Screen-Based Simulation for Training and Automated Assessment of Teamwork Skills

Comparing 2 Modes With Different Interactivity

Randolph H. Steadman, MD, MS;
Yue Ming Huang, EdD, MHS;
Markus R. Iseli, PhD;
John J. Lee, PhD;
Areti Tillou, MD, MSEd;
Maria D.D. Rudolph, MD;
Rachel Lewin, MA, PhD(c);
Alan D. Koenig, PhD;
Rukhsana Khan, MPH;
Federica Raia, PhD;
S. Michael Smith, PhD;
Yen-Yi Juo, MD;
Cameron Rice, MD;
Sophia P. Poorsattar, MD;
Noreen M. Webb, PhD

Introduction: The need for teamwork training is well documented; however, teaching these skills is challenging given the logistics of assembling individual team members together to train in person. We designed 2 modes of screen-based simulation for training teamwork skills to assess whether interactivity with nonplayer characters was necessary for in-game performance gains or for player satisfaction with the experience.
Methods: Mixed, randomized, repeated measures study with licensed healthcare providers block-stratified and randomized to evaluation—participant observes and evaluates the team player in 3 scenarios—and game play—participant is immersed as the leader in the same 3 scenarios. Teamwork construct scores (leadership, communication, situation monitoring, mutual support) from an ontology-based, Bayesian network assessment model were analyzed using mixed randomized repeated measures analyses of variance to compare performance, across scenarios and modes. Learning was measured by pretest and posttest quiz scores. User experience was evaluated using χ^2 analyses.
Results: Among 166 recruited and randomized participants, 120 enrolled in the study and 109 had complete data for analysis. Mean composite teamwork Bayesian network scores improved for successive scenarios in both modes, with evaluation scores statistically higher than game play for every teamwork construct and scenario ($r = 0.73$, $P = 0.000$). Quiz scores improved from pretest to posttest ($P = 0.004$), but differences between modes were not significant.
Conclusions: For training teamwork skills using screen-based simulation, interactivity of the player with the nonplayer characters is not necessary for in-game performance gains or for player satisfaction with the experience.
(*Sim Healthcare* 00:00–00, 2020)

Key Words: Simulation, screen-based simulation, virtual simulation, experiential learning, teamwork training, assessment, automated assessment.

Teamwork and communication failures between healthcare team members are responsible for up to 70% of medical errors.[1,2] Applying team skills in medical practice remains challenging as members of healthcare teams come from separate disciplines and isolated educational programs.

Training in teamwork skills has the potential to improve teamwork, clinical performance, and patient outcomes.[3–5] Recent reviews show that healthcare team training is effective for a variety of healthcare outcomes, including trainees' perceptions of the usefulness of team training, acquisition of knowledge and

From the Department of Anesthesiology and Critical Care (R.H.S.), Houston Methodist Hospital, Houston, TX; Department of Anesthesiology and Perioperative Medicine (Y.M.H.), David Geffen School of Medicine at UCLA; UCLA Simulation Center (Y.M.H.); National Center for Research on Evaluation, Standards, and Student Testing (CRESST) (M.R.I.); CRESST (J.J.L., A.D.K.), Los Angeles, CA; Department of Surgery (A.T.), David Geffen School of Medicine at UCLA; UCLA Center for Advanced Surgical & Interventional Technology Accredited Education Institute (A.T.), Los Angeles; Consultant, (M.D.D.R.) Claremont, CA; UCLA Graduate School of Education and Information Studies (R.L.); David Geffen School of Medicine at UCLA Dean's Office/UCLA Simulation Center (R.K.); UCLA Graduate School of Education and Information Studies (F.R.); Department of Medicine (F.R.), David Geffen School of Medicine at UCLA, Los Angeles, CA; Department of Culture & Communication (S.M.S.), Linköping Universitet, Linköping, Sweden; Department of Surgery (Y.-Y.J.), David Geffen School of Medicine at UCLA, Los Angeles; Community Memorial Health System (C.R.), Ventura; Department of Anesthesiology and Perioperative Medicine (S.P.P.), David Geffen School of Medicine at UCLA; and UCLA Graduate School of Education and Information Studies (N.M.W.), Los Angeles, CA.

Correspondence to: Randolph H. Steadman, MD, MS, Houston Methodist Hospital, 6565 Fannin St, B452, Houston, TX 77030 (e-mail: rsteadman@houstonmethodist.org).

The authors declare no conflict of interest.

Work should be attributed to the Department of Anesthesiology and Perioperative Medicine, David Geffen School of Medicine at UCLA.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.simulationinhealthcare.com).

DOI: 10.1097/SIH.0000000000000510

skills, demonstration of trained knowledge and skills on the job, and patient and organizational outcomes.[6] Neily et al[7] demonstrated an association between healthcare team training and reduced surgical mortality rate by as much as 18%.

Despite a growing literature examining the effect of teamwork training, there are limited data regarding how best to teach teamwork and communication skills to healthcare providers. Traditional team training sessions have consisted of classroom-based didactic presentations and/or resource-intense, immersive simulator-based programs requiring in person attendance and facilitated debriefing.[8,9] The publicly available Team Strategies to Enhance Performance and Patient Safety (TeamSTEPPS) curriculum developed by the Agency for Healthcare Research and Quality has finite registration capacity to train instructors.[10] The TeamSTEPPS online course can reach a wider audience.[11] However, there are limited opportunities to apply newly acquired skills within relevant contexts, repeat practice and feedback, and follow-up to assess skills acquisition and retention.

Our project addresses team training through development of an interactive screen-based serious game, a simulation platform with the capability for asynchronous customized learning and easy accessibility.[12] Our design is based on educational principles of cognitive engagement and builds on relevant models for skills acquisition and long-term retention, including simulation features that lead to effective learning.[13–15] We emphasize best practices and principles of team training effectiveness as well as game development as reviewed by experts in the field.[16–21] However, game design can be time consuming and costly, depending on the degree of interactivity the game player is given. Knowing whether a more interactive interface results in improved in-game performance and player satisfaction would be helpful in guiding game design.

