Title

Outcomes assessment pitfalls: challenges to quantifying knowledge gain in a sex education game.

Authors

Bertozzi, Elena
Bertozzi-Villa, Amelia
Padankatti, Swathi
et al.

Publication Date

2020

DOI

10.12688/gatesopenres.13129.3

Peer reviewed

RESEARCH ARTICLE

REVISED Outcomes assessment pitfalls: challenges to quantifying knowledge gain in a sex education game [version 3; peer review: 2 approved, 1 not approved]

Elena Bertozzi[1], Amelia Bertozzi-Villa[2], Swathi Padankatti[3], Aparna Sridhar[4]

[1]Game Design & Development, Quinnipiac University, Hamden, Connecticut, 06518, USA
[2]Institute for Disease Modeling, Bellevue, Washington, 98005, USA
[3]Department of Pediatrics and Neonatology, Sundaram Medical Foundation, Chennai, India
[4]Department of Obstetrics and Gynecology, David Geffen School of Medicine, Los Angeles, CA, 90095, USA

v3 **First published:** 30 Jun 2020, **4**:73
https://doi.org/10.12688/gatesopenres.13129.1
Second version: 13 Nov 2020, **4**:73
https://doi.org/10.12688/gatesopenres.13129.2
Latest published: 16 Feb 2021, **4**:73
https://doi.org/10.12688/gatesopenres.13129.3

Abstract

Background: The use of videogames as a public health tool is rapidly expanding. Accurate assessment of the efficacy of such games is complicated by many factors. We describe challenges associated with measuring the impact of playing a videogame with information about human sexual anatomy and reproduction and discuss motivations for, and solutions to, these challenges.
Methods: The My Future Family Game (MFF) is a validated tool for collecting data about family planning intentions which includes information about human anatomy and sexual reproduction. We sought to assess the efficacy of the game as a tool for teaching sexual education using a pre-post model which was deployed in three schools in and around Chennai, India in summer of 2018.
Results: The MFF game was successfully modified to collect data about players' pre-gameplay knowledge of sexual anatomy and processes. The post gameplay assessment process we used did not effectively assess knowledge gain. Designing assessments for games dealing with sexuality presents challenges including: effectively communicating about biological parts and processes, designing usable and intuitive interfaces with minimal text, ensuring that all parts of the process are fun, and integrating assessments into the game in a way that makes them invisible.
Conclusion: Games can be an effective means of gathering data about knowledge of sex and reproduction that it is difficult to obtain through other means. Assessing knowledge about human sexual reproduction is complicated by cultural norms and taboos, and technical hurdles which can be addressed through careful design. This study adds to the sparse literature in the field by providing

Open Peer Review

Reviewer Status ✓ ✗ ✓

	Invited Reviewers		
	1	**2**	**3**
version 3 (revision) 16 Feb 2021			✓ report
version 2 (revision) 13 Nov 2020	✓ report		? report
version 1 30 Jun 2020	✗ report	✗ report	

1. **Melissa Gilliam**, University of Chicago, Chicago, USA
2. **Claudia Regina Furquim de Andrade**, University of Sao Paulo, Sao Paulo, Brazil
3. **Fabio Buttussi**, University of Udine, Udine, Italy

Any reports and responses or comments on the article can be found at the end of the article.

information about pitfalls to avoid and best practices in this evolving area.

Keywords
games for health, serious games, sexual education, outcomes assessment, family planning, India

Corresponding author: Elena Bertozzi (elena.bertozzi@quinnipiac.edu)

Author roles: Bertozzi E: Conceptualization, Funding Acquisition, Investigation, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Bertozzi-Villa A**: Data Curation, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Padankatti S**: Investigation; **Sridhar A**: Funding Acquisition, Investigation, Methodology, Project Administration, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work is supported by the Bill and Melinda Gates Foundation through its Grand Challenges in Global Health program which funded the development of the My Future Family Game [OPP1161938].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Bertozzi E, Bertozzi-Villa A, Padankatti S and Sridhar A. **Outcomes assessment pitfalls: challenges to quantifying knowledge gain in a sex education game [version 3; peer review: 2 approved, 1 not approved]** Gates Open Research 2021, **4**:73 https://doi.org/10.12688/gatesopenres.13129.3

First published: 30 Jun 2020, **4**:73 https://doi.org/10.12688/gatesopenres.13129.1

REVISED **Amendments from Version 2**

Many thanks to this reviewer for his bibliographic recommendations which provided us with useful examples of other work in the field. These articles allowed us to expand the discussion section to explore how others have addressed lack of participation in post-test assessments and better use of in-game metrics to assess knowledge gain for participants who don't complete the post-test.

The explanation of how the game works was expanded to include more details and there is now a link to video of gameplay.

Throughout the paper, we standardized the use of term "post-test" and consistently use "serious" rather than "applied" game. The scale of scores was also standardized to facilitate clarity and ease of comparison.

Any further responses from the reviewers can be found at the end of the article

Introduction

The acceptance of games as useful and effective tools for collecting data, educating players, and achieving positive behavior change is growing due to an increase in rigor in the deployment and assessment of serious games (Boyle *et al.*, 2016; Coovert *et al.*, 2017; Zammitto, 2009). Embedding outcomes assessment within the game itself is often described as an important design principle in building games, largely due to the fact that most games incorporate some form of player feedback and metrics as part of gameplay (Ifenthaler *et al.*, 2012; Van Staalduinen & de Freitas, 2011). There are situations, however, in which such assessment is quite difficult.

The *My Future Family Game* (MFF) game was initially developed as a tool for collecting information about family planning intentions among adolescents in Mysore, India in 2017. The original goal was to gather information about desired family size and spacing, influencers of the decision-making process, and other data points[i]. Focus group participant feedback during early-stage planning was crucial to the success of the project. Serious games are only effective if in addition to achieving their stated goals, they are also intrinsically motivating (fun) for players (Malone, 1981). Analysis of focus group feedback showed that although sex education is included in the standard curriculum for adolescents, many young people do not have basic knowledge about human reproduction (Bertozzi *et al.*, 2018). Including this information in the game would strongly motivate adolescents to play, and was supported by parents and educators as a way of communicating sensitive information.

The first beta of the game (MFF_2017) was successfully tested on 480 adolescents in summer of 2017 and proved to be a very effective tool for gathering information from a population about which little accurate information is available from other sources (Ismail *et al.*, 2015; Weiser, 2015).

[i]The project was funded by a Grand Challenges in Global Health grant (https://gcgh.grandchallenges.org/grant/childbearing-intentions-and-family-planning-game).

Post-game paper questionnaires and interviews demonstrated that the game was considered fun and well-accepted by student players. Analysis of the MFF_2017 deployment suggested that the game could function not only as a method of collecting data about family planning intentions, but also as a means of communicating information about human sexual anatomy and reproduction in an innovative way (Bertozzi *et al.*, 2018). For the second deployment of the game (MFF_2018) on a different population we sought to validate the utility of the game as a means of measuring pre-intervention knowledge of game content and quantifying knowledge gain after having played the game. Our goal was also to develop an assessment that did not require post-intervention interviews or paper questionnaires to facilitate large-scale deployment of the game as an educational tool in low-resource settings.

A literature search was conducted to assess best practices for outcomes assessment of videogame efficacy particularly relating to games for health and knowledge gain. There is very sparse literature in the area and several of the papers discussed this lack and the need for more work in the field (Baranowski *et al.*, 2019; DeSmet *et al.*, 2015). The search terms 'videogame' AND 'knowledge' were used in the PubMed database resulting in 38 results. Titles and abstracts were screened for studies that measured outcomes of playing videogames about health resulting in 16 papers that discussed methodologies. Of these, papers that reported assessment of knowledge gain as a result of videogame play dealt with topics that either were integrated with existing assessments (such as diabetes knowledge measures (Joubert *et al.* 2016)) and/or were not controversial and easily assessed with pre-post questionnaires (Del Blanco *et al.*, 2017; Hieftje *et al.*, 2019).

