EDUCATION
RESEARCH SERVICE

Original Article

Assessment of a generalizable methodology to assess learning from manikin-based simulation technology*

Dominic A. Giuliano, DC and Marion McGregor, DC, PhD

Objective: This study combined a learning outcomes-based checklist and salient characteristics derived from wisdom-of-crowds theory to test whether differing groups of judges (diversity maximized versus expertise maximized) would be able to appropriately assess videotaped, manikin-based simulation scenarios.
Methods: Two groups of 3 judges scored 9 videos of interns managing a simulated cardiac event. The first group had a diverse range of knowledge of simulation procedures, while the second group was more homogeneous in their knowledge and had greater simulation expertise. All judges viewed 3 types of videos (predebriefing, postdebriefing, and 6 month follow-up) in a blinded fashion and provided their scores independently. Intraclass correlation coefficients (ICCs) were used to assess the reliability of judges as related to group membership. Scores from each group of judges were averaged to determine the impact of group on scores.
Results: Results revealed strong ICCs for both groups of judges (diverse, 0.89; expert, 0.97), with the diverse group of judges having a much wider 95% confidence interval for the ICC. Analysis of variance of the average checklist scores indicated no significant difference between the 2 groups of judges for any of the types of videotapes assessed ($F = 0.72$, $p = .4094$). There was, however, a statistically significant difference between the types of videos ($F = 14.39$, $p = .0004$), with higher scores at the postdebrief and 6-month follow-up time periods.
Conclusions: Results obtained in this study provide optimism for assessment procedures in simulation using learning outcomes-based checklists and a small panel of judges.

Key Indexing Terms: Education; Manikins; Reproducibility of Results

J Chiropr Educ 2014;28(1):16–20 DOI 10.7899/JCE-13-31

INTRODUCTION

In today's healthcare education, high-fidelity manikin-based simulation provides necessary, structured opportunities for practice before real patient interaction.[1–3] Understanding the benefits of such practice, however, requires strong evaluation procedures and comparative assessment of learning retention. The assessment of clinical performance in a simulated learning environment remains an area of concern requiring further investigation.[4] The uniqueness of the simulated teaching environment has made the assessment of learning and the development of an evaluation tool challenging to accomplish and to quantify.[4,5]

It has been suggested by Murray et al[6] that evaluation of simulated patient care encounters is relatively reliable and shows improved learning retention. This assertion was based largely on the work of Schwid et al[7] and Sica et al[8] in cardiac life support and radiology, respectively. Generalizability of these results to the wider simulation-based community, such as primary care medicine, chiropractic care, nursing, and other medical specialties, remains unknown.

Indeed, reliability of assessment and learning retention are key constructs in understanding the cost–benefit ratios associated with implementing simulation-based clinical education. While it has been suggested that high-fidelity manikin-based simulation laboratories are extremely expensive, Harlow and Sportsman[9] pointed out that any economic analysis on their use should include consideration of externalities, such as increased clinical competence

*This paper was selected as a 2013 Association of Chiropractic Colleges-Research Agenda Conference Prize Winning Paper - Award funded by the National Board of Chiropractic Examiners.

and patient safety. Society's concern with public safety, cost effectiveness, and effectiveness of treatment has called for more efficacious ways to teach and train health care providers.[5] Simulation is suggested as a tool to accomplish this by immersing students into simulated clinical scenarios without requiring interaction with actual patients.[4,5]

Unfortunately, there are peculiar challenges to assessment of clinical competencies related to manikin-based simulation experiences, making a single form of evaluation difficult to achieve. Simulation experiences, while potentially based on the same unfolding events (eg, a cardiac arrest), may have more than one successful outcome, and may require different competencies depending on the scope of practice of the discipline using the simulation in its training programs. Simulation scenarios usually are created to match the unique concerns of individual teaching environments. Thus, the correct responses of an anesthesiologist to a cardiac crisis would be expected to be different from the correct response by the chiropractor who identifies the event in his or her office. Again, the adequacy of clinical performance evaluation in a simulated learning environment remains an area of concern requiring further investigation.[4]

In an effort to manage this concern, a generalizable methodology for evaluation of manikin-based simulation events is proposed. An evaluation method that can be transferred from scenario to scenario, then, would not require individualized reliability and validity measures before implementation. The purposed methodology for this investigation is based on the creation of a checklist tool related to the intended learning outcomes for a given simulation, and the use of a panel of judges to score the checklist. It was hypothesized for this investigation that by defining the learning outcomes and using theory related to the wisdom-of-crowds,[10] sufficient reliability between examiners and validity associated with measuring different learning outcomes could be achieved. Support for this hypothesis would provide evidence for the use of such an evaluation method for consideration in a variety of clinical contexts.

