Review of: "How Effective are Tabletop Role-Playing (Serious) Games in Understanding and Validating the Predictive Capabilities of Disaster Response Agent-based Models?"

Elaine L. Sigalet

Potential competing interests: No potential competing interests to declare.

Many thanks for the opportunity to review this important topic: the effectiveness of tabletop role-playing games in validating agent-based models. This is an important area for research, and I commend the authors for conducting this research. I hope my comments will broaden the authors' abilities to continue this important line of research, with more attention to the organization of the paper, analysis, and presentation/clarity of terms used to describe concepts and data.

The purpose of the research, according to the authors, was threefold: to train and prepare responders in emergency preparedness and response, and to gain insight into the use of tabletop role-playing games (RPGs) for validating current agent-based models (AGMs) like NetLogo; by observing interactions between players in the game and comparing them to agents in AGMs.

Authors present a good overview of the literature but miss a few important papers that focus on different NetLogo AGM models that present important information, such as: "Improving human behaviour in macroscale city evacuation agent-based simulation," Beth Barnes * , Sarah Dunn, Christopher Pearson, Sean Wilkinson, School of Engineering, Newcastle University, Newcastle, NE1 7RU, UK, International Journal of Disaster Risk Reduction, Volume 60, 15 June 2021, 102289, Sena A, Forde F, Yu C, Sule H, Masters MM. "Disaster Preparedness Training for Emergency Medicine Residents Using a Tabletop Exercise." *MedEdPORTAL*. 2021;17:11119. https://doi.org/10.15766/mep_2374-8265.11119; and Araz, Ozgur M., and Megan Jehn. "Improving public health emergency preparedness through enhanced decision-making environments: A simulation and survey-based evaluation." *Technological Forecasting and Social Change* 80.9 (2013): 1775-1781.

This is an ambitious and important research topic. My suggestions moving forward include:

1. Define and provide information about NetLogo, as it is the AGM the study is trying to validate, yet there is very little information about it.
2. Provide more information about the relationship between comparing TTRPGs to AGMs, NetLogo, the oil and gas disaster, and the ambulance policy in the introduction and abstract. "Example: This study used two variations of an ambulance policy dictating procedure post large oil and gas medical disaster to compare the effectiveness of using table-top role-playing games based on the same scenario for informing, refining, and validating an agent-based model developed for this simulated scenario; NetLogo.

3. A larger sample population should resonate with a real-world population. Not sure master students fit with this. It would be helpful to have more clarity about the participants and groups. In the abstract, results are based on 8 participants - but in materials and design, it states 10 participants and two observers in the preliminary TTRPG and goes on to say doctors were compared to students. In results, it says there were 12 participants with two observers and only one physician. There is a statement stating that the students were compared to the doctors- again, I found this confusing. I would suggest describing group A as those with some medical background. Furthermore, were the master students pre-med or in a faculty related to emergency response? It is always important when comparing populations to compare real populations, those actually in these roles in the real world, to get an accurate accounting of knowledge, attitudes, and behaviors.
4. It would be helpful to have more information about each role in a disaster response at the operational center level, including the game master. Was the game master someone who would normally work as the emergency center operations director or commander in charge?
5. Better organization of concepts: Some background was provided in the introduction, the literature review, and further on in the materials and design section.
6. Suggest starting with introduction, where you introduce the purpose of the study and why it is important. Then background- you do not need a literature review but should use that review to let the reader know what we know today about TTRPGs and how they can be used to validate AGMs in scenarios like major disasters- more information on NetLogo would be helpful. Research design: Method: Quasi-experimental research design used to look at validity of AGM through a comparison of an oil and gas disaster scenario and player knowledge and behavior- Who was your target audience? Sample: Population- sample size? Educational intervention: Preparation for TTRPG. Data analysis: Excel…. Data: Outcome variables: Questionnaires (sample) previously validated or undergo reliability and validity testing, time to task, performance; definition of concepts- scoring rubric- again, reliability and validity? Results, Discussion, Limitations, Conclusion.
7. Research Design: Quasi-experimental- as there is no randomization of population to groups
8. Study outcomes include pre- and post-assessment on knowledge using a MCQ questionnaire, but I was unable to actually see that data in the paper. There are many statistical programs that can be used to analyze this type of data. T-tests with programs like SPSS or a similar test using Excel should be used to look at differences within groups, whereas independent t-tests should be used to look at differences between groups. Results should be described by presenting the mean/SD, confidence interval, and t statistic. With a small sample size, it is important to use a non-parametric test to confirm findings, as it is unlikely that data are normally distributed, which is a requirement for using parametric testing. Suggestions, again with SPSS, would include the Wilcoxon rank test for differences in groups and the Mann-Whitney U test for differences between groups.
9. Again, helpful to include a small table of findings for all analyses where it is discussed in the manuscript. Was reliability and/or validity tested with the questionnaires? What about observer training and interrater agreement- again, these concepts would strengthen findings. Group A had a higher knowledge score post-RPG compared to group B- that could be related to their background- so significance is unclear. Both groups did improve their scores, so that does suggest the TTRPG was effective for building the level of knowledge. Interestingly, the group of students only

increased it 3.1 %- what was their baseline- much lower than the medical group?

10. Providing clear descriptions of character variables is important to reader understanding: example, time on task described by 2<t<4- I was confused by these variables.
11. The analysis of performance between in-game and user feedback, which I assumed was from the debrief, is challenging to compare- the concepts evaluated were different but overlap. It would be helpful to see the same concepts compared in both game interaction and participant feedback in the debrief. Teamwork is a broad concept that usually includes communication, collaboration, and coordination (from the military definitions) and in medicine: leadership, roles, situation awareness, communication, and resource utilization.
12. It would be interesting to know if the authors suggested any changes to the NetLogo AGM based on preliminary testing using their disaster scenario in the TTRPG. Could results be used to develop a new AGM for this particular scenario? The study brings forth some interesting questions.
13. Comparing saving lives and deaths between the TTRPG and AGM reinforces the importance of this type of research so that we can eventually have AGMs that are valid for predicting future outcomes from serious events like disasters, but as the authors suggest, further research is needed to ensure AGMs are valid. I would disagree with the authors' claim of partial validity; the comparison is either valid or not unless authors want to describe the different types of validity: content, test-retest, predictive, etc…. Perhaps stating that the overlap in confidence intervals between the AGM and TTRPG suggests researchers are on the right track to developing predictive AGMs would be a better way to phrase findings.

Again, the authors put considerable effort into this research study, and I hope my comments will motivate them to continue, with more attention to areas in my comments. I look forward to seeing further publications from these authors.