A search using 'videogame' AND 'sex yielded 51 non duplicative results of which only two were assessments of a videogame about sex from the same research group (Fiellin *et al.*, 2017; Gariepy *et al.*, 2018). Neither of these attempted a pre-post model for knowledge gain. The primary outcome measure was behavioral (initiation of sexual activity). Players of the sex ed game demonstrated increased sexual knowledge (assessed using interviews) compared to a control group that played generic games.

The literature provided us with little guidance in determining how to assess knowledge gain in a videogame about sexual education that did not include in-person interviews, beyond the importance of embedding the assessment as much as possible within the game itself.

Methods

Inclusion Criteria and Deployment Structure

MFF_2018 was deployed in Chennai, India as part of research conducted by Dr. Swathi Padankatti and her team from the International Alliance for the prevention of AIDS in collaboration with the U.S. based game development team (Dr. Bertozzi's group at Quinnipiac University) and Dr. Aparna Sridhar at U.C.L.A's School of Medicine. Dr. Padankatti and her team identified three schools willing to participate in the study who could provide a total of 419 student players (208 males and 211 female) ranging from 13 to 16 years of age. This was a

convenience sample due to the sensitive nature of the game content. Schools were selected based on the research team's pre-existing relationships with administrators, with whom they had previously worked on AIDS education initiatives and were open to curriculum initiatives relating to sexual knowledge. The fact that these educators and students had previously participated in HIV education initiatives made it likely that the sample audience would have more knowledge about sex and reproduction than the population at large.

For game deployment in each participating school, 30 android tablets and headsets were set up in a school classroom, and groups of students in the target age group successively cycled through to play the game and discuss their experience. Participants were invited to play the game if they desired to do so and could stop at any time. No students were excluded from participation. Groups were not segregated by sex. To ensure comfort and privacy, students were able to move freely around the room to find their preferred space to sit and play.

In the MFF_2017 deployment post-game questionnaires were used for assessment. These were paper forms filled out by students after playing the game, asking students to qualitatively self-assess knowledge gain and provide feedback on the process of gameplay. While these questionnaires provided valuable feedback and indicated high rates of self-assessed knowledge gain, they were not efficient data-collection strategies. Because forms were filled out on paper, response rates were low and it was not possible to link student feedback to specific test-takers. Positive self-assessment of knowledge gain was encouraging, but not a rigorous method for determining game efficacy. The absence of an evaluative framework for the game was the primary motivation for development of the pre-post testing process in MFF_2018.

Once the participants began playing, they were first asked to indicate their sex and age, after which the pre-test was triggered prior to initiating the main game. The post-test, with exactly the same structure and questions as the pre-test, appeared after completing the game.

Given that MFF_2018 was the first field deployment of the pre-post assessment, the deployment team reported issues to the U.S. development team after each play session. The issues were collected and organized into topics. Below we describe the challenges encountered in both the MFF_2017 and MFF_2018 tests and how they were addressed or are planned to be addressed in future revisions of the game.

MFF (original and modified versions of the game are available here: https://osf.io/gtfu5/wiki/home/ (Bertozzi-Villa *et al.*, 2020). The apks can be installed on any Android tablet or phone.). A video of gameplay can be viewed here: https://familyplanninggame.qu.edu/ which includes all the minigames.

Pretest Challenge One: Communication

Discussion of human anatomy and behavior regarding sex and reproduction is problematic in India (Ismail *et al.*, 2015). Many adolescents receive very little information from their parents or teachers due to cultural taboos (Khubchandani *et al.*, 2014). In designing the MFF_2017 game, we were very careful to introduce explicit material slowly and through a process in which it was revealed in context. The game was constructed so that at each point where explicit material is available for the player, the player was asked whether or not they wanted to see it, and then if they agreed, the material was presented in a context that made sense based on the information being gathered.

For example, when players were asked information about when they planned to start dating a possible partner, they were provided with information about the anatomy of the opposite sex[ii]. When they were asked about the age they planned to marry, after consenting (Figure 1), they were given information about how intercourse works via the animation in Figure 2.

The addition of a pre-test that collects knowledge of human sexual anatomy and reproduction before the game begins

[ii]Due to cultural taboos in India which would have made it impossible to deploy the game at all, same-sex marriage was not an option in the game.

Figure 1. Consent screen from Getting Married Milestone.

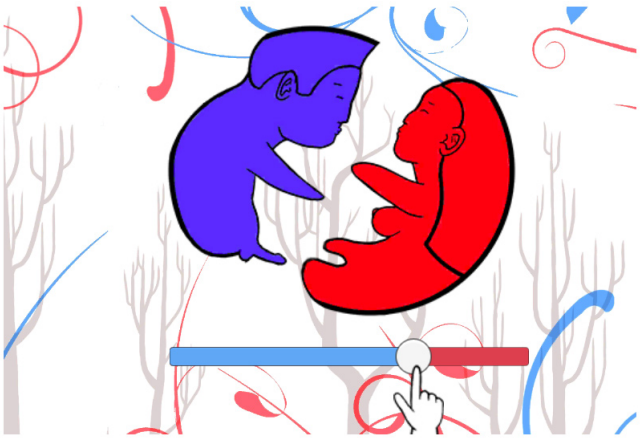

Figure 2. Still from animation that demonstrates sexual intercourse.

required exposing participants to images and terms before we were able to prepare them the way we had with gameplay without assessment. To be as accessible as possible to players at any reading comprehension level, the game includes as little text as possible and communicates most information through graphics, audio and animation. This is especially important when discussing information about sexuality because these terms may not be familiar to students. However, the inclusion of the pre-test introduced a great deal of technical vocabulary in English before the gameplay began. All of the schools included in the MFF_2018 study had instruction in English, but it was unclear if terms like testicles, ovaries, urine and feces were well-understood by players (Figure 3). Although education about sexual functions is technically part of the educational curriculum for all students in India, the content is not actually taught in many schools due to cultural reluctance to discuss sexuality.

Solution: We sought to address the challenge of not startling participants by designing an interface that introduces the sensitive content in the pre-game assessment carefully. Images are largely outlines with enough detail to communicate but not enough to be offensive. Biologic terms relating to specific body parts are shown in limited number on the screen where they belong so even if the player does not know exactly what the term means they can start to associate the correct term with the appropriate body parts. Players are taught the correct terms and their associated body parts when they play the game.

Pretest Challenge Two: Interface Design

A usability issue encountered during the MFF_2017 was that players lacked familiarity with the drag and drop interface commonly used on smartphones and tablets. The design team determined that the 2018 pre-test was the perfect opportunity to teach players how to use drag and drop so that they would be prepared for it when they reached the game. We also determined that we could use the interface design to reinforce the learning of anatomy by clarifying the correlation between the body parts pictured on the screen and their own bodies as explained below.

The key educational content of each milestone of the game is outlined in Table 1.

To assess knowledge of these topics while training students on a drag-and-drop interface, the pre-post test was designed to show male and female figures in outline, with internal organs visible. A series of 14 anatomy questions covering the full scope of in-game content was presented in a sequence of views. Players answered questions in the pre-test by dragging a word representing a concept (usually with an animation to help explain it) to the correct location on an image (Figure 4). The structure and content of pre and post-tests is identical.