Our research aim is to evaluate the usability, learning, and in-game performance differences between 2 modes of single-player screen-based simulated team training, which use the same scenarios but differ in interface interactivity and in user experience. The 2 modes manifest as the player observing and assessing nonplayer character performance (evaluation or EVAL mode) or as the player interacting with nonplayer characters (game play or GP mode). Our project scope is limited to comparing interactivity between 2 screen-based interfaces and does not compare with high-fidelity full-body simulator training.

METHODS

Study Design

This project, approved by the University of California Los Angeles (UCLA) Institutional Review Board, is a mixed randomized repeated measures design with an allocation ratio of 1:1 between the 2 modes of the game. For an estimated effect size of 0.5, a significance level of 0.05, and a power of 0.8, the resulting sample size calculation yielded 102 participants (51 per group).

Development

The screen-based team training application was developed in the following stages: (1) identification of teamwork constructs to be assessed; (2) identification of evidence required to infer selected teamwork constructs; (3) mapping between evidence and teamwork constructs to design an automated assessment model; (4), creation of scenarios and tasks that provide the defined evidence, using the Unity development platform (Version 5.6.5, 2017–2018; Unity Technologies, CA); and (5) iterative programming, prototyping, and testing.[18]

Based on extensive literature review, input from external consultants and focus group interviews (see Summary Report, Supplemental Digital Content 1, http://links.lww.com/SIH/A568, which details our process and findings), the team defined the 4 main teamwork constructs following the TeamSTEPPS model: leadership, communication, situation monitoring, and mutual support.[10] We identified 17 observable actions, which were independently rated according to their association with each of the 4 teamwork constructs by 12 members of the research team (see Table, Supplemental Digital Content 2, http://links.lww.com/SIH/A569, which describes the player actions). We assessed rater agreement by conducting a generalizability analysis for each teamwork construct.[22] The coefficient indicating the level of agreement among raters (index of dependability) was high for all 4 constructs, ranging from 0.90 to 0.93. The average mapping over all raters was used to define Bayesian network (BN) parameters (see Figure, Supplemental Digital Content 3, http://links.lww.com/SIH/A570, which illustrates our BN). Scenarios were scripted to engage the user in situations that assess understanding of the various teamwork actions and behaviors considering common or frequent pitfalls as well as knowledge or skill gaps.[23]

Bayesian Network Model of Assessment

For automated assessment of player performance, we chose a BN model. Bayesian networks are probabilistic graphical models that represent observable and latent (hidden) variables and their dependencies as directed graphs, where each graph node represents a variable and each arrow represents a direct dependence between a latent variable and an observable variable. The design of our BN was closely aligned with evidence-centered assessment design,[24,25] which connects a proficiency model (the 4 teamwork constructs) with an evidence model (the 17 observable actions).

Figure 1 shows the portion of the BN for the observable action closed-loop communication specifically. Closed-loop communication was mapped by 2, 12, 2, and 10 members to leadership, communication, situation monitoring, and mutual support, respectively, for a total of 26 ratings. The percentage contribution of each teamwork construct to closed-loop communication is 8% (2/26), 46% (12/26), 8% (2/26), and 38% (10/26). The total BN configuration includes similar mapping for all 17 observable actions. The average percentage contributions (arrows in BN) of each construct (averaged over all 17 observable actions) is shown in Table 1 for each scenario and mode. The comparability between GP and EVAL modes is high for each scenario.

The BN is used to calculate a player's probability of proficiency in each teamwork construct. As players proceed through the game, their performance related to each observable action updates their proficiency probability scores on the 4 constructs.

Bayesian networks fulfill our requirements of the following: (a) multidimensionality—enabling assessment of 4 different latent teamwork constructs; (b) interdependence between

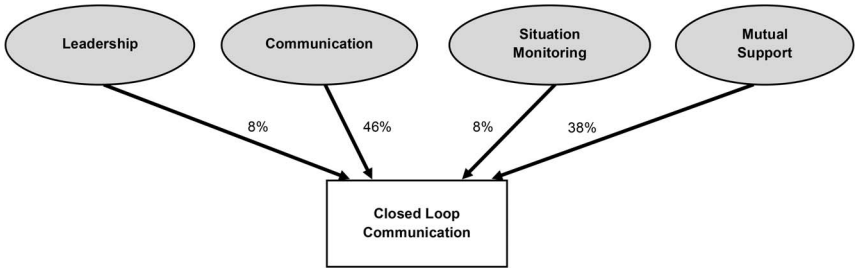

FIGURE 1. Sample BN excerpt for closed-loop communication. Dependency of closed-loop communication performance on the 4 teamwork constructs. Percentages indicate the relative contribution of each construct to closed-loop communication.

observable actions/tasks and latent constructs; and (c) probabilistic nature of our inferences. Bayesian networks have been widely used as system models in engineering applications to perform fault or diagnostic analysis[26] and as student models in educational applications[23,27] for intelligent tutoring systems.

Participant Recruitment

Eligibility criteria for participants included nursing (floor, operating room, emergency department, critical care), resident and attending physicians (anesthesiology, critical care, surgery, emergency medicine, internal medicine), and other allied health professionals including paramedics, pharmacists, and respiratory therapists. Participants had the option to perform the game on a computer at the simulation center or on their own personal computer with an Internet connection. Data were collected by an online server.

Eligible participants who responded to the recruitment e-mail were randomly assigned to either GP or EVAL mode and e-mailed the respective link to that mode. Randomization was accomplished by a block-stratified sequence generation design, based on professions and specialties, with a goal of 12 participants per group and equal distribution of the following professions: physicians, nurses, and others (paramedics/Emergency Medical Technicians, respiratory therapists, pharmacists). Physicians were additionally stratified based on level of experience (ie, attendings vs residents) and by specialty (anesthesiology, emergency medicine, surgery and internal medicine). There were 4 targeted specialties for nurses: operating room, floor, intensive care unit, and emergency medicine. Participants were randomized to 1 of 2 study conditions, EVAL or GP mode, by computerized coin flip generator. We enabled participants to contact coordinators about technical issues.

Although informed of the 2 possible modes (GP and EVAL), participants had not received descriptions for each mode and were blinded to the treatment arm to which they were assigned (they received a number code for login). The same home screen was used for both modes, and both groups went through the scenarios in the same order. Participants were asked to complete the game in one sitting to create comparable conditions. Performance ratings were generated by the BN-based automated assessment that was designed for the study and automatic scoring of the quizzes. There were no subjective ratings performed by researchers.

