Intern Emerg Med (2017) 12:651–656
DOI 10.1007/s11739-016-1486-7

EM - ORIGINAL

An Italian version of the Ottawa Crisis Resource Management Global Rating Scale: a reliable and valid tool for assessment of simulation performance

Jeffrey Micheal Franc[1,2] · Manuela Verde[2] · Alba Ripoll Gallardo[2] · Luca Carenzo[2] · Pier Luigi Ingrassia[2]

Received: 24 November 2015 / Accepted: 7 June 2016 / Published online: 16 June 2016

Abstract Objective measurement of simulation performance requires a validated and reliable tool. However, no published Italian language assessment tool is available. Translation of a published English language tool, the Ottawa Crisis Resource Management Global Rating Scale (GRS), may lead to a validated and reliable tool. After developing an Italian language translation of the English language tool, the study measured the reliability of the new tool by comparison with the English language tool used independently in the same simulation scenarios. In addition, the validity of the Italian language tool was measured by comparison to a skills score also applied independently. The correlation coefficient between the Italian language overall GRS and the English language overall GRS was 0.82 (adjusted 95 % confidence interval: 0.62–0.92). The correlation coefficient between the Italian language overall GRS and the skill score was 0.85 (adjusted 95 % confidence interval 0.68–0.94). This study demonstrated that the Italian language GRS has acceptable reliability when compared with the English language tool, suggesting that it can be used reliably to evaluate the performance during simulated emergencies. The study also suggests that the tool has acceptable validity for assessing the simulation performance. The study suggests that the Italian language GRS translation has reasonable reliability when compared with the English language GRS and reasonable validity when compared with the assessment of the skills scores. Data suggest that the instrument is adequately reliable for informal and formative type of examinations, but may require further confirmation before use for high-stake examinations such as licensing.

Keywords High-fidelity simulation · Reliability · Validity · Assessment tool · Emergency medicine

Introduction

With advances in medical technology, teaching in medicine is often shifting from the traditional apprentice-based model to more active engagement through proficiency-based training [1]. Often, this training includes simulated situations of short duration that allow for immediate feedback, reflection, and corrective actions. The evaluation of performance plays a key role in adult learning, facilitating reflection and refinement of behavior [2–4].

Levine et al. suggest that simulation is important for real-time assessment of physician competency in the modern age of medicine [5]. Indeed, over the past two decades, the role of simulation has grown in certification and credentialing [6]. Evidence suggests that simulation scenarios perform as well as OSCE (Objective Structured Clinical Evaluation) or observation of actual patient encounters [7, 8].

Unfortunately, it is all too common in medicine for evaluation tools to be used without first proving their reliability and validity. In "The Metric of Medical Education,", Downing states that, like all scientific data, assessment data must be reliable to be meaningfully interpreted [9]. In addition, he describes validity as the

Electronic supplementary material The online version of this article (doi:10.1007/s11739-016-1486-7) contains supplementary material, which is available to authorized users.

✉ Jeffrey Micheal Franc
jeffrey.franc@gmail.com

[1] University of Alberta, 790 University Terrace, 8303–1121 Street NW, Edmonton, AB T6G 2T4, Canada

[2] Universita' del Piemonte Orientale, Novara, Italy

"sine qua non of assessment," and cites that educational assessment tools have little intrinsic value without evidence of validity [10]. Thus, the burden of proof of reliability and validity of any evaluation instrument clearly rests on the examiner. The adoption of objective and validated assessment tools is paramount to ensure the correct measurement of achieved competencies.

The authors were unable to find any published Italian language simulation evaluation tools that have been documented to be reliable and valid. The English language Ottawa Crisis Resource Management Global Rating Scale (Ottawa GRS) has been thoroughly assessed for validity and reliability, and appears to contain many features that suggest an ideal evaluation tool. The overall performance field of the Ottawa GRS is rated on a 1–7 semi-anchored scale, where 1 is "Novice" and 7 is "Clearly Superior.". The scale includes anchored text at the 1, 3, 5, and 7 values that describe in more detail the requirements needed to reach each score. In addition, the Ottawa GRS includes five specific domains: leadership skills, problem-solving skills, situational awareness skills, resource utilization skills, and communication skills. The five domains use the same 1–7 scale, but with anchors at the 1, 3, 5, and 7 positions clearly describing the details in each domain to reach each score. The authors have investigated the inter-rater reliability and construct validity in a very detailed study. In this study, intraclass reliability was found to be 0.590–0.613 for the overall score [11]. In addition, the inter-rater reliability of the problem-solving, leadership, and situational awareness fields were 0.474–0.626 [11]. The fields of resource utilization and communication showed lower intraclass coefficients ranging from 0.236 to 0.384 [11].