The assessment was designed to correlate with the way information is delivered in the different milestones in the game. In the MFF_2017 deployment we noted that players had a difficult time understanding where different organs were located in the body and what their functions were. In the pre-post additions, we were careful to depict both male and female bodies as a whole at the start of the pre-test. The view then zooms in to just the abdomens of the male and female bodies. We added the whole person views in the top right and left of the screen so that players could understand which view of the body was presented to them. It is very difficult to understand how organs are laid out relative to other organs. For example, in the female

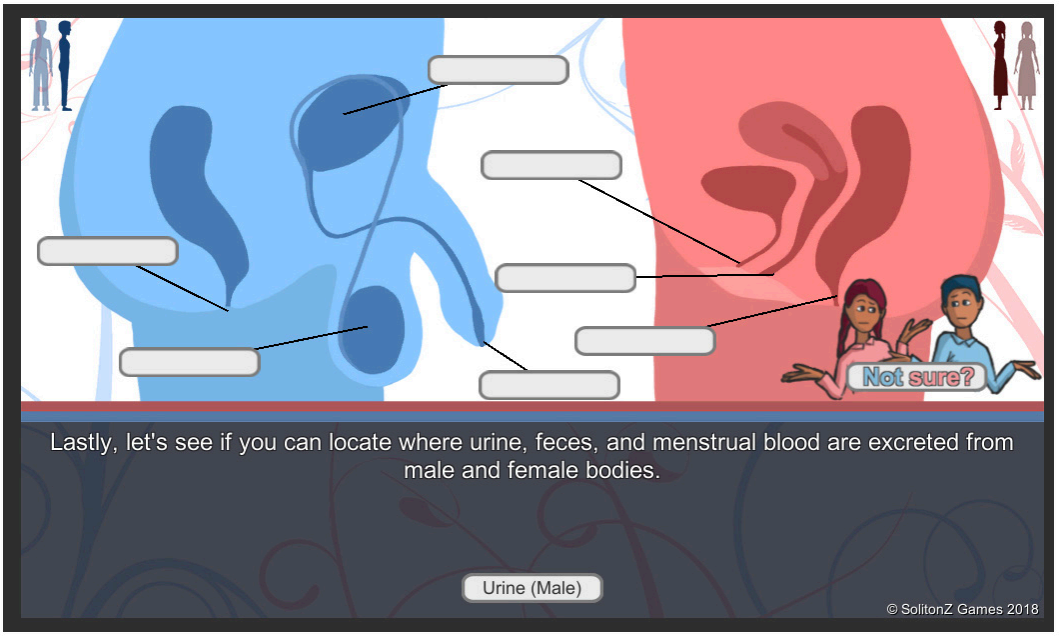

Figure 3. Final anatomy screen from pre-test.

Table 1. Key educational content for each milestone in the MFF game.

Milestone	Content
1: Puberty and bodily functions (same sex as player)	Hair growth, menstruation or ejaculation
2: Reproductive anatomy	Identification of the internal reproductive organs of male and female bodies and their functions
3: Puberty and bodily functions (opposite sex as player)	Hair growth, menstruation or ejaculation
4. Anatomy of intercourse	Act of heterosexual coitus via union of penis and vagina
5. Fertilization	Movements of eggs and sperm, fertilization of egg by sperm.

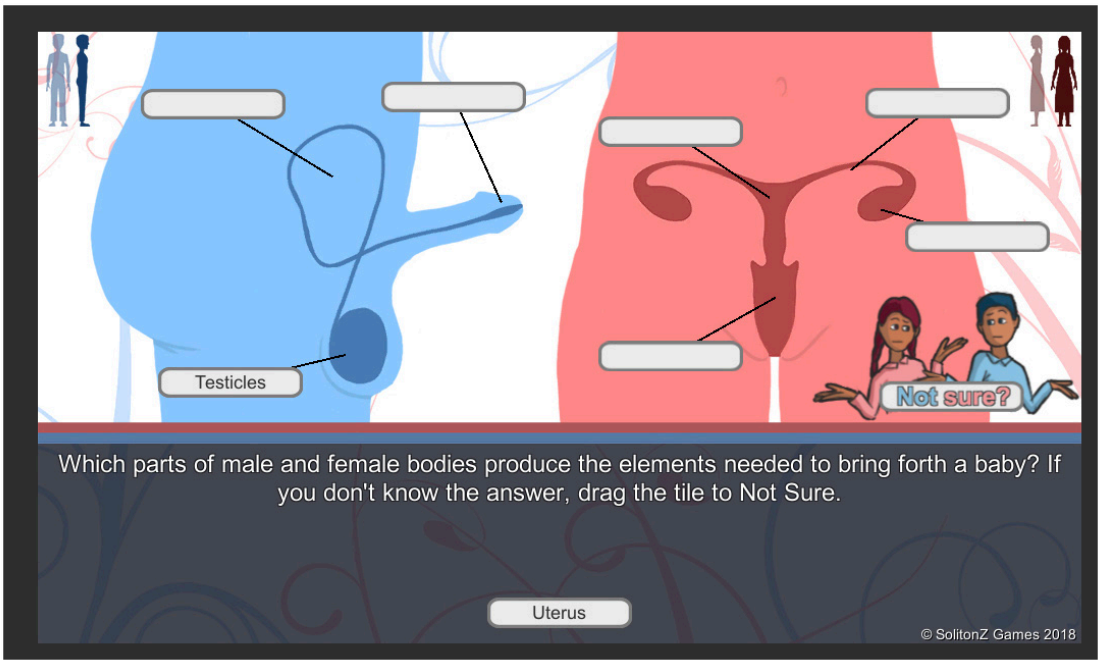

Figure 4. Example of pre-test anatomy question.

body, it can be difficult to show the positions of the three apertures of the urethra, vagina, and rectum relative to one another. The additional views were added to minimize this confusion.

Our hope was that the layout of the pre-test prior to play would prepare players to approach the anatomy section of the game where they have to drag and drop each body part to its correct location (Figure 5). During the MFF_2017 deployment of the game, it was clear that some players did not understand the difference between the front and side views of the anatomical drawings. In addition to adding the side views in the upper right and left corners of the pre-test, we also incorporated them into the minigame. These views update as each organ is dragged into the correct location in the front view.

Solution: The drag and drop training in the pre-test was successful. Players learned how to use it in the pre-test and used it in the game without the difficulties encountered in the previous deployment.

Pretest Challenge Three : Ensuring Fun and Making Assessment Invisible

During the early stages of the MFF_2017 deployment, teachers stayed in the room during gameplay. They often gave stern instructions on how to behave and ordered students to follow the instructions of the researchers. We realized that this made it impossible for students to experience playing the game as play. Due to the presence of their teachers, students experienced the session like a test that they were required to engage in. To encourage a sense of play, the research protocol was modified early in the first deployment to ask instructors to leave the room during gameplay. Additionally, language was added to the introductory scripts, encouraging students to play the MFF game as a game – they should only do the parts of it that they wanted to, and could stop playing at any time. This protocol was

Figure 5. Screenshot of anatomy minigame.

extended into the MFF_2018 deployment where considerable care was taken to encourage a sense of play and remove the pressure associated with the school environment. By introducing a pre-test, however, we recreated the circumstances under which the experience of play was potentially undermined. Students were invited into the room to play a game. However, after they are welcomed to the game, they are presented with an assessment. The 2018 deployment team reported that some students were concerned that they did not have the "right" answer and wanted to be able to go back and correct their previous answers during the pre-test. Given that the Indian system of education heavily relies on test scores and impactfully rewards those who test well, these students appeared very motivated to "do well" as soon as they realized it was an assessment.

Solution: To counter this, the researchers repeatedly stated that they should just answer what they knew and then go on to the game, but this clearly affected the experience. We learned that in future deployments, we need to add more context and less pressure to the pre-test to ensure players understand that they will not be criticized or penalized for not knowing the answers. We can actually leverage their desire to do well by encouraging them to see if they can figure out what they didn't know in the pre-test during subsequent gameplay.

Post-test Challenge: Adherence

We encountered more serious issues with the post-test. While the transferal of the post-test questionnaire into a digital framework did allow for personalized tracking of results, there were challenges to collecting post-game information. Qualitative feedback from the 2018 deployment team indicated that, when students came to the end of the game, and saw the same screen they had seen earlier for the pre-test, many simply dragged the tiles to "Not Sure" because it was the fastest way to get to the final screen. Others simply put down the tablet which meant that researchers had to exit the player from that game session (with no responses to the post-test questions) to reset the tablet for the next group of students. Due to the fact that we do not know exactly what happened in all the cases where there appear to be random answers to the post-test, we cannot determine how many students actually answered the questions intentionally.

