Second, we used the first call of each day rather than all calls on a given day. Including all calls on a given day would have the advantage of efficiently increasing the sample of observations within each condition, but would also eliminate the independence of the observations. We did not compare ratings assigned by one rater with extensive observations (e.g., preceptors assigned to the candidate for a prolonged period of time) with multiple raters each basing their rating on a small sample of performance. Future research should explore the effects of one strategy over the other as researchers in other fields conducting studies addressing this question have concluded that the latter is more psychometrically sound.[12] That is, aggregating over multiple raters is a powerful means of minimizing the contribution of measurement error.[12] Finally, the SBA included rater training, while the WBA did not. This may have contributed to the slight differences we observed in reliability. However, the results of this study (i.e., similar reliabilities despite differences in the availability of rater training) suggest that applying our process to other settings where access to raters for comprehensive rater training is not possible may not be a prohibitive factor.

Conclusion

For five of the seven dimensions believed to represent the construct of paramedic clinical performance, scores obtained in the SBA were associated with scores obtained in real clinical contexts with real patients. As SBAs are often used to infer clinical competence and predict future clinical performance, this study contributes validity evidence to support these claims as long as the importance of sampling performance broadly and extensively within each assessment strategy is appreciated and implemented.

References

1. Kane MT. An argument-based approach to validity. Psychol Bull. 1992;112(3):527–35.
2. Boulet J, Swanson D. Psychometric challenges of using simulations for high-stakes assessment. In: Simulations in Critical Care Education and Beyond. Des Plains, IL: Society of Critical Care Medicine; 2004: 119–30.
3. Swanwick T. Understanding medical education. 2010: Wiley Online Library.
4. van der Vleuten C, Swanson D. Assessment of clinical skills with standardized patients: state of the art. Teaching Learning Med. 1990;2(2):58–76.
5. Brennan RL. Educational measurement. 2006: Praeger Pub Text.
6. Ministry of Health Emergency Health Services Branch, ALS – Advanced Life Support Patient Care Standards. Queen's Printer for Ontario, 2011. Vol. 3.0.
7. Ministry of Health Emergency Health Services, BLS Standards of Care. Publications Ontario, 2007. Vol. 2.0.
8. Roch SG, Woehr DJ, Mishra V, Kieszcynska U. Rater training revisited: an updated meta-analytic review of frame-of-reference training. J Occup Organiz Psychol. 2011;85(2):370–95.
9. Tavares W, Boet S, Theriault R, Mallette T, Eva WW. Global rating scale for the assessment of paramedic clinical competence. Prehosp Emerg Care. 2012;17 (1):57–67.
10. Harden R, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. BMJ. 1975;1(5955):447.
11. Norman G, Streiner D. Biostatistics: The Bare Essentials. 2008: BC Decker, Pmph USA Ltd.
12. Lakes KD, Hoyt WT. Applications of generalizability theory to clinical child and adolescent psychology research. J Clin Child Adolesc Psychol. 2009;38(1):144–65.
13. Brennan R. Generalizability theory. Educ Measurement Issues Pract. 1992;11(4):27–34.

14. Cronbach LJ. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley, 1972.
15. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teacher. 2012;34(11):960–92.
16. Kane M. Validating score interpretations and uses. Language Testing. 2012;29(1):3–17.
17. Swanson D, Norman G, Linn R. Performance-based assessment: lessons from the health professions. Educ Res. 1995;24(5):5.
18. Boulet JR. Summative assessment in medicine: the promise of simulation for high–stakes evaluation. Acad Emerg Med. 2008;15(11):1017–24.
19. Boulet JR, Murray D, Kras J, Woodhouse J, McAllister J, Ziv A. Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. Anesthesiology. 2003;99(6):1270–80.
20. Ziv A, Wolpe PR, Small SD, Glick S. Simulation-based medical education: an ethical imperative. Acad Med. 2003;78(8):783–8.
21. Andreatta P, Saxton R, Thompson M, Annich G. Simulation-based mock codes significantly correlate with improved pediatric patient cardiopulmonary arrest survival rates. Pediatr Crit Care Med. 2011;12(1):33–8.
22. Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. Adv Health Sci Educ. 2012;18(2): 291–303.
23. Martin M, Hubble MW, Hollis M, Richards ME. Interevaluator reliability of a mock paramedic practical examination. Prehosp Emerg Care 2012;16,(2):277–83.
24. Han J, Kreiter CD, Park H, Ferguson KJ. An experimental comparison of rater performance on an SP-based clinical skills exam. Teaching Learning Med. 2006;18(4): 304–9.
25. Williams R, Klamen D, McGaghie W. Cognitive, Social and environmental sources of bias in clinical performance ratings. Teaching Learning Med. 2003;15(4):270–92.
26. Van Der Vleuten C, Schuwirth L. Assessing professional competence: from methods to programmes. Med Educ. 2005;39(3):309–17.