The wisdom-of-crowds theory suggests that the average of independent judgments will be more accurate than either an individual assessment or group consensus. Historical research shows that simply averaging the results of a number of independent and earnest attempts at estimation will yield more accurate results than accepting the evaluation of the average individual.[11–13] Certain conditions, however, apply. For example, assessments must be free from influence. Group decisions are widely understood to be biased due to a variety of factors, including the importance of group cohesion.[14] Lorenz et al[15] found that even mild social influence could affect accuracy on estimation tasks. In addition, it has been noted, that crowds or a panel of judges are only wiser than individuals when the crowd or panel is comprised of relevant expertise and contains diverse perspectives.[16]

Therefore, the purpose of this study was 2-fold. First, it was to combine the interests of learning outcomes, with the salient characteristics derived from wisdom-of-crowds theory to test whether judges would be able to appropriately assess videotaped, high-fidelity manikin-based simulated events. The second purpose was to determine if differing groups of judges (one with diverse expertise maximized and one with expertise maximized) would provide differing results. Information from this investigation is intended to inform a generalizable set of methods that can be used to guide assessment needs in the wider simulation community.

METHODS

This study was approved by the research ethics board of Canadian Memorial Chiropractic College. All 185 undergraduate student interns participated in a two-hour simulation lab as a mandatory part of the curriculum. The lab detailed a cardiac event in a practitioner's office. This was one simulation scenario from the available bank of scenarios used at this institution. All scenarios and lab procedures at this institution have been crafted based on simulation technology[1] and the theory of emotional learning.[17] A full description of the simulation scenario is available upon request of the authors.

Interns entered the lab and were assigned randomly to a role in the simulation. Roles included clinician, patient in the waiting room, receptionist, and so forth. Although an orientation was provided to the lab itself, which included discussion of lab procedures and manikin handling, interns were given no information regarding what clinical situation to expect or how to respond. All interns had been through an emergency procedures class in the usual curriculum which included the necessary information for handling the case; however, the lab was not associated with this class.

Upon completing the scenario for the first time, interns were debriefed. This consisted of providing feedback and information with respect to appropriate case management. The video recording captured before the debriefing period is hereafter referred to as the predebriefing simulation.

After the debriefing period, interns completed the same scenario a second time. Hereafter, the video recording of this time point is referred to as the postdebriefing simulation.

In addition, for this study, interns reentered the lab six months later to perform a follow-up scenario. No additional briefing or reminders regarding the intended learning were provided to interns during the time between the postdebriefing simulation and the 6-month follow-up. At follow-up, a final video was captured. All video recordings were approximately 3 to 7 minutes in length.

A total of 69 videos were available for evaluation: 23 were from the predebriefing events, 23 were from the postdebriefing events, and 23 were follow-up videos completed using the same scenario six months later. From the 69 videos, a total of 9 were chosen for review. Videos were selected based on stratified random sampling. As such, the 23 videos from the predebriefing events were placed in a sample pool. A computerized random numbers generator was used to select three of these videos for evaluation. The same procedure was used to select three

Table 1 - Average Score for Each Group Across Each of the Three Types of Videos

Group[a]		Video Type*			
		Pre Debrief	Post Debrief	6-Month F/U	Total
		$n = 3$	$n = 3$	$n = 3$	$n = 9$
Diverse[b]	Mean	7.89	15.22	15.89	13
	SE	2.75	1.18	0.29	1.55
Expert[c]	Mean	6.78	13.78	15	11.85
	SE	2.79	0.91	1.2	1.57
Total	Mean	7.33	14.5	15.44	12.43
	SE	1.77	0.74	0.59	1.08

* $F = 14.39$, $p = .0004$.
[a] $F = .72$, $p = .4094$.
[b] *ICC = 0.89.*
[c] *ICC = 0.97.*

videos from each of the postdebriefing and follow-up video pools.

Based on predefined curriculum learning outcomes for each scenario and the content covered in the debriefing part of the lab, a checklist was created in an attempt to quantify student success relative to the outcomes. Thus, for this cardiac simulation event anticipated gold standard behaviors and technical skills were identified and used. This finalized checklist is available upon request from the authors.

