Moving Beyond Fun: Evaluating Serious Experience in Digital Games

Ioanna Iacovides
UCL Interaction Centre
Gower Street
London, WC1E 6BT, UK
i.iacovides@ucl.ac.uk

Anna L Cox
UCL Interaction Centre
Gower Street
London, WC1E 6BT, UK
anna.cox@ucl.ac.uk

ABSTRACT
Games are normally considered to be "fun", though recently there is growing interest in how gameplay can promote empathy and encourage reflection through "serious experience". However, when looking beyond enjoyment, it is not clear how to actually evaluate serious experience. We present an evaluation of four games that were submitted to a student game design competition; the competition challenged teams to design a game that inspired curiosity around human error and blame culture within the context of healthcare. The entries were judged by a panel of six experts and subjected to a round of play testing by twelve participants. Methods included gameplay observation, questionnaires, post-play interviews and follow-up email questions. We discuss the utility of these methods, with particular emphasis on how they enabled a consideration of the immediate and longer term impact of serious experience on players.

Author Keywords
Games; evaluation; critical play; engagement; positive experience; negative experience; serious experience.

ACM Classification Keywords
H.5.3 Information interfaces and presentation (e.g., HCI): Miscellaneous ; K.8.0. General: Games.

INTRODUCTION
Digital games are an immensely popular leisure time activity and are increasingly being used to persuade within more serious domains such as learning, advertising, politics [3] and behavior change [e.g.16]. For instance, educators have long wanted to "harness the motivational power of games" to make learning more fun [14; p.4]. However, despite the focus on enjoyment as a significant component of the player experience [19], Marsh & Costello [18] argue that HCI and game designers need to look beyond the notion of fun to consider a wider range of emotional experience. They propose the term "serious experience" to cover experiences that are (1) *uncomfortable, negative and/or unpleasant* and/or (2) *entertaining without being exclusively fun*. As in Benford and colleagues work on uncomfortable interactions [2], the aim is not to create long term discomfort or pain, but to provide worthwhile experiences with benefits such as raised awareness or critical engagement with a serious subject matter.

In addition, Flanagan [7] argues for *critical play*; where games can be used to communicate empathy and enable reflection on different aspects of life. Board games such as *Train* (Brenda Romero) – which raises questions about complicity during the Holocaust – and digital games like *Hush* (Jamie Antonisse & Devon Johnson) – where you play a Rwandan Tutsi mother trying to sing her child to sleep as Hutu soldiers approach their home – and *A Closed World* (Gambit Game Lab) – which tackles sexuality and identity – illustrate how gameplay attempts to provoke thought on serious issues. However, the extent to which players have engaged with the take-home messages of these games is often unclear as evaluation rarely seems to be elaborated on as part of the design process.

Persuasive games with concrete goals such as an improvement in learning outcomes or a change in behavior have been assessed by a range of methods such as pre and post intervention tests [e.g. 22], surveys and in-game performance [e.g. 6], think aloud and interviews [e.g. 13]. There is usually a focus on enjoyment, where evaluation methods are used to maximise the chances that the player has a positive affective experience [21]. However, when looking beyond fun in terms of gameplay, it is not clear what the criteria for success are nor how best to evaluate games which aim to promote discussion and reflection on societal problems and challenges.

In this paper, we explore the question of "How to evaluate serious experience within games?" through presenting an assessment of four entries to a student game design competition. Our evaluation was the final step in a design process that was influenced by participatory approaches [11]. We aimed to establish which of the games was most likely to inspire reflection and encourage further exploration of the competition topics: human error and

CHI 2015, April 18 - 23 2015, Seoul, Republic of Korea

ACM 978-1-4503-3145-6/15/04…$15.00
http://dx.doi.org/10.1145/2702123.2702204

blame culture within the context of healthcare. We also reflect on our mixed method approach and consider our findings in relation to evaluating similar games outside of a competition format.

GAMES THAT AREN'T EXACTLY "FUN"

As opposed to "serious games" (i.e. non-entertainment games), Bogost [3] uses the term "persuasive games" to cover games used for any purpose (entertainment, education, activism etc.), which aim to persuade players by delivering a particular message or argument. For instance, *The McDonald's Videogame* (a critique of the global fast food industry) illustrates how corruption within the industry is a systemic problem as the player must resort to dubious business practices in order to do well in the game.

A related term, "critical play" is used to refer to "play environments or activities that represent one or more questions about aspects of human life" [7; p. 6]. Critical play can be reflected in a variety of ways, including subverting game design mechanics in order to challenge player expectations and modes of thinking. The board game *Anti-monopoly* for example, illustrates the harmfulness of monopolies by reversing familiar *Monopoly* conventions.