Translation of the Ottawa GRS into Italian may offer a solution. However, translation of an assessment instrument can be problematic. Translation of a reliable and valid English language instrument into Italian—no matter how carefully the translation is performed—is no guarantee of producing a reliable and valid Italian language tool. The subtleties of language can be extremely important, particularly when evaluation involves the very nebulous principle of assigning an objective measurement to an observed human performance. Unfortunately, direct assessment of reliability and validity of a new grading instrument can be very difficult and time consuming. Fortunately, however, comparison of an Italian translation to a proven English language tool offers a much simpler method: if the Italian language tool is reliable compared to the English language tool, then the Italian language tool can be expected to have a reliability and validity comparable to the English language tool.

The primary objective of the current study was to assess the reliability of an Italian language Ottawa GRS translation by comparing it to the English language GRS. The null hypothesis of no correlation between the Italian language GRS and English language GRS scores was tested against the two-sided alternative hypothesis of significant correlation between the scores in different languages.

A secondary objective of the study was to assess the validity of the Italian language GRS by comparing it to a skills score. The null hypothesis of no correlation between the Italian language GRS and the skills score was tested against the two-sided alternative hypothesis of significant correlation between the two scores.

Methods

The overall study design was a measure of the strength of association between the Italian language GRS compared to both the English GRS and the skills score. This association was measured by applying each of three scoring tools concurrently by separate examiners to simulation scenarios.

As all data were identified and reported in aggregate, the study was deemed exempt from formal institutional review by the Ethics Committee of the Universita' del Piemonte Orientale where all simulations took place.

To ensure face validity of the Italian language tool, the translation of the original GRS from English to Italian was performed through an iterative group process involving all authors. All authors are fluent in both English and Italian: three are native first language Italian (MV, LC, PLI) and one is native first language English (JMF). Appendix 1 shows the final tool produced by the iterative process.

The English language GRS tool was used exactly as originally published with no modifications.

A technical skills scoring sheet was also developed for each simulation scenario. These scoring sheets were simple checklists scored as 0 (no), 2 (yes), or 1 (yes, but incomplete). A sample technical skills scoring sheet is shown in Appendix 2.

The overall structure of the simulations was a convenience sample of a previously scheduled simulation session in the SimWar format as detailed by Okuda et al., where simulation participants work as a team, and teams advance to the next round of competition based on judgment using pre-determined scoring tools [12]. Study participants were recruited through mailing lists to reach residents and program directors throughout Italy. Participants registered on a voluntary basis in teams of four. Each team rotated through seven initial scenarios. Five of these scenarios involved simultaneous measurement of English and Italian GRS and were included in the study. This included three high-fidelity scenarios: neonatal (birth asphyxia), adult (acute pulmonary edema), and obstetric (eclampsia). In addition, there was a single medium fidelity station of adult cardiac life support

(cardiac arrest due to hyperkalemia). A virtual reality station scenario of triage of ten casualties following a motor vehicle collision was also included using the XVR simulator (E-semble, Delft, The Netherlands). The four highest ranked teams following the initial round of scenarios advanced to the semi-final high-fidelity simulation of traumatic brain injury management. The two highest ranking teams from the semi-final proceeded to the final scenario of cardiac arrest with peri-mortem c-section and neonatal resuscitation. Simulations were conducted in Italian in all cases.

During the simulation scenarios, teams were independently evaluated by three examiners using three separate tools: (1) Italian GRS, (2) original (English) GRS, and (3) skills score. All raters were faculty at the Universita' del Piemonte Orientale, and none were familiar with the GRS tool. Each rater was given a single tool for the entire study. Raters of the GRS in each language were not given access to the GRS of the other language. Raters using the Italian GRS were all first language Italian speakers. Raters using the English language GRS tool were all fluent in both Italian and English. Although for each scenario the three raters were in the same room, they were asked to complete their evaluations independently without consulting one another or discussing the content of their tool. The simulation raters did not change as the teams rotated through each scenario, so that in the end for each scenario all teams were rated by the same raters. Raters completed their ratings on paper that was then collected after each scenario.

Data were entered into a MySQL (Oracle, Redwood, California, USA) database by one of the study authors (MV). Statistical analysis was performed using "R: A language and environment for statistical computing." (R Development Core Team, Vienna, Austria).