Intervention

Participants in both modes progressed through 3 healthcare scenarios in different settings: emergency department (Scenario 1), operating room (Scenario 2), and intensive care unit (Scenario 3; see Screenshots, Supplemental Digital Content 4, http://links.lww.com/SIH/A571, which depict scenes from each scenario). They were asked to implement teamwork skills in 4 areas: leadership, mutual support, communication, and situation monitoring. In GP mode, participants assumed the role of the team leader, whereas in EVAL mode, participants evaluated the decisions and actions of the nonplayer character leading the team. The medical knowledge required of these clinical environments was minimal, and cues were provided, as technical skills were not evaluated (diagnoses, evaluation results, and medical management cues were provided in both modes).

In GP mode, the participant selected actions to take, determined the timing/sequence of actions, and designated nonplayer characters to perform an action. Participants made these decisions prospectively, ie, before nonplayer characters initiating action. Whenever a player action was expected (eg, a response to a nonplayer character query), a countdown clock wound down and blinked noticeably after 20 seconds. Once the player chose an action, a pop-up window appeared asking the player to choose from 4 possible dialog options (that varied from best to least good). If the player failed to take a proper action within the allotted timeout, the player was prompted by leading dialog from a nonplayer character, and if still no action occurred, finally, a pop-up window appeared, which contained the same dialog choices as described previously. Game-play scoring was affected by the action selected, action timing, and appropriate dialog choice.

In EVAL mode, the participant observed scripted action and intermittently evaluated the actions taken by the team leader. Periodically, the scenario paused, and a multiple-choice question appeared on screen to assess the most appropriate decision, action, or dialog for the situation. Evaluation mode scoring was based on participant answers to prompted questions. After each scenario, all players in both GP and EVAL modes received an identical after action review (A.A.R.) that provided reflection and contextualized learning.

TABLE 1. Relative Contributions of Teamwork Constructs

	Scenario 1, %		Scenario 2, %		Scenario 3, %	
	GP	EVAL	GP	EVAL	GP	EVAL
Leadership	39	39	37	34	26	28
Communication	32	30	34	33	31	30
Situation monitoring	13	11	16	17	17	21
Mutual support	17	19	13	16	26	20
Total	101	99	100	100	100	99

Comparability between GP and EVAL modes is high for each scenario, indicating that teamwork actions were distributed similarly between the 2 modes.

There were intrinsic differences in how similar learning objectives were presented and assessed in each of the modes. Evaluation mode participants were asked to evaluate *retrospectively* for potential improvements in communication and teamwork, whereas GP mode participants *prospectively* determined the type and timing of actions and requests. Because the timing of the evaluations in EVAL mode were predetermined, EVAL mode did not take action timing into account, while GP mode did.

Outcomes

The primary measured outcome of this study was teamwork scores from the BN assessment model designed for the study (see Figure, Supplemental Digital Content 3, http://links.lww.com/SIH/A570, which illustrates our BN). Secondary measured outcomes included scores from completed pregame and postgame quizzes and surveys. Quiz questions were drawn from the TeamSTEPPS learning benchmark question set.[28] A survey using a scale of 1 (strongly disagree) to 4 (strongly agree) measured participant reactions regarding the user interface and overall usefulness of the screen-based training.

Statistical Analysis

Mixed randomized repeated measures analyses of variance were conducted for the 4 teamwork constructs, expressed by 4 BN proficiency variables, to compare in-game performance between GP and EVAL modes and to examine changes in performance over the 3 scenarios. Data were deidentified before analysis. Values for the BN proficiency variables range from 0 to 1 and indicate our inferred probability or belief of proficiency, given observed behavior: a value of 0 means the examinee has no proficiency, a value of 1 means the person has complete proficiency, and a value of 0.5 means that there is insufficient evidence to infer either presence or absence of proficiency. In addition, the same analyses were conducted for improvement scores between adjacent scenarios. Analyses