Solution: Because the post-test was not only clearly an assessment, but also not well-integrated into the play experience, many players simply abandoned it. We failed to provide players with a compelling reason to want to engage in the final assessment as part of gameplay, which will be corrected in the next deployment.

Analysis

During gameplay, tablets kept timestamped records of every user input. Data on pre- and post- test responses were saved in .csv format for statistical analysis. These datasets include information on the id number of the tablet used, the school in which the game was deployed, the self-reported sex of the player, and a unique user id for each run-through of the game. The pre-post data contains no other personalized student data.

All analyses were run in R version 3.6.0 (Bertozzi-Villa, 2020). Overall pre- and post- test scores, as well as the percent of students who responded correctly to each question, were

calculated from individual responses. On the post-test, players who responded "not sure" to every question were logged as having a "null" post-test. Score differences between groups were assessed via two-sample t-tests, and pre- to post-test score changes were assessed via one-sample t-tests.

Ethics and consent

The study design was approved by the Institutional Review Board of the Sundaram Medical Foundation, Dr. Rangarajan Memorial Hospital, Chennai, India (IRB # IEC-09/1/2018).

Informed verbal consent was obtained from the principals of the three participating schools following consultation and a gameplay demonstration with each one. Consent was not obtained from student participants. The board deemed oral consent would suffice for the principals, and as the game covers topics which are part of the curriculum, participants' consent was not needed. As noted earlier, once in the room students could play or not play as they wished and could put down the tablet and stop playing at any time.

Results

The goal of this analysis was to test if embedding the game within a pre-post assessment would accurately assess how much players had learned over the course of the game. Results demonstrated that the pre-game assessment effectively captured players' existing knowledge of human sexual anatomy and functions. The assessment tool is very helpful in demonstrating which schools are doing better with their sex ed curriculum and specifically which topics are better understood. The post-test was not effective because many players skipped it as it was not perceived as being an integral part of the experience. Lessons learned from the deployment will guide future revisions.

Assessment of pre-test scores

The pre- and post-test were identical, each consisting of 14 anatomy questions presented via a drag-and-drop interface with figures. Students received a single point for each question they answered correctly, and no points for incorrect responses. In the results below, test scores are always presented as percentages. A total of 419 students in three schools completed the pre-test and main game. The schools were selected based on scheduling availability and willingness to participate. The researchers from the IAPA had previously worked with these schools on AIDS education initiatives. Across all schools, the pre-test score was 33.5% on average (SE 1.15%), with substantial variation between schools. In particular, students at School 2 (who had sexual education as a formal part of their curriculum) performed significantly better than students at Schools 1 or 3 (two-tailed t-test $p<0.001$). Pre-test scores were not significantly different between male and female students at any school (Figure 6).

Across all schools, students scored slightly better on pre-test questions relating to the anatomy of their own sex compared

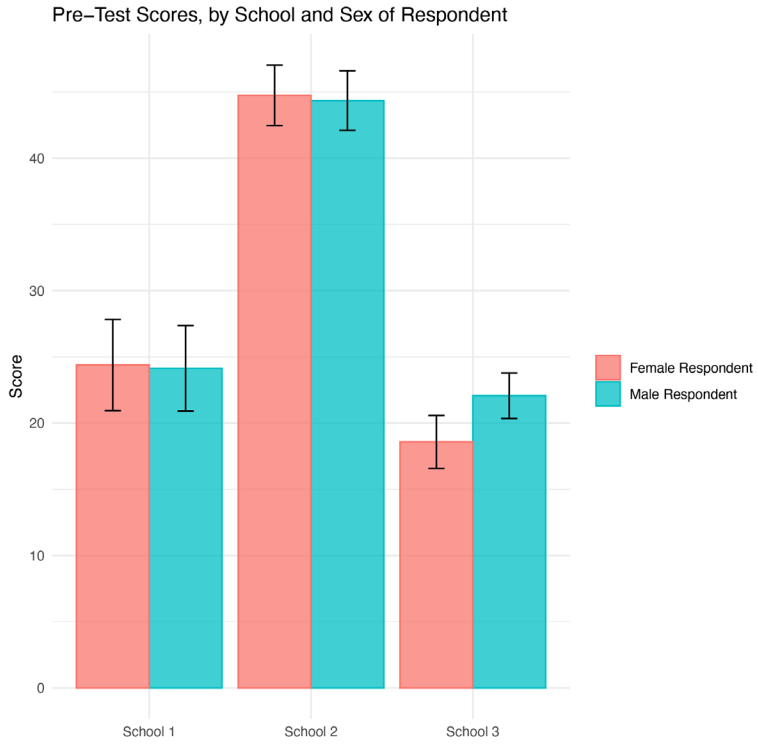

Figure 6. Pre-test scores by school and sex of respondent. Bars represent the standard error of the mean. N=419.

to the opposite sex, but this effect was not statistically significant (Figure 7). For female anatomy questions, 34.4% (SE 1.86%) of female respondents answered correctly, compared to 30.6% (SE 1.6%) of male respondents (two-tailed t-test *p=0.13*). For male anatomy questions, 33.3% (SE 1.89%) of female respondents answered correctly, compared to 36.3% (SE 1.8%) of male respondents (two-tailed t-test *p=0.25*).

As shown in Figure 8, the only questions for which a majority of responses were correct were "Where is urine excreted from a male?" (53.0% correct) and "Where does a lining build up to prepare for pregnancy?" (50.6% correct). For eight of the remaining 12 questions, the correct answer received a plurality of responses, but not a majority. The four questions for which the most frequent response was not the correct answer were "Where sperm exit the body?" (plurality answer "Not Sure", 29.8%), "where menstrual blood is excreted?" (plurality answer "Not Sure", 27.7%), "The organ that becomes erect before intercourse?" (plurality answer "Vagina", 33.7%), and "Where urine is excreted from a female? (plurality answer "Vagina", 35.6%)".

Pre-post test assessment

As described above, assessment of knowledge gain was complicated by the large number of students who did not complete the post-test or who rushed through it, answering "not sure" to all questions (173 students, 41.3% of total). We refer to this group as having a "null" post-test. While it is not possible to assess knowledge gain among those with a null post-test, among the 246 (58.7%) students who did attempt the post-test we find on average a 6.27-percentage-point score gain between pre- and post- tests (95% CI 3.8–8.75, *p<0.001*, one-sample t test, Figure 9). Scores ranged from 0 to 100 for both tests. In the pre-test, 17 students answered all questions incorrectly while one student answered perfectly. In the post-test, five students answered all questions incorrectly and five students answered perfectly.

A question-by-question breakdown of pre- vs post- test result among those who attempted the post-test shows the largest knowledge gain around topics of intercourse, egg storage, and sperm movement (Figure 10).

Discussion

Initial analysis of the results suggested that there was little knowledge gain as a result of gameplay. There was very little change between pre- and post-test results on many of the metrics. Discussions with the deployment teams and more detailed analysis of the results produced a more nuanced understanding of what happened. Many players simply did not complete