To assess the reliability of the Italian GRS, the results were compared to the English GRS using the Pearson–product correlation [13]. To assess the validity of the Italian GRS, the results were compared to the skill score, again using the Pearson–product correlation.

As suggested by Litwin, in 'How to Measure Survey Reliability and Validity', correlation coefficients greater than 0.7 were considered adequate [14]. Two-sided alternative hypotheses were used in all cases.

In total, seven statistical tests were performed. To account for multiple testing, p values were adjusted using the Holm method, and confidence intervals were adjusted using the Bonferroni method to ensure that a family-wise error rate of 0.05 was maintained [15]. The statistical hypotheses and methodology were completely specified prior to any data collection, and no modification was permitted after data collection. Post hoc exploratory data analysis was performed for hypothesis generation only and clearly identified as such. Thus, no confidence intervals or p values were obtained during the post hoc analysis.

To hold to the highest standards for reproducibility as described by Peng and to encourage further research in this area, both the raw data and the R code for the statistical analysis are available for download [16, 17].

Results

Twenty-eight residents took part in the study: seven teams of four residents each. Residents came from eight different Italian universities (Bari, Chieti, Genova, L'Aquila, Messina, Novara, Padova, and Sassari). Median resident age was 30 years and median postgraduate training years was three. Residents came from a variety of training programs including anesthesia (16 residents), emergency medicine (7), cardiology (2), internal medicine (1), pulmonary (1), and geriatrics (1).

In total, 41 simulations were evaluated: seven teams completed each of the simulations in the initial round, four teams participated in the semi-final, and two teams participated in the final. There were 123 completed scoring observations: three for each simulation. In total, 19 raters participated: 12 raters completed GRS scores (6 in Italian and 6 in English), while 7 raters performed skill scores. As not all teams completed the same number of simulations, and not all raters performed the same number of ratings, the final study design was not strictly orthogonal [18].

The correlation coefficient between the Italian language overall GRS and the English language overall GRS was 0.82 (adjusted 95 % confidence interval 0.62–0.92) as shown in Fig. 1. The adjusted p value for significant correlation was less than 0.000001. The figure also shows that

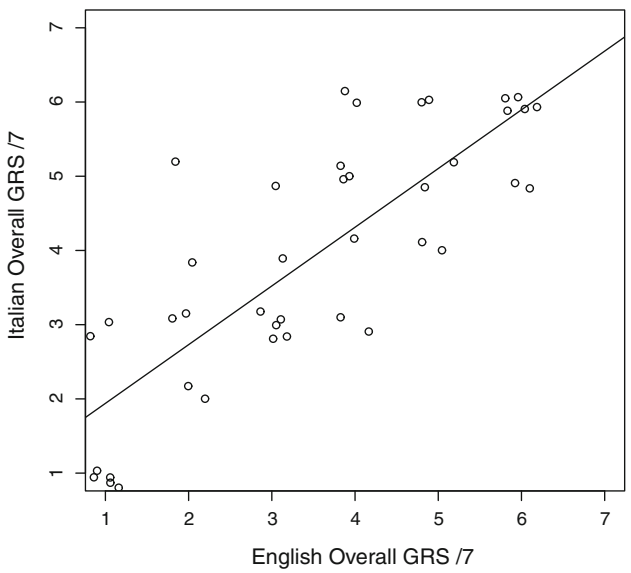

Fig. 1 Correlation between Italian and English overall global rating scale (GRS)

Table 1 Correlation between Italian and English global rating scale (GRS)

	GRS field	Correlation	Adjusted 95 % confidence interval	Adjusted *p* value
1	Leadership	0.82	0.61–0.92	<0.000001
2	Solving	0.83	0.61–0.93	<0.000001
3	Awareness	0.81	0.61–0.92	<0.000001
4	Utilization	0.73	0.46–0.88	<0.000001
5	Communication	0.71	0.42–0.87	<0.000001