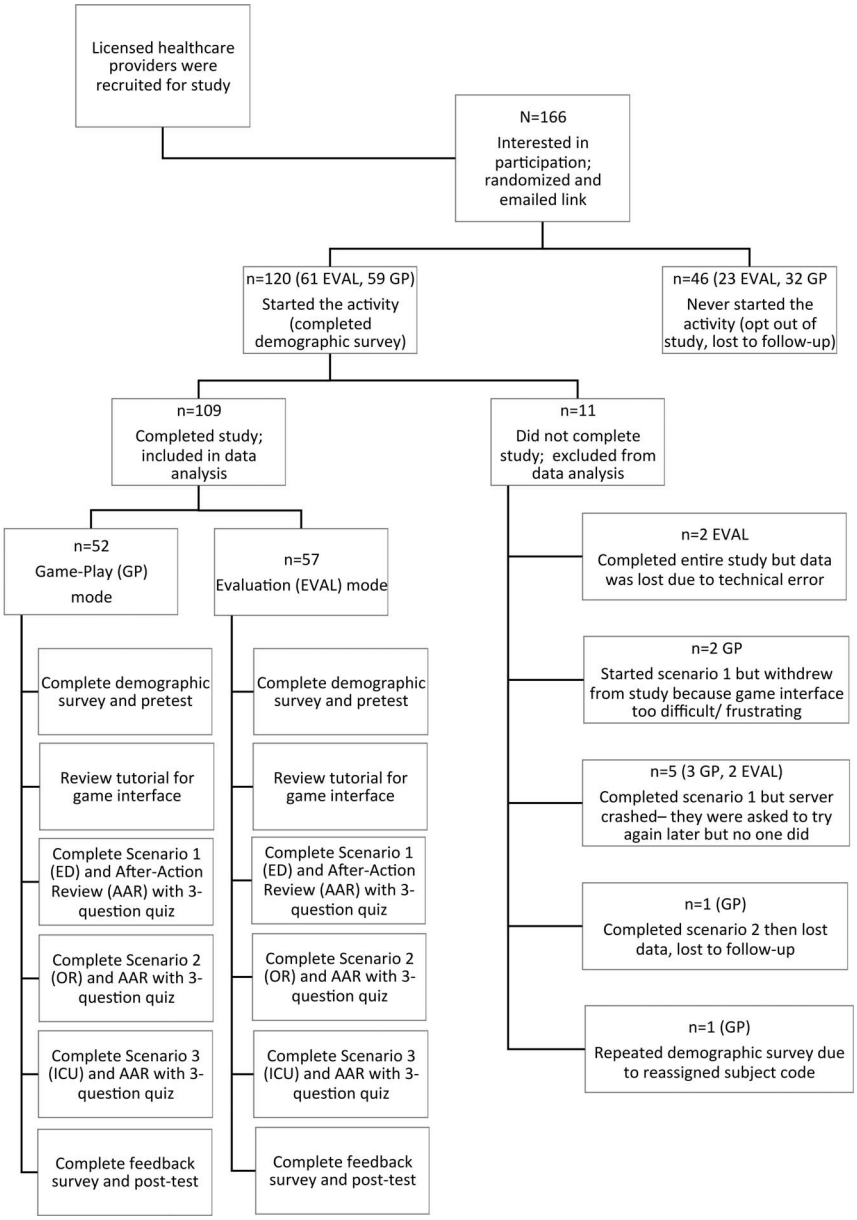

FIGURE 2. Study flow diagram.

TABLE 2. Study Participant Demographics

	GP, n = 52		EVAL, n = 57		Statistics	
	Count	% of Total	Count	% of Total	χ^2	*P*
Sex	n = 52		n = 57		0.150	0.699
Male	20	18	24	22		
Female	32	29	33	30		
Age	n = 51		n = 57		0.198	0.699
<20	0	0	0	0		
20–30	16	15	19	17		
31–40	21	19	22	20		
41–50	7	6	9	8		
51–60	3	3	4	4		
>60	4	4	3	3		
Profession	n = 52		n = 57		1.140	0.565
Physician	30	28	35	32		
Anesthesiology	16	15	17	16		
Emergency medicine	3	3	7	6		
Internal medicine	5	5	7	6		
Surgery	6	6	4	4		
Nurse	14	13	17	16		
Critical care	5	5	6	6		
Emergency medicine	3	3	3	3		
Floor	3	3	3	3		
Operating room	3	3	5	5		
Other	8	7	5	5		
Respiratory therapist	0	0	1	1		
EMT/paramedic	4	4	1	1		
Pharmacist	4	4	3	3		
Hours of video GP/week	n = 51		n = 56		0.845	0.655
None	28	26	32	29		
1–2 H	18	17	16	15		
2–5 H	3	3	6	6		
5–10 H	2	2	2	2		
>10 s	0	0	0	0		
Prior team training experience	33	30	41	38	0.651	0.420

After removing 13 subjects who did not finish the study in one sitting, the duration of the entire encounter was similar between participants in EVAL and GP modes (n = 96, EVAL: median = 1.39 hours, mean = 1.48 hours, SD = 0.5 hours; GP median = 1.19 hours, mean = 1.32 hours, SD = 0.5 hours). The difference between modes in duration was not statistically significant ($P = 0.13$).

were conducted for each of the 4 teamwork constructs separately as well as for an overall teamwork composite score (equally weighted combination of the 4 teamwork construct proficiency values). For significant mode × scenario interactions, simple main effects analyses were conducted and specific comparisons were examined with Bonferroni adjustments for multiple comparisons where appropriate.[29]

A mixed randomized repeated measures analysis of variance was conducted for average pretest and posttest quiz scores of teamwork knowledge to examine differences between the 2 modes and changes from pretest to posttest. Using data from the feedback survey, χ^2 analyses were used to compare modes in the pattern of self-reported ratings. All statistical analyses were performed using SPSS (Statistical Package for the Social Sciences) Statistics for Windows Version 25 (IBM Corp, 2017).

RESULTS

Figure 2 shows recruitment and randomization efforts. We randomized and e-mailed the link to the game to 166 licensed healthcare providers who expressed interest in study participation. Among these, 120 started the game with 109 finishing and providing complete data for analysis; 46 never started and were lost to follow-up. Of the 120 who attempted the activity, 61 were in the EVAL group and 59 in the GP group. Among the 46 who never started the activity, 23 had been randomized to EVAL and 32 to GP mode.

Participant randomization was achieved as shown in Table 2. Both EVAL and GP groups had equivalent numbers of participants in terms of age, sex, profession, hours of video gaming, and team training experience. After removing 13 subjects who did not finish the study in one sitting, the duration of the entire encounter was similar between participants in EVAL and GP modes (n = 96, EVAL: median = 1.39 hours, mean = 1.48 hours, SD = 0.5 hours; GP median = 1.19 hours, mean = 1.32 hours, SD = 0.5 hours). The difference between modes in duration was not statistically significant ($P = 0.13$).

Difference Between Modes in Teamwork Skills

Although BN proficiency values were high in both modes, scores in EVAL mode were higher than in GP mode for every teamwork construct and for every scenario (Figs. 3, 4). For the composite score (Fig. 3), differences between EVAL and GP means were 0.16 for Scenario 1 [simple main effect of mode: $F_{(1,107)} = 169.01$, $P = 0.000$, effect size (partial $\eta^2 = 0.61$)],

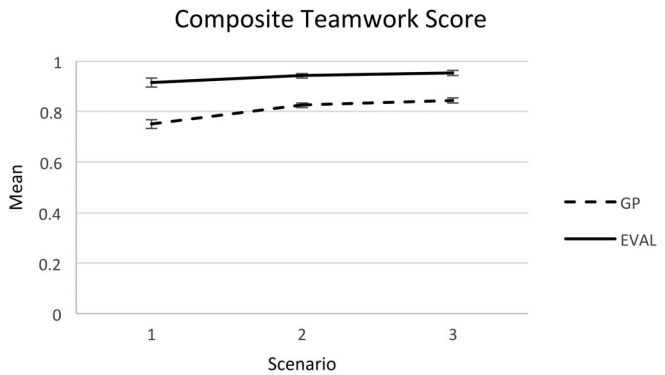

FIGURE 3. Composite teamwork score. Mean performance as inferred by BN proficiency values for the teamwork composite score (average of communication, leadership, situation monitoring, and mutual support proficiency values) for GP and EVAL modes. Error bars are 95% confidence intervals. Differences between EVAL and GP means were 0.16 for Scenario 1 ($F_{(1,107)} = 169.01$, $P = 0.00$, $\eta_p^2 = 0.61$), 0.12 for Scenario 2 ($F_{(1,107)} = 290.33$, $P = 0.000$, $\eta_p^2 = 0.73$), and 0.11 for Scenario 3 ($F_{(1,107)} = 219.37$, $P = 0.00$, $\eta_p^2 = 0.67$). η_p^2, effect size.