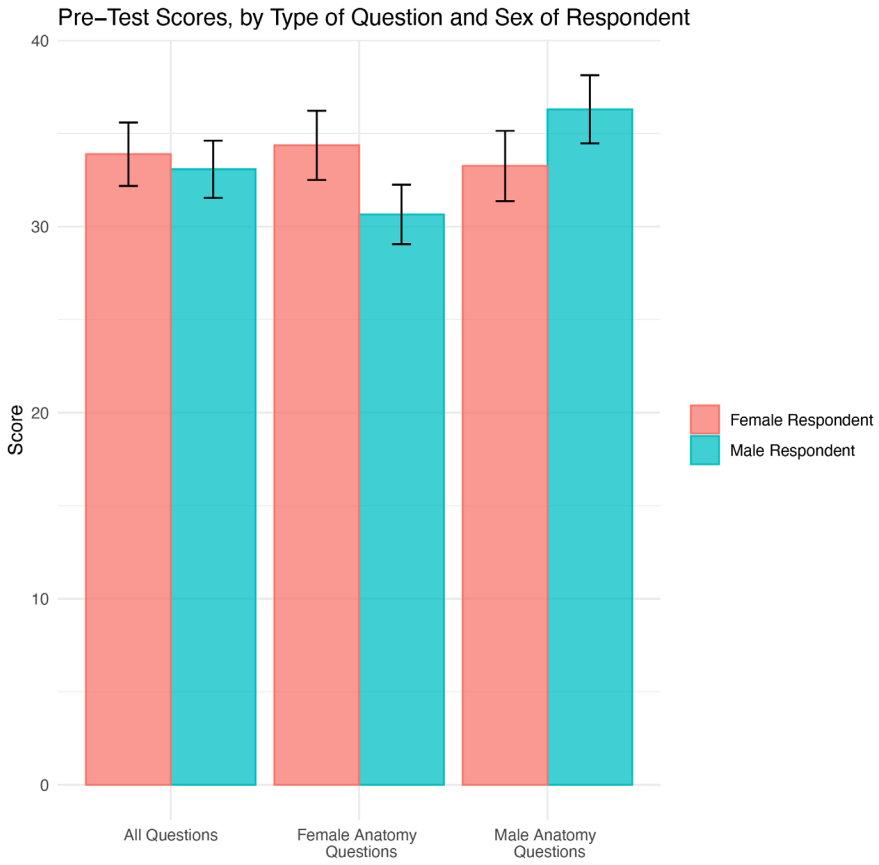

Figure 7. Pre-test scores by type of question and sex of respondent. Bars represent the standard error of the mean. N=419.

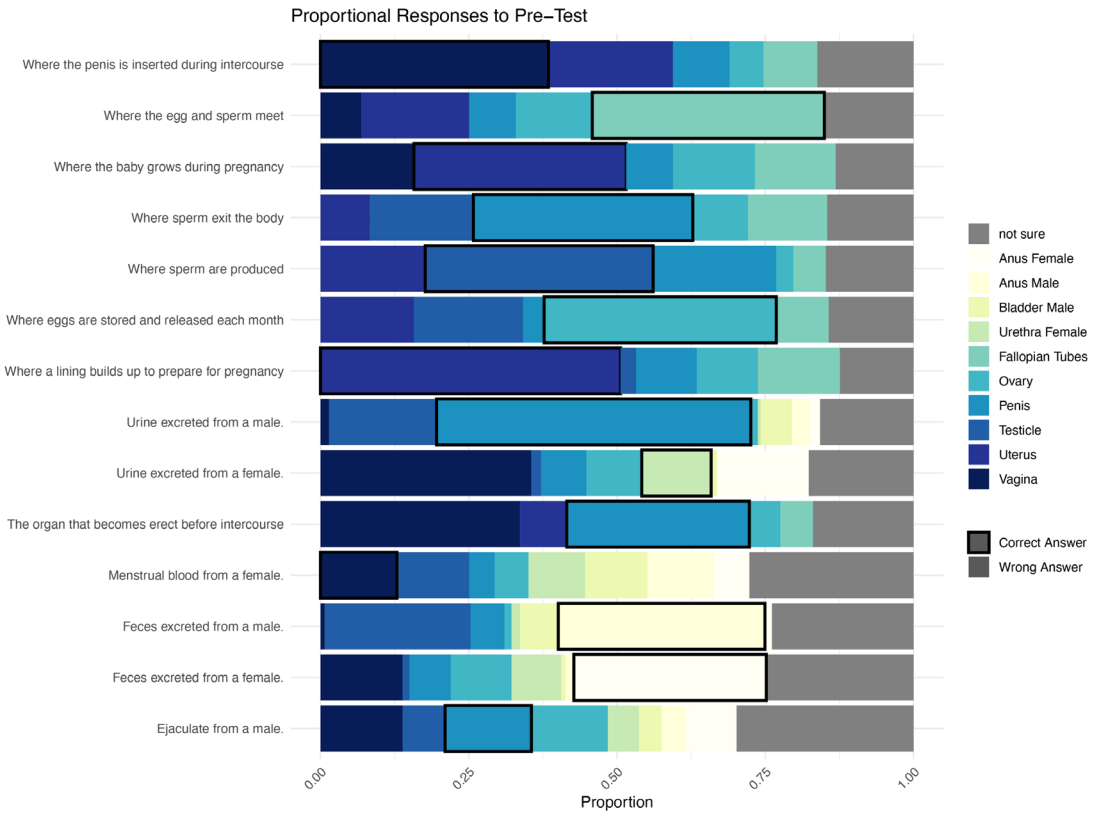

Figure 8. Pre-test responses by question. The highlighted bar shows the correct answer in every instance. N=419.

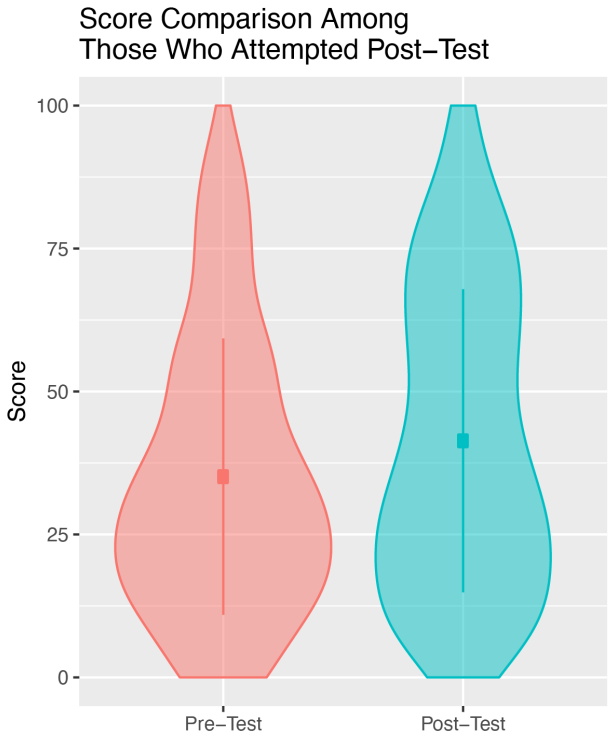

Figure 9. Violin plot of score distributions in the pre- and post-test, for those students who completed both tests (N=246). Points represent means, and bars represent two times the standard error.