Taking subversion a step further, Wilson and Sicart [24] introduce the notion of "abusive game design". Similar to a critical design approach, abusive design practices challenge standard usability paradigms by making games very difficult to play e.g. through introducing physical pain, implementing unfair game mechanics or involving embarrassment. For instance, when a ball is missed in *PainStation* (a modified version of *Pong*), the game physically punishes players with heat impulses, electric shock or an integrated miniature wire whip. The purpose is not entirely clear but the authors state that through adopting an abusive game design "attitude", play is made more personal through a dialogic relationship between the (not necessarily co-located) player and designer; the game essentially becomes "an open invitation to explore the extremes of gameplay experiences, together" [p. 46; 24].

In addition to visceral discomfort (e.g. inflicting pain), Benford and colleagues [2], discuss how "uncomfortable interactions" can be produced through cultural discomfort (e.g. having to confront challenging themes such as terrorism); control (e.g. surrendering control to others) and intimacy (e.g. employing voyeurism). Further, the authors argue that that fun can consist of more than just pleasurable sensations, for instance, the thrill and suspense of a rollercoaster ride. Within the realm of cultural experiences, they discuss how interactions that cause a degree of suffering to the user are implemented for the purposes of entertainment, enlightenment and connecting with others.

For example, the performance *Ulrike and Eamon Compliant* asks participants to assume the role of a terrorist (Ulrike or Eamon) as they walk through the city while receiving phone calls on the way to an interrogation in a hidden room. This interactive performance aims to enlighten participants through engaging them in "a dark and challenging theme, while also involving an unusual and discomforting form of sociality" [p. 2008; 2].

Drawing on the work of Benford and colleagues [2] and Montola [20] (who explored positive-negative experiences in extreme role-playing games), Marsh & Costello [18] point out that a focus on enjoyment may lead designers to see negative affect as something to avoid at all costs. Such an approach potentially makes it harder to offer players alternative experiences that are both deep and powerful. Instead, Marsh & Costello (ibid) turn to media, drama, performance, literature, music, art and film that have a longer history of shaping a wider range of experience and emotion e.g. in order to illustrate suffering and adversity. As a result, they propose the term "serious experience" to cover experiences that are (1) *uncomfortable, negative and/or unpleasant* and/or (2) *entertaining without being exclusively fun*. The latter refers to experiences that are thought-provoking or that alternate between positive and negative emotion. The authors argue that designers of serious games should aim for an appropriate rhythm between fun and seriousness and that extreme experiences that cause player discomfort can be used to raise awareness and prompt reflection. Further, they stress that, in order to fulfill a persuasive purpose the "experience with persuasive technology and games needs to resonate or linger with the user/player after an encounter" [p. 116; 18].

SERIOUS EXPERIENCE IN DIGITAL GAMES

Some early examples of digital games that attempt to provoke reflection are *Kabul Kaboom* and *September 12th*. Both are described as "socially or politically critical games" that you can never win, thus invoking "more pain than pleasure" [15]. *Kabul Kaboom* illustrates the contradiction of the US air force attacking the Taliban in Afghanistan whilst simultaneously air-dropping food; the gameplay involves controlling an on screen avatar to collect hamburgers raining down while avoiding bombs. When the avatar is inevitably hit, the final scene is littered with body parts and debris, while a voice states "Mm, yummy". In *September 12th,* the player controls a cross hair for launching missiles on an unidentified Middle Eastern town. The bombs kill terrorists but also generate collateral damage, where civilians mourning innocent victims soon turn into terrorists themselves. Lee [15] provides an insightful critique of these games and argues they are a new medium of expression, but it is not clear from his account how players react to gameplay nor whether they actually go on to critically reflect on issues such as war and terrorism.

Flanagan [7] discusses *September 12th* as a game that involves critical play and has been designed to make people think and provides *Hush* as another example. *Hush* was produced as part of the Values@Play project [8], which aims to help designers integrate human values into the design process. In the game, players must type letters

appearing on the screen from the lullaby sung by a Tutsi mother, Liliane, who is trying to calm her child to avoid detection from soldiers approaching their home in Rwanda. Belman & Flanagan [1] convincingly argue for how the game is able to foster empathy and also refer to player accounts of escalating tension and dread. However, again, there is no mention of exactly how the game was evaluated and what messages players took away from their gameplay.

Another example of critical play is *Blowtooth* [17], a pervasive mobile game that enables players to smuggle virtual drugs within real world airports by using unknowing bystanders. Linehan et al., [17] produced the game to demonstrate how the real world environment (airports, in this case) can be used to enhance the experience of pervasive games, though the authors also suggest the game was able to stimulate critical thinking about airport environments. As part of the evaluation, six participants were recruited to play the game whilst travelling through airports. After doing so, they were asked to fill in two questionnaires (one on game enjoyment, the other on levels of anxiety and awareness). Open-ended questions were also included and deemed "equally, if not more valuable" (p. 2701) due to the small sample size. Given the emphasis of security within airports and the controversial subject matter of the game, the authors were expecting the game to alternate between positive and negative experience. However, they were surprised to find that while players generally enjoyed playing the game, they reported low levels of anxiety, as well as low awareness of security and other passengers. On the basis of the open-ended responses, (and despite the quantitative findings) the authors argue there was evidence the game led to critical thinking, at least in terms of players being more aware of the airport environment e.g. where passengers are made to wait.