The checklist was comprised of 16 items, 15 requiring a yes or no response and the last evaluating overall intern performance (ranging in score from 0 = poor to 3 = very good). Each of the first 15 items was scored as 1 for yes and 0 for no, based on the recognition that each of the procedures and technical skills were equally important in completing the gold standard protocol. The final question was weighted more highly to allow an overall subjective assessment of the interns' performance beyond the gold standard items, and in recognition that a variety of subjective considerations could be used. A total possible score, therefore, for the checklist was 18 for each video.

To evaluate the impact of differing expertise on evaluation, two groups of three judges were chosen to implement the checklist assessment. Three judges per group were deemed appropriate based on the work of Larrick and Soll.[18] These investigators pointed out that as long as there is one instance in a group where some judges score higher and some lower than the actual score (bracketing the truth), the average of the group should be more accurate than the average judge alone. In particular, Larrick and Soll[18] note that if there are only two judges whose errors are unbiased, normally distributed, and uncorrelated, the likelihood of bracketing the truth is 50%. In this investigation then, including three judges enhances the chance of a more accurate score. In addition, as per the suggestion of Libby and Glass,[19] the small panel size reflects a balance between the cost of judges and concerns regarding the risk of error. For this study, the risk of error was considered low, reflecting one of many forms of performance judgment that students undergo, while the cost of judges, who were all clinicians with considerable content knowledge of the scenario, was considered high.

All judges had the necessary clinical experience and understanding to evaluate students in this scenario. All were licensed practitioners and faculty members of the institution. Judges were chosen, however, on the basis of their expertise in simulation experiences. The first group of three judges, hereafter referred to as the diverse group, was comprised of one member with considerable simulation knowledge, a second member with minimal simulation experience, and a final member with no knowledge regarding the simulation process. The second group of three judges, hereafter referred to as the expert group, was comprised of three judges, all with the same high level of experience with the simulation process. Although none of the expert participants was credentialed in the creation and formation of simulation events, all had participated in scenario development, and had already observed a variety of simulated events in the lab.

To prevent even minimal social bias,[15] all judges in each group viewed the prerecorded videos in an independent and blinded manner. For example, judges were isolated from each other and, therefore, had no knowledge of each other's scores while viewing the videos. In addition, judges did not discuss scores or observations of intern performance before or after their assessments.

In addition, the nine prerecorded videos were assessed by all judges in the same random order to minimize the impact of any potential learning effects. Each group of judges was given instruction on the content and scoring of the checklist. Examples were provided to explain the basis of their potential responses during the actual assessment of the recordings, and judges had the opportunity to clarify any areas of concern. Judges also provided feedback after completing their assessments, regarding the clarity of the checklist and its ease of use. Members of the diverse group commented during feedback about the order of the items and thoughts regarding orientation enhancements. Small changes were made on the basis of this feedback when the assessments were rerun with the expert group.

Scores out of 18 on the checklist were converted to percentages to complete the analysis of interest. Analysis was based on individual and average intraclass correlation coefficient (ICC) scores and confidence intervals (CIs) for each of the two groups of judges. As well, a 2-way analysis of variance (ANOVA) was conducted to determine if there was a statistically significant difference in the scoring patterns between the two groups or across the three different videos times (predebriefing, postdebriefing, and 6-month follow-up).

RESULTS

Table 1 provides the average score for each group across each of the three types of videos (predebriefing, postdebriefing, and 6-month follow-up). The ANOVA indicated that, while there was a statistically significant

difference between the types of videos ($F = 14.39$, $p = .0004$) there was no difference in the average scores between the two groups of judges ($F = 0.72$, $p = .4094$). Post hoc Scheffe test[20] (on the main effect of video type) indicated a statistically significant difference between the predebrief and postdebrief scores ($p = .002$), and between the predebrief and 6-month follow-up scores ($p = .001$). From Table 1 it can be seen that for the diverse and expert groups, the predebrief scores were almost half those of the postdebrief and follow-up videos.

Reliability of the groups was assessed using ICC values and CIs. The ICC for the diverse group between individual measurements was 0.74 (95% CI, 0.36–0.93). The individual ICC for the expert group was 0.92 (95% CI, 0.79–0.98). When considering only the average measurement, the ICC for the diverse group was 0.89 (95% CI, 0.62–0.97) and the ICC for the expert group was 0.97 (95% CI, 0.92–0.99).