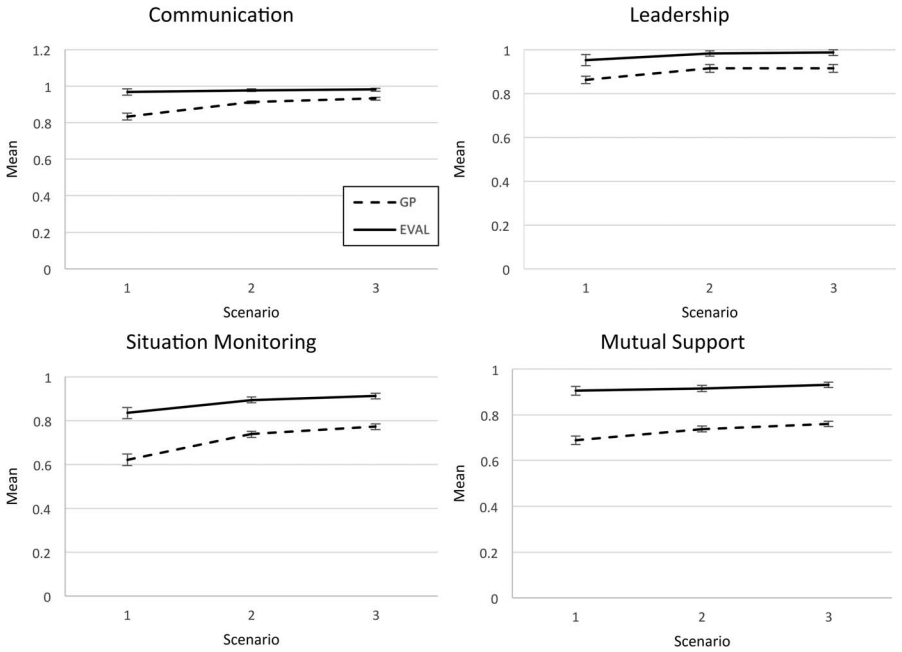

FIGURE 4. Teamwork construct scores. Mean performance as measured by BN scores for 4 measures of teamwork skills: communication (A), leadership (B), situation monitoring (C), and mutual support (D) for GP and evaluation modes. Error bars are 95% confidence intervals. Differences between EVAL and GP means ranged from 0.09 to 0.22 for Scenario 1 ($F_{(1,107)}$ = 22.25, P = 0.000 to $F_{(1,107)}$ = 259.91, P = 0.000, η_p^2 = 0.17 to 0.71), 0.07 to 0.18 for Scenario 2 ($F_{(1,107)}$ = 61.99, P = 0.000 to $F_{(1,107)}$ = 392.40, P = 0.000, η_p^2 = 0.37 to 0.79), and 0.05 to 0.17 for Scenario 3 ($F_{(1,107)}$ = 62.37, P = 0.000 to $F_{(1,107)}$ = 402.22, P = 0.000, η_p^2 = 0.37 to 0.79). η_p^2, effect size.

0.12 for Scenario 2 ($F_{(1,107)}$ = 290.33, P = 0.000, effect size = 0.73), and 0.11 for Scenario 3 ($F_{(1,107)}$ = 219.37, P = 0.000, effect size = 0.67). For the separate teamwork constructs (Fig. 4, communication, leadership, situation monitoring, mutual support), differences between EVAL and GP means ranged from 0.09 to 0.22 for Scenario 1 (simple main effects of mode: $F_{(1,107)}$ = 22.25, P = 0.000 to $F_{(1,107)}$ = 259.91, P = 0.000, effect sizes = 0.17–0.71), 0.07–0.18 for Scenario 2 ($F_{(1,107)}$ = 61.99, P = 0.000 to $F_{(1,107)}$ = 392.40, P = 0.000, effect sizes = 0.37–0.79), and 0.05–0.17 for Scenario 3 ($F_{(1,107)}$ = 62.37, P = 0.000 to $F_{(1,107)}$ = 402.22, P = 0.000, effect sizes = 0.37–0.79).

Quiz scores were high in both modes (Fig. 5) and improved from pretest to posttest overall (main effect for time: $F_{(1,68)}$ = 8.75, P = 0.004), although there was no significant difference between modes in the change from pretest to posttest (nonsignificant effect for mode: $F_{(1,68)}$ = 1.21, P = 0.276; nonsignificant mode × time interaction effect: $F_{(1,68)}$ = 1.71, P = 0.196).

Improvement in Teamwork Skills

Inferred BN proficiency values improved over the course of the scenarios (Figs. 3, 4), although the improvement pattern was nonlinear. Teamwork proficiency improved more from Scenario 1 to 2 than from Scenario 2 to 3. Moreover, the improvement pattern was different in the 2 modes; the quadratic effect for the scenario × mode interaction was statistically significant for the overall teamwork composite and each teamwork construct except leadership: $F_{(1,107)}$ = 10.32, P = 0.001 to $F_{(1,107)}$ = 39.21, P = 0.000. Specifically, the improvement

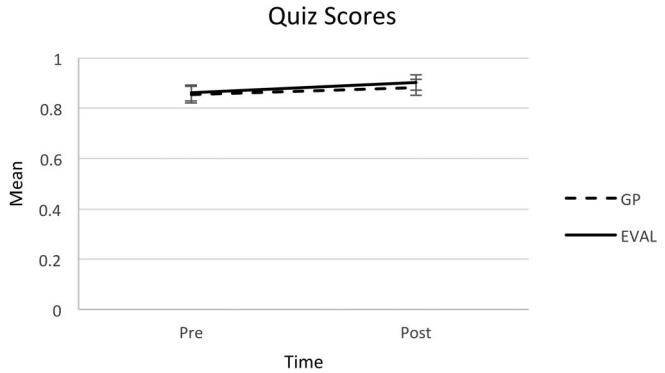

FIGURE 5. Quiz scores. Mean performance for quiz pretest and posttest scores. Error bars are 95% confidence intervals. Quiz scores were high in both modes and improved from pretest to posttest overall (main effect for time: $F_{(1,68)}$ = 8.75, P = 0.004), although there was no significant difference between modes in the change from pretest to posttest (mode effect: $F_{(1,68)}$ = 1.21, P = 0.276; mode × time interaction effect: $F_{(1,68)}$ = 1.71, P = 0.196).