In a different approach, Ruggiero [22] carried out a large scale quantitative evaluation of *Spent*, a persuasive game about poverty and homelessness where players have to try and survive as a single parent on $1000 a month. The evaluation involved 5139 participants in 200 classrooms across four US states. The study used the Affective Learning Scale [ALS; 23], which essentially appears to assess attitude changes towards particular content. In an immediate post-test, the game group (who played *Spent*) and reading group (who read an article on homelessness) scored significantly higher than the control group (who only took the tests), though there did not seem to be a significant difference between the game and reading groups. However, at a three week post-test, while all scores decreased, the game group still had significantly higher scores than both other groups. The game was found to significantly improve attitudes towards poverty and homelessness, but it is unclear exactly what players felt during the game or whether the game led to any form of critical reflection. Further, the focus on ALS scores and the lack of description relating to the game make it difficult to consider the ways in which it was able to influence affective learning.

Marsh and Costello [18] mention plans to evaluate serious experience within a Great Barrier Reef game, but this appears to be work in progress. Apart from stressing the need to consider lingering experiences in addition to moment-by-moment play, it is not clear how the authors plan to assess their game. While serious experience may lead to reflection on particular issues, we don't yet know how to establish whether this reflection actually takes place.

In summary, there are few examples of how serious experiences in games have been evaluated, particularly in relation to uncomfortable experiences and how effectively these may raise awareness and provoke thought on specific societal issues. In order to be able to judge the entries to a student game design competition, we developed an approach involving expert judging and play-testing to establish which of the entries was likely to inspire reflection and encourage further exploration of the competition topics. The competition enabled us to develop and assess techniques for evaluating serious experience in games. The following section outlines the competition before introducing our methods.

THE GAME DESIGN COMPETITION

Overview

As part of the CHI+MED research project, a persuasive game design competition was held to create games for an accompanying project website, Errodiary.org. CHI+MED is investigating ways to reduce errors in the domain of healthcare and improve patient safety; Errordiary is a public engagement portal for human error and related topics. For the competition, the student teams were challenged to design a game that inspired curiosity and reflection on human error and blame culture e.g. that got players thinking about the fact that individuals get blamed when the wider system is at fault. A kick-off event was held in February 2014, with presentations from experts in human error, blame culture, healthcare and game design, followed by a Q&A session and a game design workshop. A website with information about the competition and further resources was developed to support the teams during the design process. The teams had to fill in a submission form, describing their game and how it was designed, before the entries were evaluated. Prizes were awarded at a final showcase in May 2014.

Nine student teams registered for the competition and four submitted entries before the deadline. The four teams consisted of 2-4 undergraduate and postgraduate students from five universities, across departments in Computer Science, Communication, Psychology and Medicine within the UK. The entries are presented below in alphabetical order.

The entries

Medical Student Errors (Figure 1) was created by Devon Buchanan and Angela Sheard. It is an interactive fiction

about a day in a life of a junior doctor. Through a text-based interface the players is presented with a number of scenarios relating to how people make and communicate errors. The player can move backwards and forwards through the narrative, exploring different dialogue options and finding out more about particular concepts through hyper-links.

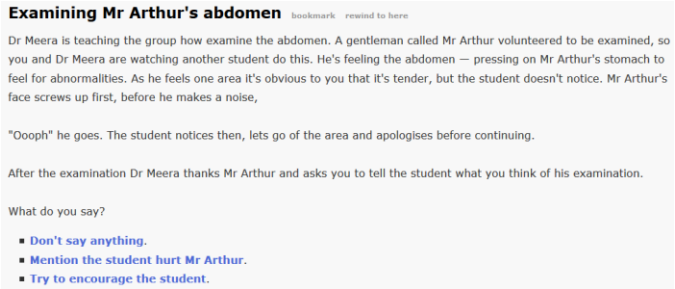

Figure 1: Medical Student Errors

Figure 2: Nurse's Dilemma

Nurse's Dilemma (Figure 2) was created by Adam Afghan, Andrew Gorman, Natasha Trotman and Jining (Kea) Zhang. The player is cast in the role of a nurse faced with a series of challenges during her daily tasks. The game uses a text-based interface with simple audio and graphics. The designers describe it as an empathy based game that aims to shed light on the pressures, constraints and stresses that nurses are expected to deal with every day.

Patient Panic (Figure 3) was created by Cameron Kyle-Davidson, Lydia Pauly, Benjamin Williams and Connor Wood. The game is set during a natural disaster where the player is a local doctor who was to treat multitudes of patients before they expire. Like *Tetris*, there is no win state, the game gradually increases in difficulty until the player runs out of lives and is fired for their inability to cope. The game employs a simple point and click interface, animations and a soundtrack involving ambulance sirens.