DISCUSSION

The assessment of simulation events in healthcare education poses a unique problem as the skills and scenarios being addressed can be highly variable, and judgments regarding student behavior and skill level can be very subjective. Simulation processes, however, are costly and time-consuming, and, therefore, it is critical that educational benefits be documented to manage curriculum evolution appropriately.[21–23]

Gaba et al[24] measured the reliability of multiple judges viewing taped simulation experiences of malignant hypothermia and cardiac arrest, using checklist instruments, as far back as 1998. Finding fair to excellent agreement among assessors, the team suggested the continued need for multiple judges to decrease measurement error. In 2002, Murray et al,[6] however, challenged the need for multiple judges, and concluded from their data that multiple assessments of student achievement would not significantly decrease measurement error. This is consistent with the work by Swanson et al,[25] indicating that measurement error also was dependent on the task being performed, and the interaction between person and task. Murray et al[6] have proposed that it is more important to present students with a diverse set of clinical scenarios from which to evaluate performance rather than evaluate performance using multiple raters.

While it is true that health care students should encounter many and diverse clinical scenarios, it is equally true that subjectivity and single-rater error can bias our understanding of learning achievement and retention significantly in any given situation. It has been shown that averaging the independent assessments of a panel of judges provides a superior estimate of truth when compared to individual assessment.[18] This is critical, not only as feedback to individual students, but also for decision-making in curriculum change and reform. Thus, the current research was focused on understanding the impact of panel diversity on the reliability of simulation evaluation by means of an outcomes-based checklist.

It has been suggested that, based on wisdom of crowds theory, errors are canceled out when a diverse set of judges are used.[16,26] Our results showed that both groups of judges (diverse and expert) had very high ICCs (individual ICC, 0.74 and 0.92; average ICC, 0.89 and 0.97, respectively). It was observed, however, that the diverse group had a much wider 95% CI for the ICC (0.36–0.93 and 0.62–0.97) than the expert group (0.79–0.98 and 0.92–0.99). The ANOVA detected no difference in the mean scores associated with group. Both groups scored the predebrief videos significantly lower than either the postdebrief or the 6-month follow-up videos.

Our results confirmed that a group of three independent judges comprised of diverse expertise in the simulation environment (but still with content expertise) can, on average, assess with equal accuracy as a group of three judges with greater and more homogeneous expertise. Given the wide CIs associated especially with the diverse group, the results of this study disagreed with the notions put forward by Murray et al[6] that suggest a single assessor is sufficient. It is understood that cost–benefit ratios require careful consideration of the use of manpower for assessment needs. Our work, however, indicated that even a small group of three judges is sufficient to provide accurate scores, reflecting new and retained learning. In addition, because the diverse group of judges performed equally as well as the expert group, where cost is an issue, savings can be considered by including lesser levels of expertise when a panel of judges is used. Finally, we concluded that the evaluation methodology used in this research should be reproduced in the more widespread simulation community. Creating a predefined, learning outcomes-based checklist and panel of three judges to evaluate video experiences can be expected to provide scores that are able to discriminate between levels of learning that are not topic-dependent.

Although the results of this investigation are highly promising, study limitations must be considered. Although we have no reason to believe that the group reliability results would differ from scenario to scenario, certainty can be derived only through future research. In addition, small modifications were made to the order of the checklist and to the instructions provided to the judges in the expert group. Although these differences were perceived to be minor, this may provide an alternate explanation for the smaller CI in that group. Nonetheless, such changes would not be expected to impact the average scores. Finally, the sample size of videos was quite small (9 simulation events). Given the consistency of the results between groups, however, the small sample size does not appear to have had a substantial impact.

CONCLUSIONS

In simulation, accurate and reliable tools to assess learning are needed, and the diversity of simulation scenarios and learning outcomes make generalizing evaluation methods difficult. The consistent assessment results obtained in this study provide optimism for a checklist tool combined with a panel of judges (either diverse or expert)

to be used to evaluate clinical performance and learning retention in a simulation event. Assessment of learning and learning retention are crucial to maintain in health care education, and these study results suggest optimism for accurate evaluation methods in the simulation environment.

CONFLICTS OF INTERESTS

There were no external sources of funding for this study, and no conflicts of interests were identified within this investigation.