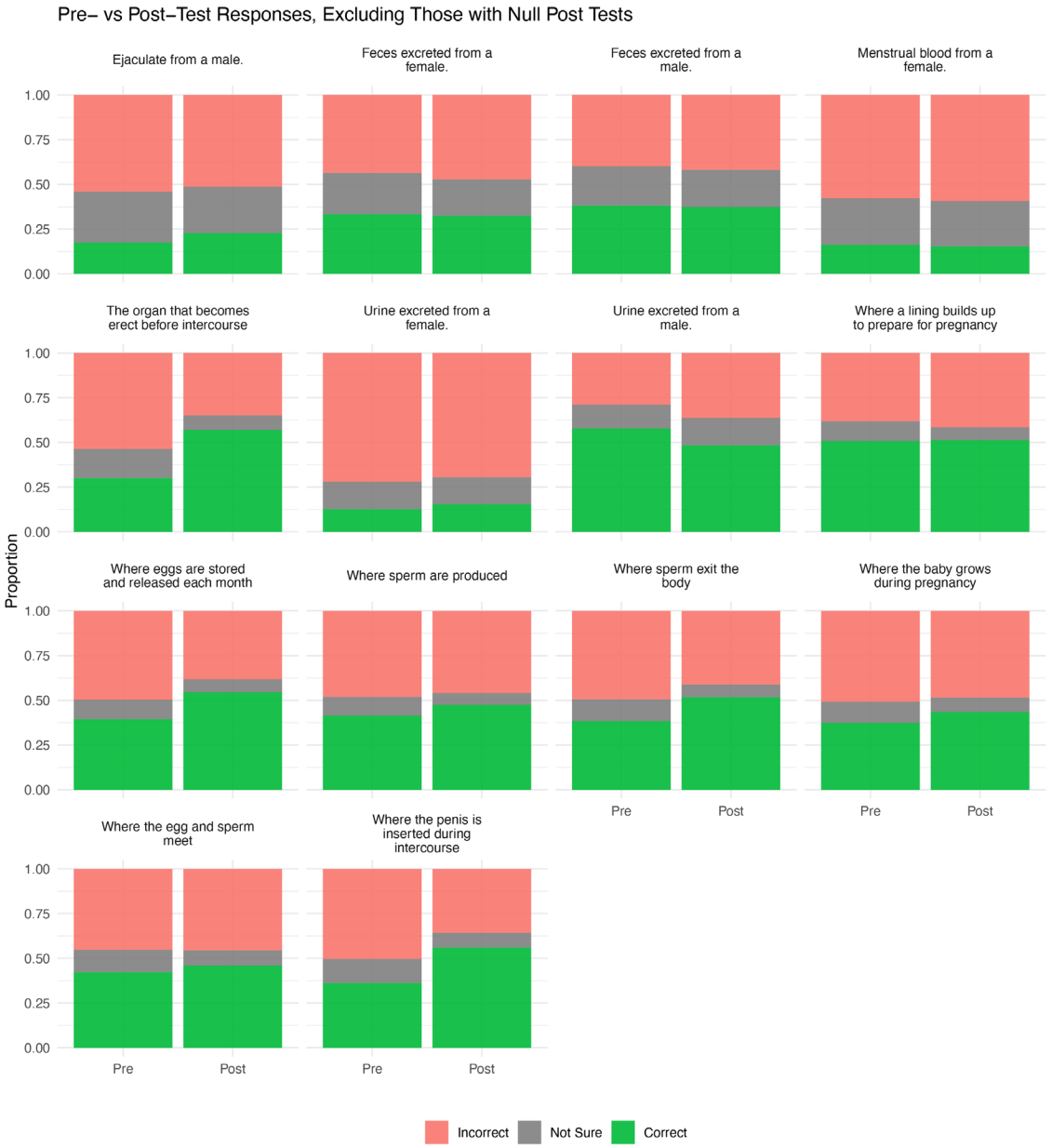

Figure 10. Question-by-question comparison of pre- and post- test responses, by proportion, excluding those with "null" post-tests (N=246).

the post-test or randomly swiped to finish as quickly as possible. When the data for players who did complete both the pre- and post-tests with intention was analyzed separately, there was a modest but notable increase in knowledge. Additionally, we determined that the pre-test was a useful tool for assessing prior knowledge and therefore the efficacy of sexual education programs at different schools.

We learned a great deal about the difficulty of creating effective pre-post assessments for a game that includes sensitive topics. Adolescents offered a game of this type are already nervous and excited about it. The process of setting up a context in which their current knowledge is assessed needs to be approached carefully. We encountered several pitfalls that complicated the assessment process and which affected the validity of the assessment data. We are able to conclude that using a game to assess current knowledge of reproductive anatomy and processes can be very effective. In order to assess knowledge gain after gameplay, students need to be motivated to fully engage in the post-test assessment. For future deployments of the game, we plan to change the deployment protocol to address the issues discussed in this report and

better integrate the pre-post testing process in the overall experience.

Other researchers have addressed the problem of lack of participation in post-testing in creative ways. Chittaro & Buttusi (2020) included a pre-post test in their widely distributed aircraft cabin safety game. In addition to the data collected from players who completed the pre-post, they collected data about the kinds and frequency of errors that players made while playing the game and included that analysis in their reporting on efficacy and knowledge gain. There are two minigames in the MFF game where players are dragging and dropping objects to learn how different reproductive systems work. In future deployments, we can track how often objects are misplaced or dropped and integrate these metrics into our efficacy assessment. Additionally, we can track how often players replayed minigames and how this correlates to performance on pre-post tests. In their meta-analysis of digital learning through games, Clark *et al.* (2016) suggest several ways to enhance leaning and assessment including: allowing players to have unlimited access to the game, embedding the game within existing classroom assessment measures, and including supplemental non-game instruction. None of these options were available to us in the India deployments, but they can guide our attempts to distribute the game in the future.

Other ways to more accurately assess the impact of the game include expanding the scope of assessment to include more qualitative metrics. Other researchers have noted that playing games about sensitive topics can result in players feeling more comfortable about the topic and reducing stigma (Arnab *et al.*, 2013; Ferchaud *et al.*, 2020). The deployment teams during both deployments noted that prior to playing the game, the atmosphere in the room was somewhat tense, but that after everyone had played the game players were relaxed and laughing and appeared more comfortable asking questions about reproduction and sexuality. In future deployments, we will seek to measure this impact and include self-efficacy and affective outcomes as described in several of the studies reported by (Boyle *et al.*, 2016)

It is standard practice in serious game development to seamlessly integrate assessment into the existing structure of the game (Klopfer *et al.*, 2018; Serrano-Laguna *et al.*, 2018). As we have shown, this is difficult in a game that deals with a sensitive topic. Our plan going forward is to address this challenge openly in the introduction to the game experience. After players open the tablet, we will have an animated character appear who discusses the fact that what will follow is a game about sexuality and that this is a difficult topic for many people to talk about. After normalizing the idea of embarrassment, the character will then introduce the idea that knowledge is power and that the game will help players learn about things that are important to their future. Then the pre-post will be presented as a challenge…" let's see how much you know now and then see if after you play the game you know all the answers to things you didn't know before." Hopefully, with this context, we will avoid the pitfalls of our Chennai deployment.

Conclusion

This deployment demonstrated that a game-based tool can be an effective means of gathering information. We learned that many adolescents in these schools lack basic knowledge of human anatomy and sexuality, especially given that the students chosen had already received baseline training in HIV prevention and are likely better informed than other students. We observed that in addition to collecting data and educating players, games of this type also have the potential for de-stigmatizing conversations about sex and sexuality which we will seek to quantify in future deployments. The deployment also provided us with important information for improving the tool.

Data availability

Underlying data

Open Science Framework: Outcomes Assessment Pitfalls: Challenges to Quantifying Knowledge Gain in a Sex Education Game. https://doi.org/10.17605/OSF.IO/WMHCD (Bertozzi-Villa *et al.*, 2020)

This project contains the following underlying data

- prepost.csv (Questions and responses to all pre- and post- tests administered, along with timestamps and other metadata)

Extended data

Pre- and post-test data were analyzed and visualized using R version 3.6.0. All code is available from GitHub (https://github.com/bertozzivill/india-family-planning) and archived with Zenodo (http://doi.org/10.5281/zenodo.3822455 (Bertozzi-Villa, 2020))

Software availability

An installable and playable version of the game and all data used for analysis is publicly available at Open Science Framework, as described below.

Archived source code at time of publication: https://doi.org/10.17605/OSF.IO/WMHCD (Bertozzi-Villa *et al.*, 2020)

Acknowledgements

We are grateful to the International Alliance for the Prevention of AIDS deployment team:

Nivetha Jagadeesh, Sheema Gopi, Arulraj Louis, N. Gayathri.

This project would not have been possible without the work of the game development team: Zachary Kohlberg, Christopher Blake, and Jacob Kohlberg.