St. Error Hospital (Figure 4) was created by Charmian Dawson and Subhan Shaffi. The game utilizes a bird's eye view of a hospital where players take on a management role: balancing a budget, directing staff, organizing ward areas and implementing strategies that aim to reduce the likelihood of errors (resilience strategies). In addition to an overview of the ward, the game also displays information reports and graphs to provide the player with feedback on their performance. In terms of audio, a background hum is present throughout the game to indicate ward activity.

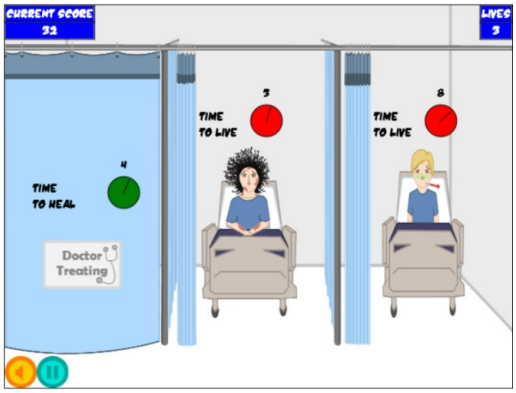

Figure 3: Patient Panic

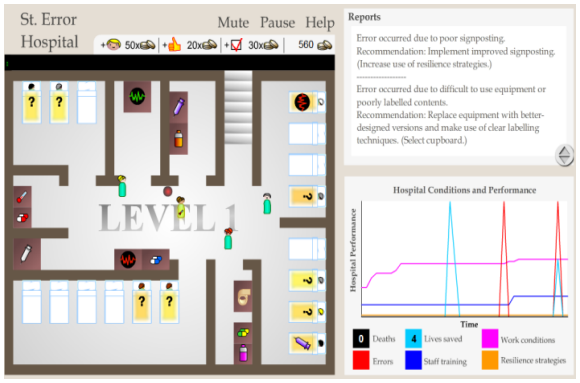

Figure 4: St. Error Hospital

METHODS

Our approach involved a mix of methods to establish which of the entries was likely to inspire reflection and encourage further exploration of the competition topics.

Expert judging

Six judges, with expertise across human error, user experience, game design and healthcare were asked to play the games and fill in a feedback form for each one. The form asked for their general impressions; the extent to which they thought the game had potential to inspire curiosity about the competition topics; and for them to rank the games according to their overall impression of the extent to which each entry addressed the competition aims.

Play-testing

Design: This was an observational study of gameplay that included a post-play interview and follow-up emails.

Participants: Twelve participants (9 female, 3 male) were recruited through a university participant pool (mean

age=23.3; sd=3.3). The only requirement was that they at least occasionally played video games.

All the participants had started playing games by the age of 13, with the most frequent age range being "8-10 years" (5/12). Frequency of play ranged from once every 3 months to daily, with the most common range being "2-3 times a week" (5/12). Gameplay sessions lasted from "less than ½ hour" to "between 2-3 hours", with "1-2 hours" being most common (4/12). The most frequently used gaming platforms were mobile phone (11/12), PC/laptop (9/12) and tablet (6/12). When asked about the games they had recently played, players mentioned a range of titles from casual games such as *Candy Crush* and *Flappy Bird* (9/12) to hardcore games like *Bioshock Infinite* and *Call of Duty* (5/12). While most were familiar with the term "human error", none of the participants had any expertise in human error research and only one had visited Errordiary before the play-testing sessions (P1).

Materials: The evaluation took place in a lab, where the games were played on a Windows laptop. Screencast-o-matic was used to record the gameplay. Participants filled in a questionnaire about their gaming habits and preferences before the session began. An additional questionnaire was filled in after each game, which included open-ended questions about what they liked most and least about each game, and how many stars they would award it out of five.

Procedure: Sessions lasted no more than 2 hours, where participants played each game for up to 15 minutes (order counterbalanced) and answered a short questionnaire on each. The session concluded with a final interview where players were asked to rank the games in terms of (1) gameplay quality and (2) how well they inspired curiosity and reflection on the competition topics. Two days after the session, players were sent a follow up email to assess whether any of the games led to "lingering" experiences. The email asked whether they had discussed the games with anyone else, whether they had been thinking about any of the games and whether they had gone on to explore the Errordiary website. Participants were paid £10 after the session and sent a £10 Amazon voucher after replying to the email questions.

FINDINGS

Expert judging: rankings

Ranking	Mode	Median
Nurse's Dilemma	1	1.5
St. Error Hospital	2	2
Medical Student Errors	3	2
Patient Panic	4	4

Table 1: Judges' ranking

Table 1 indicates how the judges ranked the games, where *Nurse's Dilemma* was considered to be the competition favorite, closely followed by *St. Error Hospital*.

Play-testing: ratings and rankings

While the post-play questionnaire requested quantitative information – the star ratings for each game – the questionnaire's main purpose was to allow players to note down their initial reactions that could then be used as a prompt during the interview. The star ratings are provided in Table 2, where *St. Error Hospital* scored highest.