About the Authors

Dominic Giuliano is the interim director of integrative learning and educational coordinator of the simulation lab, and Marion McGregor is the director of education for year II, both at Canadian Memorial Chiropractic College. Address correspondence to Dominic A. Giuliano, 6100 Leslie Street, Toronto, Ontario M2H 3J1, Canada; e-mail: dgiuliano@cmcc.ca. This article was received October 2, 2013, revised December 3, 2013 and Dec 10, 2013, and accepted on December 26, 2013.

REFERENCES

1. McGregor M, Giuliano D. Manikin-based clinical simulation in chiropractic education. *J Chiropr Educ*. 2012;26(1):14–23.
2. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach*. 2005;27:10–28.
3. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003–2009. *Med Educ*. 2010;44:50–63.
4. Harder BN. Use of simulation in teaching and learning in health sciences: a systematic review. *J Nurs Educ*. 2010;49(1):23–28.
5. Leape L, Berwick D, Clancy C. Transforming healthcare: a safety imperative. *Qual Saf Health Care*. 2009; 18(6):424–428.
6. Murray D, Boulet J, Ziv A, Woodhouse J, Kras J, McAllister J. An acute care skills evaluation for graduating medical students: a pilot study using clinical simulation. *Med Educ*. 2002;36(9):833–841.
7. Schwid HA, Rooke GA, Ross BK, Sivarajan M. Use of a computerized advanced cardiac life support simulator improves retention of advanced cardiac life support guidelines better than a textbook review. *Crit Care Med*. 1999;27:821–824.
8. Sica GT, Barron DM, Blum R, Frenna TH, Raemer DB. Computerized realistic simulation: a teaching manual for crisis management in radiology. *Am J Roentgenol*. 1999;172:301–304.
9. Harlow KC, Sportsman S. An economic analysis of patient simulators clinical training in nursing education. *Nurs Econ*. 2007;25(1):24–29.
10. Surowiecki J. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economics, Societies and Nations*. New York, NY: Random House Inc; 2004.
11. Galton F. Vox populi. *Nature*. 1907;75:450–451.
12. Knight H. *Reliability of Judgments: A Comparison of Group and Individual Judgments* [Master's thesis]. New York, NY: Columbia University; 1921.
13. Treynor JL. Market efficiency and the bean jar experiment. *Financ Anal J*. 1987;43:50–52.
14. Janis IL. Groupthink. *Psychol Today*. 1971;5(6):43–46, 74–76.
15. Lorenz J, Rauhut H, Schweitzer F, Helbing D. How social influence can undermine the wisdom of crowd effect. *Proc Natl Acad Sci U S A*. 2011;108(22):9020–9025.
16. Larrick RP, Mannes AE, Soll JB. The social psychology of the wisdom of crowds. In: Krueger JI, ed. *Frontiers of Social Psychology: Social Psychology and Decision Making*. New York, NY: Psychology Press; 2011:227–242.
17. Gordon J, Hayden E, Ahmed R, Pawlowski J, Khoury K, Oriol N. Early bedside care during preclinical medical education: can technology-enhanced patient simulation advance the Flexnerian ideal. *J Assoc Am Med Coll*. 2010;85:370–377.
18. Larrick RP, Soll JB. Intuitions about combining opinions: misappreciation of the averaging principle. *Manag Sci*. 2006;52(1):111–127.
19. Libby E, Glass L. The calculus of committee composition. *PLoS One*. 2010;5(9):e12642.
20. Scheffe H. A method for judging all contrasts in the analysis of variance. *Biometrika*. 1953;40:87–104.
21. Alinier G, Hunt B, Gordon R, Harwood C. Effectiveness of intermediate-fidelity simulation training technology in undergraduate nursing education. *J Adv Nurs*. 2006;54(3):359–369.
22. Gaba DM. The future vision of simulation in health care. *Qual Saf Health Care*. 2004;13(suppl 1):i2–i10.
23. Tuoriniemi P, Schott-Baer D. Implementing a high-fidelity simulation program in a community college setting. *Nurs Educ Perspect*. 2008;29(2):105–109.
24. Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anaesthesiology*. 1998;89:8–18.
25. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. *Educ Res*. 1995;24:5–11.
26. Simmons JP, Nielson LD, Galak J, Frederick S. Intuitive biases in choice versus estimation: implications for the wisdom of crowds. *J Consum Res*. 2011; 38(1):1–15.