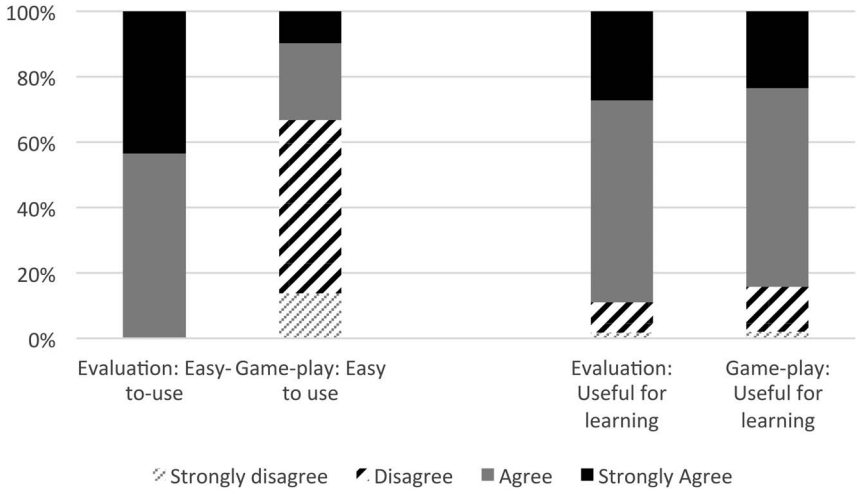

FIGURE 6. User experience. Self-reported user reactions show that EVAL mode was easier to navigate, with only a third of respondents in GP mode agreeing with the statement "The interface was easy to use" ($\chi^2(3) = 53.73$, $P = 0.000$), but both EVAL and GP modes were equally useful for learning ($\chi^2(3) = 0.25$, $P = 0.969$).

from Scenario 1 to 2 was greater in GP mode than in EVAL mode for the teamwork composite score and for each teamwork construct except leadership (pairwise comparisons comparing modes in improvement scores from Scenario 1 to Scenario 2 for communication, situation monitoring, mutual support: $F_{(1,107)} = 18.15$, $P = 0.000$ to $F_{(1,107)} = 53.49$, $P = 0.000$).

Participants' Ratings of the Training Experience

Whereas all respondents in EVAL mode agreed or strongly agreed with the statement that the interface was easy to use, only a third of respondents in GP mode agreed or strongly agreed ($\chi^2(3) = 53.73$, $P = 0.000$; Fig. 6). Despite the differences in reported ease of use, participants in the 2 modes did not differ significantly in their assessment of whether the training helped them learn teamwork skills ($\chi^2(3) = 0.25$, $P = 0.969$). Moreover, reported ease of use did not affect BN scores (Tables 3, 4). In GP mode, participants who reported the interface to be difficult to use and participants who reported the interface to be easy to use obtained very similar (and not statistically significantly different) BN scores. Similarly, BN scores of participants in EVAL mode did not differ according to their perceptions about the user interface.

DISCUSSION

We created and compared 2 modes of screen-based learning for the acquisition of teamwork skills. The primary finding of our study is that a high degree of interactivity may not be necessary for performance gains: learning occurred in both modes; however, participants in the less interactive mode (EVAL) exhibited higher levels of in-game performance. The level of interactivity did not affect participants' view of the usefulness of the experience. This is consistent with systematic reviews comparing observational roles to hands-on participation in scenario-based simulation.[30,31] Learner outcomes and role satisfaction for observers were as good or better than hands-on

TABLE 3. Mean BN Scores for GP Mode According to Participant Responses to the Survey Item: "The Interface Was Easy to Use"

GP Mode	Strongly Disagree (n = 7)	Disagree (n = 26)	Agree (n = 12)	Strongly Agree (n = 5)	*F*	*P*
	M (SD)	M (SD)	M (SD)	M (SD)		
Scenario 1						
Teamwork	0.76 (0.07)	0.75 (0.09)	0.77 (0.07)	0.74 (0.08)	0.25	0.86
Communication	0.84 (0.07)	0.83 (0.11)	0.85 (0.07)	0.83 (0.10)	0.11	0.95
Leadership	0.87 (0.11)	0.86 (0.11)	0.88 (0.11)	0.88 (0.06)	0.21	0.89
Situation monitoring	0.63 (0.06)	0.62 (0.10)	0.64 (0.10)	0.58 (0.09)	0.51	0.68
Mutual support	0.69 (0.09)	0.68 (0.10)	0.71 (0.07)	0.68 (0.11)	0.18	0.91
Scenario 2						
Teamwork	0.82 (0.05)	0.83 (0.04)	0.84 (0.03)	0.82 (0.04)	0.49	0.69
Communication	0.90 (0.04)	0.91 (0.04)	0.92 (0.03)	0.91 (0.04)	0.35	0.79
Leadership	0.90 (0.07)	0.91 (0.06)	0.93 (0.05)	0.93 (0.02)	0.44	0.72
Situation monitoring	0.73 (0.07)	0.74 (0.07)	0.76 (0.06)	0.72 (0.06)	0.61	0.61
Mutual support	0.73 (0.06)	0.74 (0.06)	0.75 (0.03)	0.73 (0.07)	0.24	0.87
Scenario 3						
Teamwork	0.83 (0.06)	0.85 (0.05)	0.86 (0.03)	0.85 (0.03)	0.57	0.64
Communication	0.91 (0.07)	0.94 (0.04)	0.94 (0.02)	0.94 (0.02)	0.84	0.48
Leadership	0.90 (0.07)	0.92 (0.05)	0.92 (0.05)	0.93 (0.03)	0.33	0.80
Situation monitoring	0.76 (0.09)	0.78 (0.06)	0.79 (0.05)	0.76 (0.04)	0.36	0.79
Mutual support	0.75 (0.05)	0.76 (0.05)	0.78 (0.04)	0.76 (0.07)	0.92	0.44

Two participants in GP mode did not complete the survey (total n = 50).