Game	Mean	SD
St. Error Hospital	3.1	1.1
Nurse's Dilemma	2.8	1.5
Patient Panic	2.6	1.3
Medical Student Errors	2.3	1.1

Table 2: Star ratings for each game

Players were also asked to rank the games in order of their most to least favorite in terms of gameplay (Table 3) and in order of most to least likely to lead to reflection about human error and blame culture (Table 4). *St. Error Hospital* was seen as the most game-like of the entries, though *Nurse's Dilemma* came out on top in terms of inspiring curiosity and reflection. The rankings were used as further prompts for discussion in the post-play interview.

Ranked in terms of gameplay	Mode	Median
St. Error Hospital	1	1
Patient Panic	2	2
Nurse's Dilemma	3	3
Medical Student Errors	4	4

Table 3: Gameplay ranking

Ranked in terms of reflection	Mode	Median
Nurse's Dilemma	1	1.5
St. Error Hospital	2	2
Medical Student Errors	2	2
Patient Panic	4	4

Table 4: Reflection rankings

Qualitative analysis

The judges' submission forms, the open ended answers from the participant questionnaires and the participant interview transcripts were collated in Nvivo 8 and coded for: (1) discussion of topics related to human error and blame culture within healthcare; (2) positive and negative comments about each entry; and (3) players' emotional reactions. The games dealt with serious topics and as such there was less emphasis on how "fun" people thought the games were and if they would play them again, and more on whether the game led to a consideration of human error and/or blame culture and how the player felt after playing them. The feedback is summarized below, where each game is presented in alphabetical order. Participants are referred to by number e.g. P1 refers to Participant 1, as are judges e.g. J1 refers to Judge 1.

Medical Student Error

The judges praised the game for being simple to play, the interactive fiction format, its focus on communication and the links provided to the Errordiary website. The final scene was noted for providing players with an opportunity to reflect and investigate further. For instance, J6 stated "*I also liked that the game gave me summary of what I had learned at the end and provided links to Errordiary to find out more about errors and blame culture*." However, the amount of text and specialist language meant there were concerns about the intended player audience and whether non-medical students would find the game relevant e.g. "*It might be interesting for perhaps first year medical students to learn about ethical issues but I'm not sure that the general public would find it much fun to play*" (J4).

In the play-testing, participants generally found the game easy to play, and liked that they could go back and change options. For example, in response to the question on what they liked about the game, P7 stated *"the player gets to choose from a lot of different options"*. Some also commented that that they could learn a lot from playing the game about working in a medical context and the experience of junior doctors, while others appreciated that it allowed them to reflect on their own behavior in different circumstances. P9, for instance, said "*it gives me the chance to think how I will behave in certain situations and my reasons for doing it*". Within the play-testing session, many of the players clicked on the various links provided and some spent the remainder of the 15 minute session reading articles on Errordiary. As a result of the game and the links they explored, a few of the players did go on to reflect on the issues related to the competition topics such as the frequency of errors in a medical context (P8: "*It is interesting to spot so many human errors in the hospital*"); blame culture (P6: "*sometimes you may forget too, so you should not have that blame culture of it because all people will make mistakes from time to time*") and resilience strategies (P4: "*so I was thinking how people do their things in their everyday life and the strategies they use*").

However, many of the players had trouble reconciling their expectations of games in general with the interactive fiction format of *Medical Student Errors* – many didn't see it as particularly game-like or engaging e.g. "*While the information given was excellent it was very dull, like reading lecture notes. There was no game aspect just reading information*" (P5). Some of the reasons for this include the lack of graphics, the amount of information/description provided and the fact the game did not have a clear goal. For players such as P4 this was disconcerting: "*I went through it and at the end, I thought I started from somewhere, I came to something else and somehow I felt there was no connection in between*". Further, while the inclusion of different options was seen as a positive, P10 noted that the consequences of different actions were not always clear: "*there weren't any real results from choosing particular options – there weren't any "right" options. Previous options/choices did not affect how the next scene played out*".

Nurse's Dilemma:

The judges were generally impressed by how the game was able to create an emotionally compelling experience. For instance, J5 stated "*the game was able to engage me on an emotional level and I was genuinely torn about what I should do in some of the scenarios. The end was also very good at explaining the game and how to find out more*". While the game was able to effectively communicate how individuals have to deal with wider systemic issues and blame culture, it could have gone further in terms of linking these issues back to specific occurrences of human error. Further, not everyone appreciated having to install the Unity plug-in to play the game (even if it was "*worth it*", J6) and one of the judges found the game "*too slow and depressing*." (J1).