References

Arnab S, Brown K, Clarke S, *et al.*: **The development approach of a pedagogically-driven serious game to support Relationship and Sex Education (RSE) within a classroom setting.** *Computers & Education.* 2013; **69**: 15–30.
Publisher Full Text

Baranowski T, Ryan C, Hoyos-Cespedes A, *et al.*: **Nutrition Education and Dietary Behavior Change Games: A Scoping Review.** *Games Health J.* 2019; **8**(3): 153–176.
PubMed Abstract | Publisher Full Text | Free Full Text

Bertozzi-Villa A: **bertozzivill/india-family-planning: Game data analysis and pre-post assessment.** 2020.
http://www.doi.org/10.5281/zenodo.3822455

Bertozzi E, Bertozzi-Villa A, Kulkarni P, *et al.*: **Collecting family planning intentions and providing reproductive health information using a tablet-based video game in India [version 2; peer review: 2 approved].** *Gates Open Res.* 2018; **2**: 20.
PubMed Abstract | Publisher Full Text | Free Full Text

Bertozzi-Villa A, Bertozzi E, Sridhar A: **Outcomes Assessment Pitfalls: Challenges to Quantifying Knowledge Gain in a Sex Education Game**. 2020.
http://www.doi.org/10.17605/OSF.IO/WMHCD

Boyle EA, Hainey T, Connolly TM, *et al.*: **An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games.** *Computers & Education.* 2016; **94**: 178–192.
Publisher Full Text

Chittaro L, Buttussi F: **Learning Safety through Public Serious Games: A Study of "Prepare for Impact" on a Very Large, International Sample of Players.** *IEEE Trans Vis Comput Graph.* 2020; PP.
PubMed Abstract | Publisher Full Text

Clark DB, Tanner-Smith EE, Killingsworth SS: **Digital Games, Design, and Learning: A Systematic Review and Meta-Analysis.** *Rev Educ Res.* 2016; **86**(1): 79–122.
PubMed Abstract | Publisher Full Text | Free Full Text

Coovert MD, Winner J, Bennett W, *et al.*: **Serious Games are a Serious Tool for Team Research.** *International Journal of Serious Games.* 2017; **4**(1).
Publisher Full Text

Del Blanco Á, Torrente J, Fernández-Manjón B, *et al.*: **Using a videogame to facilitate nursing and medical students' first visit to the operating theatre. A randomized controlled trial.** *Nurse Educ Today.* 2017; **55**: 45–53.
PubMed Abstract | Publisher Full Text

DeSmet A, Shegog R, Ryckeghem DV, *et al.*: **A Systematic Review and Meta-analysis of Interventions for Sexual Health Promotion Involving Serious Digital Games.** *Games Health J.* 2015; **4**(2): 78–90.
PubMed Abstract | Publisher Full Text

Ferchaud A, Seibert J, Sellers N, *et al.*: **Reducing Mental Health Stigma Through Identification With Video Game Avatars With Mental Illness.** *Front Psychol.* 2020; **11**: 2240.
PubMed Abstract | Publisher Full Text | Free Full Text

Fiellin LE, Hieftje KD, Pendergrass TM, *et al.*: **Video Game Intervention for Sexual Risk Reduction in Minority Adolescents: Randomized Controlled Trial.** *J Med Internet Res.* 2017; **19**(9): e314.
PubMed Abstract | Publisher Full Text | Free Full Text

Gariepy AM, Hieftje K, Pendergrass T, *et al.*: **Development and Feasibility Testing of a Videogame Intervention to Reduce High-Risk Sexual Behavior in Black and Hispanic Adolescents.** *Games Health J.* 2018; **7**(6): 393–400.
PubMed Abstract | Publisher Full Text | Free Full Text

Hieftje KD, Fernandes CSF, Lin IH, *et al.*: **Effectiveness of a web-based tobacco product use prevention videogame intervention on young adolescents' beliefs and knowledge.** *Subst Abus.* 2019; **42**(1): 47–53.
PubMed Abstract | Publisher Full Text

Ifenthaler D, Eseryel D, Ge X: **Assessment for Game-Based Learning**. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*. New York, NY: Springer New York, 2012; 1–8.
Publisher Full Text

Ismail S, Shajahan A, Rao TSS, *et al.*: **Adolescent sex education in India: Current perspectives.** *Indian J Psychiatry.* 2015; **57**(4): 333–337.
PubMed Abstract | Publisher Full Text | Free Full Text

Joubert M, Armand C, Morera J, *et al.*: **Impact of a Serious Videogame Designed for Flexible Insulin Therapy on the Knowledge and Behaviors of Children with Type 1 Diabetes: The LUDIDIAB Pilot Study.** *Diabetes Technol Ther.* 2016; **18**(2): 52–58.
PubMed Abstract | Publisher Full Text

Khubchandani J, Clark J, Kumar R: **Beyond controversies: sexuality education for adolescents in India.** *J Family Med Prim Care.* 2014; **3**(3): 175–179.
PubMed Abstract | Publisher Full Text | Free Full Text

Klopfer E, Haas J, Osterweil S, *et al.*: **Resonant games: design principles for learning games that connect hearts, minds, and the everyday**. Cambridge, MA: MIT Press, 2018.
Reference Source

Malone TW: **Toward a theory of intrinsically motivating instruction.** *Cognitive Sci.* 1981; **5**(4): 333–369.
Publisher Full Text

Serrano-Laguna Á, Manero B, Freire M, *et al.*: **A methodology for assessing the effectiveness of serious games and for inferring player learning outcomes.** *Multimed Tools Appl.* 2018; **77**(2): 2849–2871.
Publisher Full Text

Van Staalduinen JP, de Freitas S: **A game-based learning framework: Linking game design and learning outcomes.** *Learning to play: exploring the future of education with video games.* 2011; 29–54.
Reference Source

Weiser S: **India: Family Planning and the Population Dilemma.** *Pulitzer Center on Crisis Reporting.* 2015.
Reference Source

Zammitto V: **Game research, measuring gaming preferences.** *Appl Artif Intell.* 2009.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✘ ✔

Version 3

Reviewer Report 23 March 2021

https://doi.org/10.21956/gatesopenres.14451.r30420

✔ **Fabio Buttussi**
Human-Computer Interaction Lab, Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy

The authors addressed all the minor issues pointed out in the previous review and further improved their paper.

In particular, the link to the video is very useful to understand the methodology, the literature is now considered and discussed, and the minor problems concerning presentation were fixed.

In my opinion, the paper is ready to be approved.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research interests concern human-computer interaction, virtual reality, artificial intelligence, and serious games. I explore different application domains for my research such as aviation safety, health and fitness, emergency medicine, and sing languages, with a particular focus on training.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 24 December 2020

https://doi.org/10.21956/gatesopenres.14413.r29986

?

Fabio Buttussi

Human-Computer Interaction Lab, Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy

This article investigates the use of a serious game in a new interesting application domain, i.e. to teach adolescents about sex education. While the first version presented the research as a user evaluation that was affected by several methodological issues, the revised version successfully refocused the research as series of lessons learned from two user evaluations and the problems encountered by the authors. Some of these problems are peculiar to the chosen domain, while others could be generalized to other domains where serious games can be used. This latter aspect makes the contribution interesting also for computer science researchers as well as educators that are not working in the specific domain. Despite the new version cannot resolve the methodological issues in the design of the user evaluation (this is the motivation for my “Partly” answer to the question about technical soundness), the new focus makes those issues not critical, so I think that they will no longer prevent the acceptance of the paper.

The main reason why I will not yet approve the article instead concerns the very limited discussion of the results against the literature. I acknowledge that the considered application domain is largely unexplored, but serious games have been used in many other domains and some of the problems encountered by the authors have already been addressed in previous work. One peculiar example I am familiar with is the problem of players not devoting proper attention to assessment tests, which was described in a recent article.[1] In particular, that article described the criteria applied by the authors to identify a subset of users who completed pre- and post-tests before and after playing a serious game among the millions of users who played it. Moreover, that article shows how in-game metrics could be used to assess improvement in the taught topic also among users who did not complete the pre- and post-tests, and the authors of the submitted paper or other researchers might be interested in applying such an approach to serious games for sex education in their future work. More generally, I think the article will benefit from a discussion of the results that compare authors’ findings with those that emerged from previous research on serious games (e.g., see the systematic reviews about serious games).[2,3,4]

While the authors provided the links to code and installers of the serious game, I think that methods could be improved by adding some more details about how the serious game works, because several readers may not be willing to install the serious game on their devices and the article provides only one example of an included mini-game. A video of the gameplay would also be very useful.

As a minor issue, there are some problems with the terminology used in the paper. For example, pre- and post-test are called in different ways in different parts of the paper (post test, post-test, post- test, post assessment, post-game assessment). Similarly, serious games are sometimes referred to as applied games, and this can make the readers wonder if there are any differences between the two terms. It would also be useful that the authors clearly report the scale of scores of pre- and post-tests: the pre-test score is first reported as a percentage, but later the gain is reported in points, and the maximum score is reported only in Figure 9.