TABLE 4. Mean BN Scores for EVAL Mode According to Participant Responses to the Survey Item: "The Interface Was Easy to Use"

EVAL Mode	Agree (n = 31)	Strongly Agree (n = 24)	*F*	*P*
	M (SD)	M (SD)		
Scenario 1				
Teamwork	0.92 (0.03)	0.91 (0.06)	0.22	0.64
Communication	0.97 (0.02)	0.97 (0.03)	0.90	0.35
Leadership	0.95 (0.08)	0.95 (0.08)	0.03	0.87
Situation monitoring	0.84 (0.07)	0.83 (0.10)	0.14	0.71
Mutual support	0.91 (0.02)	0.91 (0.04)	0.60	0.44
Scenario 2				
Teamwork	0.95 (0.01)	0.94 (0.03)	0.95	0.34
Communication	0.98 (0.01)	0.98 (0.02)	0.66	0.42
Leadership	0.98 (0.02)	0.98 (0.02)	0.12	0.73
Situation monitoring	0.90 (0.03)	0.90 (0.04)	1.09	0.30
Mutual support	0.92 (0.02)	0.92 (0.03)	0.97	0.33
Scenario 3				
Teamwork	0.96 (0.01)	0.95 (0.02)	1.04	0.31
Communication	0.98 (0.01)	0.98 (0.01)	0.39	0.54
Leadership	0.99 (0.01)	0.99 (0.02)	0.46	0.50
Situation monitoring	0.92 (0.02)	0.91 (0.03)	1.16	0.29
Mutual support	0.94 (0.01)	0.93 (0.03)	1.21	0.28

Two participants in EVAL mode did not complete the survey (total n = 55). None of the participants responded with disagree or strongly disagree.

roles when the learner was engaged and given tools to direct their observation.[30] Although our EVAL mode provided more of an observer role for the learner, it was not passive learning. We directed learners to evaluate the role of the leader in the game, which activated the learner to pay attention to specific actions and answer questions. Role clarity may have been enhanced in the EVAL group, where learners were directed to discrete issues and tasks, rather than in GP mode, where learners had the option to decide which issue to tackle first. The more complex interface of the more "immersive" GP mode yielded a higher level of interactivity, which may have contributed to an increased extraneous cognitive load. Cognitive load theory has posited that collaborative learning and team training work best when task-unrelated transactive activities are minimized (in our case, the rules of engagement within the virtual team and navigating a new user interface) or when learners have prior knowledge or experience with those tasks.[32] Thus, the opportunity to apply teamwork skills during the in-game interactions may have been enhanced in EVAL mode because of role clarity and reduced cognitive load.

The positive analysis of EVAL mode has important implications for online scenario development, as EVAL mode is significantly less time consuming to script and program than GP mode. In GP mode, the script is dynamic with branch points determined by individual choices made by players. Although each branch point can be constrained to a finite number of options, the timing and sequence of players' selections are not predictable, which leads to even greater script and programming complexity in GP mode compared with EVAL mode, where the scenario is linear without branching. In addition, the user interface for GP mode is more complex and must contain the affordances sought by the player, listed in a way that is intuitive and unambiguous.

One strength of our study is the BN model for automated assessment, which was used in both modes. Another feature of our study is the choice of participants to allow generalizability of the findings to a variety of licensed healthcare providers. With appropriate modifications of the scenarios and learning objectives, we suspect that additional scenarios could be designed for nonlicensed individuals working in healthcare environments, including for onboarding to familiarize individuals with challenging situations and institutional expectations.

In addition, this study incorporates features not often present in research on screen-based simulation: an experimental design, moderately large sample size, and both affective reactions and learning outcomes. A recent systematic review assessing the use of virtual training for nontechnical skills showed that there were few studies published on this topic (median of 2 articles per year from 2010 to 2017). The average number of study participants in those studies were 40 and very few incorporated a pretest/posttest or group comparison, with most of the studies measuring usability and acceptability but not learning outcomes.[33]

Limitations

We elected not to compare traditional in-person training with screen-based training,[34,35] as our primary focus was to compare 2 distributive training modalities with different characteristics. As a result we are not aware of how the in-game performance gains (measured by BN scores), we report would compare with other types of team training of a similar duration. Whether live simulations would result in greater performance gains than screen-based simulation was not tested. It is possible that the EVAL approach yielded higher performance within the virtual environment but the GP approach could lead to improved teamwork-specific behaviors in actual practice. Although baseline imbalances in the participant demographics and/or teamwork skills may exist, stratified randomization by profession/discipline was used to ensure group comparability.

CONCLUSIONS

Online serious games or simulations, designed for healthcare learning, offer a number of advantages: they are scenario-based, engaging, accessible on demand and can be programmed to provide automated scoring and feedback. Online teamwork training may serve as an asynchronous simulation modality and primer for the more resource-intense in-person team training simulation sessions. However, obstacles exist in creating virtual simulations: programming needs are significant, development times are lengthy, and the necessary development expertise is extensive, including a mix of subject matter experts from the worlds of medicine, instructional design, game design, computing, and business. In addition, GP environments are not intuitive to first-time users. Efforts to identify necessary design elements can help tip the scales in favor of online gaming's advantages by making the process more efficient and less costly. We have shown that for training teamwork skills with a short (1–2 hour) screen-based simulation, interactivity of the player with the nonplayer characters is not necessary for in-game performance gains or for player satisfaction with the experience. Future work in screen-based simulation should be directed to longer duration encounters or repeated encounters integrated over the course of a curriculum. In addition, taking into consideration cognitive load

theory, work is needed to assess the features of screen-based simulation that improve the user experience and mimic more realistic interactivity, such as the ability to verbally speak with virtual team members (eg, using natural language processing) and enhancing engagement with immersive virtual reality. Understanding which game features affect clinical behaviors and outcomes will inform future game developers and determine return on investment.