In the play-testing, *Nurse's Dilemma* was seen as easy to play and the game was most likely to elicit an emotional reaction from the players, where many were seen to exhibit empathy with nurses and the decisions they have to make. For instance, "*it was very emotionally engaging as you were reading it*" (P11). While player experiences were not necessarily comfortable, the music and the way the text appeared added to the compelling nature of the game e.g. P3: "*there was something about the sentence by sentence that came up with the music … it's like it painted a picture, like you're in that world of it*". Further, when discussing the game during the interview, participants would engage in topics related to human error such as blame culture (P6: "*I think it's quite common to have the blame culture inside a hospital, but I think your colleagues should understand about it because they all suffer, they all experience the same situation*"); demands on nurse's time (P1: "*it makes you think a nurse's job requires all sorts of things and that you can't just focus on one task at a time*"; and ethics (P10: "*it actually raises a lot of issues in terms of the difficult moral choices that the nurses have to make, and then at the end it's got that dialogue explaining all the issues*").

In terms of negative feedback, a minority of players disliked the text format of the game and the amount of reading required while there were some issues with the text e.g. "*the words are small*" (P3). In general, P12 particularly disliked text-based games as "*I don't enjoy reading so I found it really boring*". For some, the game was also seen as being too depressing e.g. "*it's really sad and really helpless*" (P9). Further, despite being considered an accurate representation by the healthcare judge, one or two participants were unsure as to how realistic the game was and questioned whether the situation was actually that bad for nurses e.g. P11 stated: "*I felt the options were restrictive and unrealistic as well as the scenario*."

Patient Panic

The judges were positive about how the game was simple to play, the look and feel of it, and the fact you could chose different difficulty levels. For instance, J2 noted "*there were quite a few creative touches – like the title 'Patient Panic', having optional music, a tutorial, beginner/ advanced options*". However, some of the judges did not find the game to be engaging and there was a general concern about whether the game went far enough in terms of relating the gameplay to the competition topics. For instance, "*The games gives an idea of the stresses involved in being an A&E doctor but does not give a lot of detail about the background to the situation*." (J4). While the ending did hint at the problem of blame culture it did not give the players any way of finding out more about the topic nor did it explain the game's negative ending (being fired for incompetence).

In the play-testing, *Patient Panic* was seen as one of the more game-like competition entries, where many players appreciated how the game had clear goals, timers, different levels of difficulty, points, and replay value. It was described by P1 as "*it's a very simple, easy game, you could probably play with it on the phone as well, and it's fun*". Some also reported that the game was effectively able to induce a sense of being "*panicked*" (P2). Further, it was seen to have replay value as many played it several times during the 15 minute sessions so they could try and do better. A couple of players also engaged in discussion about competition relevant topics in relation to the game, such as demands on doctors' time (P4: "*The doctor can only do what he can do as he's only one doctor in the hospital as per the situation. I'm curious about if they would have more staff, more doctors to treat patients, that we could have saved more lives*"), and over stretched resources (P9: "*because there are so many patients at the same time, so sometimes I think a doctor can only choose maybe the most urgent ones. He doesn't have many choices*").

However, the play-testing sessions did reveal gameplay issues as the game was not seen to be engaging for all players. For instance, P5 described the game as "*the whack a mole, it just seemed a bit pointless, there wasn't really much information on errors or anything, it was just pressing and then it got really tired of clicking all the time*". While the instructions were generally seen as useful, it sometimes took a while for players to notice elements such as the number of lives left and players were confused about how points were calculated. The "difficult" level was also found to be "*impossible*" (P6). In addition, even for those who enjoyed playing the game, the experience only occasionally led to further discussion about the competition topics. The final screen left many feeling confused about why they were being declared unfit and participants did not feel they had learnt much from playing the game e.g. P10 says the ending "*just feels like something that's thrown in because it's related to the game … nothing in the game actually makes you wonder about real life situations."*

St. Error Hospital

The judges praised the entry for its engaging gameplay and the way in which it was able to highlight the complexity of human error. It received positive feedback about the style of the game and how it was able to incorporate concepts such as resilience strategies, staff training, and quality of work environment e.g. J3: "*First impressions is this is great and they have made a real effort to engage with the concepts… generally this seemed very deep and ambitious*". However, the game was also found to be quite difficult to play and there was a concern that it might be too ambitious, where "*players will be put off by the complexity of the game (and will miss things, like the headlines at the top)*" (J4). Further, in advanced mode, it was noted players can actually get quite far in the game after firing all but one nurse.

During the play-testing, *St. Error Hospital* was rated as the most game-like out of the entries. Participants found it to be an engaging experience, appreciating the graphics and "*being given a challenge*" (P12). The game was seen as a positive spin on human error as, while it showed how things could go wrong, it also gave them opportunities to improve e.g. "*it's not only not to let the patient die, it's to improve the way the staff move as well*" (P7). It was also found to have replay value since the goals are clear and there were multiple variables to play with. Further, during the interview, participants would discuss the game in relation to relevant human error topics such as training (P4: "*you're more curious about if they've not been trained, have they been lazy or they don't know what they're doing, or there's this budget problem or they don't have the resources?*"), and staff levels (P8: "*Then when people were dying and I couldn't control it, it's caused by external factors like human errors. It was mostly due to the lack of nurses*").