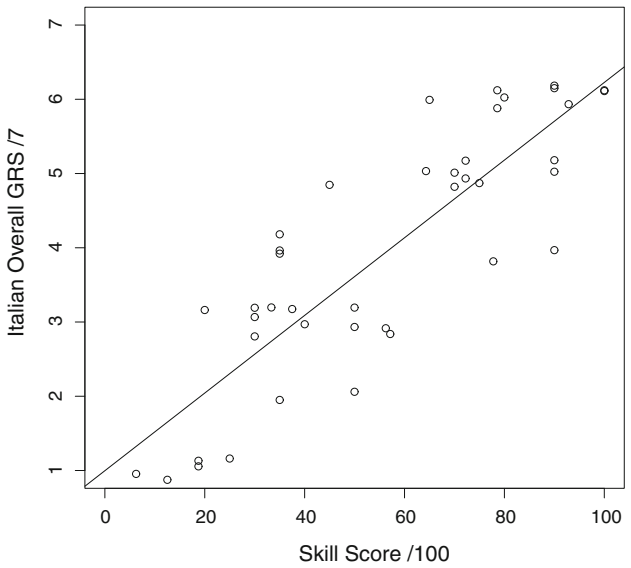

Fig. 2 Correlations between Italian overall global rating scale (GRS) and skill score

the range of assigned values was distributed well across all values from 1 to 6, and that correlation appears to hold for values throughout the range. Correlations between the other five domains of the Italian GRS and the English GRS are seen in Table 1.

The correlation coefficient between the Italian language overall GRS and the skill score was 0.85 (adjusted 95 % confidence interval 0.68–0.94) as shown in Fig. 2. The adjusted *p* value for significant correlation was <0.000001. As the total number of questions on each technical skills sheet was not the same, the sheets were scored as a percentage of the maximum total marks. The figure also shows that values on the skill score were also distributed quite evenly across the range from low to high and that correlation tends to hold for values throughout the range.

Chronbach's alpha for the Italian language GRS was 0.97, although this calculation was not part of the originally planned study analysis.

Discussion

For the primary study objective of assessing the reliability of the Italian language GRS in comparison to the English language GRS, the null hypothesis of no correlation was rejected. Thus, the study demonstrates that the Italian version of the Ottawa GRS is sufficiently reliable in comparison to the English language tool. Given the lack of current Italian language gold standard, this represents an important achievement.

Interpretation of the value of the correlation coefficient is somewhat problematic. As there was only one rater in each language, it is impossible to determine if differences between the scoring of the two raters is due to differences in the tool (different language), or due to normal inter-rater differences. While the estimates of correlation between the Italian language and English language varied in this study from 0.71 to 0.82, the English language GRS alone was found to have intraclass correlation coefficients between 0.24 and 0.62 [11]. Notably, although the authors stated that the Ottawa GRS had "acceptable inter-rater reliability," in none of the evaluation fields was the intraclass correlation greater than 0.7 [11]. Thus, this study suggests that correlation between the Italian language and English language GRS is similar or better than the correlation between two English language observers, suggesting that the Italian language GRS is at least as reliable as the English language tool. The adoption of objective and validated assessment tools is paramount for correct measurement of performance in any medical field. Downing (2004) suggests that evaluation instruments should have a reliability of >0.9 for "high-stake" evaluations such as licensing, 0.8–0.89 for summative examinations such as end of year evaluations, and 0.7–0.79 for informal formative classroom-type examinations. Overall, it appears that the Italian language GRS has acceptable reliability which is equal to the English language Ottawa GRS [9].

For the secondary study objective of assessing validity of the Italian language GRS, the null hypothesis of no correlation between the Italian language GRS and the skill scores was rejected. Thus, it appears that there is significant correlation between the two. The point estimate of correlation (0.85) suggests good correlation, although the lower limit of the confidence interval (0.68) is slightly below the desired lower limit of 0.7. Behavioral performance and adequacy of actions taken from a technical perspective have been studied in detail in the past decade [19–21]. The findings of this study are consistent with those of Riem et al., who showed that residents who demonstrated good performance in technical skills also showed good performance when exposed to an inter-operative crisis scenario [22]. Brunckhorst et al. demonstrated a similar correlation

between technical and non-technical skills [23]. Overall, the study suggests that the Italian language GRS tool has adequate reliability for assessing simulation performance.

Post hoc analysis of Chronbach's alpha, which should be used for hypothesis generation only as it was not initially part of the planned study analysis, was high (0.97). This suggests that the Italian language GRS may meet the criteria for excellent reliability in the form of internal consistency, adequate for even high-stake examinations [9]. These post hoc analyses should be of course confirmed in dedicated studies.