REFERENCES

1. Kohn LT, Corrigan JM, Donaldson MS. *To Err Is Human: Building a Safer Health System. Committee on Health Care in America. Institute of Medicine*. Washington, DC: National Academy Press; 1999.
2. Joint Commission. Sentinel Event Data - Root Causes by Event Type 2004-2Q 2014. Available at: http://www.jointcommission.org/assets/1/18/Root_Causes_by_Event_Type_2004-2Q_2014.pdf. Published 2014. Accessed May 29, 2020.
3. Capella J, Smith S, Philp A, et al. Teamwork training improves the clinical care of trauma patients. *J Surg Educ* 2010;67(6):439–443.
4. Thomas L, Galla C. Building a culture of safety through team training and engagement. *BMJ Qual Saf* 2013;22(5):425–434.
5. Weaver SJ, Dy SM, Rosen MA. Team-training in healthcare: a narrative synthesis of the literature. *BMJ Qual Saf* 2014;23(5):359–372.
6. Hughes AM, Gregory ME, Joseph DL, et al. Saving lives: a meta-analysis of team training in healthcare. *J Appl Psychol* 2016;101(9):1266–1304.
7. Neily J, Mills PD, Young-Xu Y, et al. Association between implementation of a medical team training program and surgical mortality. *JAMA* 2010;304(15):1693–1700.
8. Baker DP, Salas E, King H, Battles J, Barach P. The role of teamwork in the professional education of physicians: current status and assessment recommendations. *Jt Comm J Qual Patient Saf* 2005;31(4):185–202.
9. Marshall SD, Flanagan B. Simulation-based education for building clinical teams. *J Emerg Trauma Shock* 2010;3(4):360–368.
10. Agency for Healthcare Research and Quality. TeamSTEPPS. TeamSTEPPS. Available at: https://www.ahrq.gov/teamstepps/index.html. Accessed May 29, 2020.
11. Agency for Healthcare Research and Quality. TeamSTEPPS 2.0 Self-Paced Course. TeamSTEPPS 2.0 Self-Paced Course. Available at: https://www.ahrq.gov/teamstepps/instructor/onlinecourse.html. Published March 2019. Accessed May 29, 2020.
12. Wang R, DeMaria SJr., Goldberg A, Katz D. A systematic review of serious games in training health care professionals. *Simul Healthc* 2016;11(1):41–51.
13. O'Neil HF, Baker EL, Perez RS. *Using Games and Simulations for Teaching and Assessment: Key Issues*. 1st ed. Routledge, New York; 2016.
14. Iseli MR, Jha R. Computational Issues in Modeling User Behavior in Serious Games. In: *Using Games and Simulations for Teaching and Assessment: Key Issue*. 1st ed. Routledge, New York; 2016:21–40.
15. Issenberg SB, Mcgaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach* 2005;27(1):10–28.
16. Salas E, Reyes DL, McDaniel SH. The science of teamwork: progress, reflections, and the road ahead. *Am Psychol* 2018;73(4):593–600.
17. Mitchell P, Wynia M, Golden R, et al. *Core Principles & Values of Effective Team-Based Health Care*. Discussion Paper, Institute of Medicine, Washington DC; 2012.
18. Olszewski AE, Wolbrink TA. Serious gaming in medical education: a proposed structured framework for game development. *Simul Healthc* 2017;12(4):240–253.
19. Drummond D, Hadchouel A, Tesnière A. Serious games for health: three steps forwards. *Adv Simul (Lond)* 2017;2:3.
20. Avila-Pesantez D, Rivera L. Approaches for serious game design: a systematic literature review. *Comput Educ* 2017;8(3):1–11.
21. Catalano CE, Luccini AM, Mortara M. Guidelines for an effective design of serious games. *Int J Serious Games* 2014;1(1):1–13.
22. Shavelson RJ, Webb NM. *Generalizability Theory: A Primer*. Vol 1. Sage Publications, Newbury Park; 1991.
23. Koenig A, Iseli M, Wainess R, Lee JJ. Assessment methodology for computer-based instructional simulations. *Mil Med* 2013;178(10S):47–54.
24. Almond RG, Mislevy RJ, Steinberg LS, Yan D, Williamson DM. *Bayesian Networks in Educational Assessment*. Springer, New York; 2015.
25. Mislevy RJ, Haertel G, Riconscente M, Rutstein DW, Ziker C. Evidence-centered assessment design. In: *Assessing Model-Based Reasoning Using Evidence- Centered Design*. SpringerBriefs in Statistics. Springer International Publishing Cham, Switzerland; 2017:19–24.
26. Cai B, Huang L, Xie M. Bayesian networks in fault diagnosis. *IEEE Trans Ind Inf* 2017;13(5):2227–2240.
27. Culbertson MJ. Bayesian networks in educational assessment: the state of the field. *Appl Psychol Meas* 2016;40(1):3–21.
28. Agency for Healthcare Research and Quality. TeamSTEPPS Learning Benchmarks. Available at: http://www.ahrq.gov/teamstepps/longtermcare/sitetools/learnbench.html. Accessed May 29, 2020.
29. Tabachnick BG, Fidell LS. *Using multivariate statistics*. 7th ed. Pearson, New York; 2019:203–256.
30. O'Regan S, Molloy E, Watterson L, Nestel D. Observer roles that optimise learning in healthcare simulation education: a systematic review. *Adv Simul (Lond)* 2016;1:4.
31. Delisle M, Ward MAR, Pradarelli JC, Panda N, Howard JD, Hannenberg AA. Comparing the learning effectiveness of healthcare simulation in the observer versus active role: systematic review and meta-analysis. *Simul Healthc* 2019;14(5):318–332.
32. Kirschner PA, Sweller J, Kirschner F, Zambrano RJ. From cognitive load theory to collaborative cognitive load theory. *Int J Comput Support Collab Learn* 2018;13(2):213–233.
33. Bracq MS, Michinov E, Jannin P. Virtual reality simulation in nontechnical skills training for healthcare professionals: a systematic review. *Simul Healthc* 2019;14(3):188–194.
34. Youngblood P, Harter PM, Srivastava S, Moffett S, Heinrichs WL, Dev P. Design, development, and evaluation of an online virtual emergency department for training trauma teams. *Simul Healthc* 2008;3(3):146–153.
35. Haerling KA. Cost-utility analysis of virtual and mannequin-based simulation. *Simul Healthc* 2018;13(1):33–40.