However, the game was also seen as being the most difficult game to play. While the tutorial was helpful, for many it didn't go far enough in terms of explaining how to play and players had difficulty with certain actions e.g. P1 "*there was stuff that I could click on, but I didn't know what I was clicking or what I was doing. It took me a few trials to understand that I had to click on the red tick to deduct money*". Further, while the game provides a lot of useful information it was clear from the sessions that players weren't always able to take it all in. For instance, P3 (thinking a nurse was leaving to go on a break, rather than quitting due to poor work conditions) picked up a member of staff whilst stating "*No breaks! Where are you going missy?*" and placed her back in the ward to continue working. This behavior indicates that the message of the game did not always come across clearly. Unfortunately, there was further evidence from the sessions that the game could lead to a sense that human error can be eradicated through the constant surveillance of staff: "*at the same time there's the message of human error, it doesn't really feel that way, you feel more omnipotent*" (P11).

Follow-up emails

Nurse's Dilemma was most frequently mentioned in the follow-up emails by players (6/12) and was the most likely to resonate with players in terms of getting them to think about topics related to human error e.g. "*I have been thinking about how much effort a nurse would need to take to do his/her jobs well*" (P9). *St. Error Hospital* was mentioned in the follow-up emails by 5/12 players. Though sometimes referred to in relation to thinking about human error related topics such as staffing issues, this was to a lesser extent than *Nurse's Dilemma* as the game was also mentioned in relation to "*thinking about the strategies of playing that game*" (P6).

Medical Student Errors was mentioned in the follow-up emails by 3/12 participants: where P10 mentioned discussing the game with medical student friends. *Patient Panic* was mentioned in the follow-up emails by 2/12 players, where one stated wanting to play it again and to share all the games online (P1), and another discussed all the games with a classmate (P11).

Final decision

The methods adopted allowed for a consideration of domain relevance and potential to promote reflection (expert judging), gameplay experience and engagement with competition themes (play-testing and interviews) and longer term resonance (follow-up emails). In terms of the final decision, greater emphasis was placed on how the games impacted players; as evidenced by consideration of human error and related topics in both the post-play interviews and email responses.

On the basis of the evaluation, *Nurse's Dilemma* won first prize while *St. Error Hospital* was awarded runner-up. *Nurse's Dilemma* was most likely to have an immediate and longer term impact on players; where the game enabled empathy with nurses and an understanding of how a system can affect individuals. While *St. Error Hospital* was ambitious in scope, the complexity of the game meant that players were not always able to connect the gameplay to a consideration of the competition topics. At the prize-giving and showcase, *Nurse's Dilemma* was voted the People's Choice by the audience.

The evaluation also revealed that the judges and participants had their own preferences concerning which games they liked and what they got from them. Thus we decided to make all the games available on Errordiary (bit.ly/ErrorGames) to showcase the different ways in which the teams approached the competition challenge.

DISCUSSION

Despite recent interest in how games and technology can be used to promote empathy and encourage reflection, it is not clear how to evaluate different forms of serious experience. As the final component of the competition design process, we explored this issue when evaluating the impact of games created to raise awareness and lead to reflection on human error and blame culture within the context of healthcare.

In relation to Benford and colleagues work on uncomfortable interactions [2], the focus has been mostly on interactive, often public, performances rather than video games. Thus it is not entirely clear how to evaluate a potentially uncomfortable experience involving a single-player game played on a PC or console. As opposed to relying on expert analysis [1; 15], using only questionnaires with closed and open-ended questions [17] or an affective learning scale to assess attitudes [22], our evaluation consisted of a mix of expert judging, play-testing, and post-play assessment

This combination of evaluation methods allowed us to collect rich feedback and to investigate whether the expert opinions of the judges were reflected in the experiences of players. Similar points were raised by both groups, but the judges were able to consider whether the games presented an accurate interpretation of the competition topics, while the play-testing revealed the extent to which the game led to a consideration of those topics in practice. Given the sensitive nature of human error and blame culture within healthcare, where mistakes can lead to significant harm, the play-testing also allowed us to explore how the players reacted emotionally to each of the games.

Our approach provides further evidence that notions of fun are not necessarily applicable to considering games that involve "serious experience" [18]. Some of the participants had strong reactions to playing *Nurse's Dilemma* in particular, such as feeling sad or helpless, but it is precisely this negative emotional reaction that impacted on the player. In this case, uncomfortable experiences that made players think were more important than whether or not they thought the game was fun. Asking players to rate games and what they liked best would not have elicited the fact that while some negative experiences such as boredom should be avoided, others can lead to reflection on serious issues. The star ratings alone could not capture the qualitative differences between each game. The post-play interviews provided the most useful information for understanding the immediate impact of the gameplay, particularly in terms of the extent to which each game inspired curiosity and reflection on the competition themes.

In addition, the email questions were instrumental for considering longer term impacts such as the extent to which serious experiences actually resonated with players after the gameplay sessions. As argued by Marsh and Costello [18], if the aim is to raise awareness and get people thinking, then the evaluation needs to tap into whether a game leads to further thought or discussion about the game topics.