Unfortunately, there are several limits to the present study. Firstly, although the null hypotheses were rejected, and the point estimates for all correlation coefficients were greater than 0.7, the confidence intervals were very wide, and in all cases the lower limit of the estimated correlation was below 0.7. As the p values were small in all cases, the study suggests that although significant correlation exists, the present study may have been too small to be entirely confident that the correlation coefficient was above 0.7. Clearly, further research would be needed. In addition, the study was performed in the context of team simulations, where the raters graded the simulation team as a whole. It is possible that there is additional variability introduced into the study, as different team members likely had different levels of performance. Repeating the study with evaluation of individuals rather than teams may be insightful to further categorize this variability. Finally, although the study shows that the Italian GRS tool has good reliability in comparison to the English language tool—and should share similar validity and reliability—further studies to measure directly the reliability and validity of the Italian tool are clearly required before use in high-stake situations.

Conclusion

The present study suggests that the Italian language GRS translation has reasonable reliability when compared with the English language GRS, and reasonable validity when compared with the assessment of the skills scores. Data suggest that the instrument is adequately reliable for informal and formative type of examinations such as classroom assessment, and perhaps end of year examinations, but may require further confirmation before used for high-stakes examinations such as licensing.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

Statement of human and animal rights All procedures performed in studies involving human participants were in accordance with the ethical standards of the institution an/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Ericsson KA, Krampe RT, Tesch-Römer C (1993) The role of deliberate practice in the acquisition of expert performance. Psychol Rev 100:363–406
2. Gibbs G (1988) Learning by doing: a guide to teaching and learning methods. FEU
3. Grant J (1992) Training senior house officers by service-based learning. Joint centre for education in medicine, London
4. Kolb DA (2014) Experiential learning: experience as the source of learning and development. Pearson Education
5. Levine AI, Schwartz AD, Bryson EO, Demaria S (2012) Role of simulation in US physician licensure and certification. Mt Sinai J Med 79:140–153
6. Buyske J (2010) The role of simulation in certification. Surg Clin North Am 90:619–621
7. Gordon JA, Tancredi DN, Binder WD, Wilkerson WM, Shaffer DW (2003) Assessment of a clinical performance evaluation tool for use in a simulator-based testing environment: a pilot study. Acad Med 78:S45–S47
8. Epstein RM (2007) Medical education—assessment in medical education. N Engl J Med 356:387–396
9. Downing SM (2004) Reliability: on the reproducibility of assessment data. Med Educ 38:1006–1012
10. Downing SM (2003) Validity: on the meaningful interpretation of assessment data. Med Educ 37:830–837
11. Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J (2006) A pilot study using high-fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: the University of Ottawa Critical Care Medicine, high-fidelity simulation, and crisis resource management I study. Crit Care Med 34:2167–2174
12. Okuda Y, Godwin SA, Jacobson L, Wang E, Weingart S (2014) SimWars. J Emerg Med 47:586–593
13. Devore J (2011) Probability and statistics for engineering and the sciences. Thomson Brooks/Cole, USA
14. Litwin MS (1995) How to measure survey reliability and validity. SAGE Publications, USA
15. Hsu J (1996) Multiple comparisons: theory and methods. CRC Press, USA
16. Peng RD (2011) Reproducible research in computational science. Science 334:1226–1227
17. Franc JM (2016) Reveal project. http://www.medstatstudio.com/studies/project.php?pid=15. Accessed 18 Mar 2016
18. Montgomery DC (2009) Design and analysis of experiments. Wiley, USA
19. Flin R, Yule S, Paterson-Brown S, Maran N, Rowley D, Youngson G (2007) Teaching surgeons about non-technical skills. Surgeon 5:86–89
20. Powers KA, Rehrig ST, Irias N, Albano HA, Malinow A, Jones SB, Moorman DW, Pawlowski JB, Jones DB (2008) Simulated laparoscopic operating room crisis: an approach to enhance the surgical team performance. Surg Endosc 22:885–900

21. McCulloch P, Mishra A, Handa A, Dale T, Hirst G, Catchpole K (2009) The effects of aviation-style non-technical skills training on technical performance and outcome in the operating theatre. Qual Saf Health Car 18:109–115
22. Riem N, Boet S, Bould MD, Tavares W, Naik VN (2012) Do technical skills correlate with non-technical skills in crisis resource management: a simulation study. Br J Anaesth 109:723–728
23. Brunckhorst O, Shahid S, Aydin A, Khan S, Mcllhenny C, Brewin J, Sahai A, Bello F, Kneebone R, Shamim Khan M, Dasgupta P, Ahmed K (2015) The relationship between technical and nontechnical skills within a simulation-based ureteroscopy training environment. J Surg Educ 72(5):1039–1044