Overall, I think that the revision solved almost all critical issues and a minor revision would be sufficient to discuss the results against the literature and fix some minor language or presentation

issues.

References

1. Chittaro L, Buttussi F: Learning Safety through Public Serious Games: A Study of. *IEEE Trans Vis Comput Graph*. 2020; **PP**. PubMed Abstract | Publisher Full Text
2. Connolly T, Boyle E, MacArthur E, Hainey T, et al.: A systematic literature review of empirical evidence on computer games and serious games. *Computers & Education*. 2012; **59** (2): 661-686 Publisher Full Text
3. Boyle E, Hainey T, Connolly T, Gray G, et al.: An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education*. 2016; **94**: 178-192 Publisher Full Text
4. Clark DB, Tanner-Smith EE, Killingsworth SS: Digital Games, Design, and Learning: A Systematic Review and Meta-Analysis.*Rev Educ Res*. 2016; **86** (1): 79-122 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
No

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research interests concern human-computer interaction, virtual reality, user-adaptive and context-aware systems, mobile devices, and serious games. I explore different application domains for my research such as aviation safety, health and fitness, emergency medicine, and sing languages, with a particular focus on training.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 09 December 2020

https://doi.org/10.21956/gatesopenres.14413.r29984

Melissa Gilliam
Department of Obstetrics and Gynecology, University of Chicago, Chicago, IL, USA

The changes are adequate. Thank you for the opportunity to review again.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: ASRH, game-based learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 15 September 2020

https://doi.org/10.21956/gatesopenres.14316.r29409

Claudia Regina Furquim de Andrade
Department of Physiotherapy, Speech Therapy and Occupational Therapy, Faculty of Medicine FMUSP, University of Sao Paulo, Sao Paulo, Brazil

I would like to thank you for the opportunity to review the manuscript entitled “Outcomes assessment pitfalls: challenges to quantifying knowledge gain in a sex education game” that has been submitted to Gates Open Research.

The study investigated and discussed the challenges associated with incorporating educational assessment before and after the use of an educational game entitled 'My Future Family Game', delivered to students of three schools in and around Chennai, India. The authors investigated the efficacy of a pre- and post-test assessment. I consider this topic relevant and interesting, and in need of further investigation, especially with the application of experimental designs. Therefore, the authors are commended for the novel and trendy approach. However, I note several methodological concerns. Among these, relevant literature regarding the use of educational games, particularly concerning outcomes assessment, is not reviewed (and was not used to guide the research design); inclusion and exclusion criteria, for both participants and schools, are not specified and may not have been controlled; and the study methods are not described in sufficient detail. For these reasons, I am concerned that it is not possible to fairly review the results and conclusions reached. Furthermore, I would suggest publishing this manuscript as a research note, since it mostly describes the unexpected observations and lab protocols.

Comments on the Introduction:

- In general, the background/rationale for this study is not sufficiently developed as written. The study is not hypothesis-driven and predicted outcomes are not provided.

- The introduction, albeit interesting, is essentially a description of the game used for this study, detailing its design and development.

Comments on the Methods:

- I believe most of the information concerning the game development and pilot studies could be presented, in short, in the methods section, with citations to the previous published studies. In addition, the methods section should clearly and succinctly explain the study methods, so that it can be duplicated.
- Inclusion and exclusion criteria, for both participants and schools, are not provided (for example, were the schools chosen only by convenience? What about the student participants in each school, how were they selected?). Demographic information about included participants is not very substantial either. The design of the pre- and post-test assessment is also not clearly described.

Comments on the Results:

- The manuscript depicts a lack of interest from many participants in completing the educational assessments. I suggest that the authors elaborate on what they believe was causing this lack of interest. Could it be influencing the results?

Comments on the Discussion:

- The manuscript would benefit if the authors discussed their findings on relevant literature. The authors did not cite and engage with the pertinent literature regarding educational

assessment.

Is the work clearly and accurately presented and does it cite the current literature?
No

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: I am a Full Professor at the Faculty of Medicine at the University of São Paulo, and published 155 articles in indexed journals and 376 papers in the annals of events. I have additional 96 publications, between book chapters and books, and 361 items of technical production. Between 1989 and 2020 I participated in 24 research projects, 22 of which I coordinated. I am also the coordinator of three specialization courses.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 28 Oct 2020

Elena Bertozzi, Quinnipiac University, Hamden, USA

Thank you very much for the detailed and thorough feedback in this review. We have implemented your suggestions as detailed below.

1. Added literature review.
2. Added inclusion and exclusion criteria.
3. The study is not hypothesis driven. The paper seeks to add to the scarce literature in the field relating to assessment of the efficacy of serious games to educate players about sexual anatomy and reproduction.
4. The introduction and other sections of the paper have been revised to clarify the paper's purpose and arguments.
5. The results section was edited to clarify the reasons for the lack of participation in the

post test and add suggestions to resolve the issues we encountered.
6. The discussion section was edited to clarify connections to the pertinent literature.

The paper was reorganized to clarify the purpose of the study and how publication of the issues we met and resolved could be useful to other researchers.

Competing Interests: No competing interests were disclosed.

Reviewer Report 17 August 2020

https://doi.org/10.21956/gatesopenres.14316.r29297

Melissa Gilliam
Department of Obstetrics and Gynecology, University of Chicago, Chicago, IL, USA

Thank you very much for the opportunity to review this paper. In it, the authors consider how to integrate knowledge assessment into a sexual health game intervention entitled 'My Future Family Game' delivered to young people in Chennai, India.

This game-based intervention aims to collect information about family planning intentions. Here, the authors consider the efficacy of a pre-post assessment and the post-test assessment in particular. They describe the development process and some of the pitfalls of the approach they took.

Overall, the paper is interesting and I enjoyed reading about the logistics and decision making that the team underwent. Working internationally on game development can be difficult and it was interesting to see how the team tackled field research. That said, this seems to be a report about processes and decisions that were made rather than a generalizable paper couched in the larger literature.

Specific points:

Introduction:

- If this paper is about game assessment, then the introduction should provide a literature review on that topic.

- The introduction turns into a description of pilot game design and development. Instead, this information perhaps belongs in the methods section.

Methods:

- The methods section should be the planned study or evaluation. Instead it seems almost like a results section in that you state how many people were involved etc.

- The section on challenges from first deployment again seems like the result section. Regardless, it is a bit confusing as to how this relates to an assessment of knowledge.

Results:
- It seems that the many pitfalls are avoidable and others have successfully done game-based assessment. It is not clear whether these are particular to the ways the researchers set up the assessment or inherent to assessment, I would argue for the former.

Discussion:
- Statements such as “our response to the results was dismay” is quite casual and instead perhaps the discussion could be used to review the literature on game assessment.

- It is not clear that pre post assessment issue is particular to sensitive topics and can arise for other reasons

Is the work clearly and accurately presented and does it cite the current literature?
No

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: ASRH, game-based learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 28 Oct 2020

Elena Bertozzi, Quinnipiac University, Hamden, USA

Thank you for this thoughtful review and for the suggestions that guided our revisions to the paper.

Specific Points:

Intro: We added a literature review and analysis of results which we used to focus the introduction of the paper. The introduction now clarifies that the topic of the paper is refining the process of outcomes assessment for a serious game, specifically one that addresses sexual anatomy and reproduction.

Methods: We reorganized the paper so that the methods section explains the process that we undertook based on the first and second deployments of the game and problems associated with the addition of the pre/post assessment.

Results: There is a growing body of research on game-based assessment, however there is little published work in the area of games related to education about sex and reproduction. The assessment of knowledge gain without in person interviews is complicated by numerous factors and our paper seeks to provide guidance in this evolving area.

Discussion: We have revised this section to remove unprofessional language and clarify the connections between the assessments and issues relating to serious game deployments.

The paper has been reorganized for clarity.

Competing Interests: No competing interests were disclosed.