On the basis of Gaver et al., [9], Douglas & Wilson [24] suggest that prolonged engagement over time is one indicator of a game's success. While this may be true of a more complex game such as *St. Error Hospital* where there

are multiple variables to consider and multiple actions that can be taken, *Nurse's Dilemma* shows how a one-off play experience can have more impact through delivering a simple yet powerful message.

Limitations

One of the potential limitations of our study was the fact that, due to time constraints, follow-up emails were sent only two days after the gameplay sessions. While we did receive useful data from the participants, a longer wait would have allowed participants more time to think about their experiences and discuss the game with others.

While the majority of participants engaged with the play-testing and noted positives as well as negatives regarding the different games, one participant in particular struggled with the process:

P12: To be honest I found them quite boring and also probably because I don't really enjoy reading.
Interviewer: Yes, you've rated them all, I think, one star?
P12: Yes.
Interviewer: No, that's fair enough. Were they not what you were expecting?
P12: Yes, I don't know, maybe it's just that I prefer to have games that are more adventurous and more challenging rather than just like, I don't know...

This exchange highlights the fact that engagement normally starts off as a choice [4], and is influenced by multiple micro and macro level factors [12]. Regardless of subject matter, for those that expect to engage in more lightweight and familiar forms of gameplay and who aren't willing to revise their initial expectations, serious experiences will not lead to engagement, let alone further reflection.

Similarly, while the participants were told about the aim of the games prior to playing them, not everyone was familiar with the idea of using games for serious purposes. For instance, P2 noted "*I think this is a new kind of game because even though before we have seen some scary context like you explore in a dark room and you feel scared and something like that, but games on this topic, it's my first time*". There were further tensions expressed between player expectations of gameplay and the experience of playing persuasive games about serious issues. Even one of the judges raised questions about "*Is this a game or a story though? Can you lose or do you get points?*" (J3 on *Nurse's Dilemma*). Similarly, P11 noted in relation to the text-based entries (*Nurse's Dilemma* and *Medical Student Errors*) "*the two middle ones, they didn't really feel like games, they felt like I was going through one of those storybooks you had when you were a kid where you got to pick your ending*". The discussion of what makes a game is beyond the scope of this paper, save to say that the competition had a broad remit, but it would be worth exploring how people's expectations of what a game should be, influence their subsequent interpretations of gameplay.

Further research

In terms of game design mechanisms, *Nurse's Dilemma* suggests a short game with a simple message that is able to elicit an uncomfortable yet compelling experience through narrative, audio and simple graphics is more effective than pure text, compulsive gameplay or a complex simulation. *Nurse's Dilemma* is not a fun game, but through its negative emotional impact it is able to expose tensions in an underlying system and lead to reflection on normally taken for granted assumptions about responsibility and blame within the context of healthcare. Arguably, the information in the final scene acted as a debrief to participants, helping them to contextualize their experience and relate it to the real world. This process appears to be similar to the final stage of dénouement described by Benford et al [2] as it allows for experiences to be assimilated and reflected upon. Further research could investigate these mechanisms in more depth to understand how particular game elements are able to support different forms of serious experience that result from games and other forms of interaction.

Our evaluation approach could also be used for games that are focused on raising awareness and promoting reflection on other types of serious issues e.g. the environment, unemployment etc. A similar comparative methodology (involving domain experts; play-testing with target audience and follow-up assessments) could help select between games or prototype designs. Even when evaluating a single game, it would be important to include expert judging for assessing domain relevance; play-testing with post-play interviews for understanding the experience of play and how players engage with domain topics; and follow-up assessments for considering longer term resonance. While star ratings are relatively simplistic there may be more nuanced questionnaires that could help assess the impact of gameplay on players. The evaluation approach could also be adapted for longer games e.g. having several play-testing sessions and gaming diaries. Finally, the combination of methods may be useful for comparing and evaluating other forms of technology that result from reflective and critical design practices [e.g. 5].

CONCLUSION

Assessing the entries to a game design competition allowed us to explore how to evaluate serious experience in games. Through combining judging with play-testing we were able to assess domain relevance and whether expert opinion was reflected in player experience. While simple ratings were not found to be useful, asking players to rank the games in different ways led to a discussion that indicated the extent of engagement with the competition themes. In particular, the discussion enabled a consideration of the games in terms of gameplay and in terms of reflection on domain concepts. Finally, the use of post-play email questions was vital for establishing how the games resonated with players. We argue these methods will help designers and evaluators who wish to move towards serious experiences that aim to

promote reflection as part of a transformative learning process [10].

ACKNOWLEDGMENTS

We would like to thank all the teams that took part in the competition, the participants from the play-testing sessions and the expert judges. Special thanks to members of CHI+MED who took part in discussions about the competition. The authors are supported by the EPSRC funded CHI+MED project (EP/G059063/1).

REFERENCES

1. Belman, J., & Flanagan, M. Designing Games to Foster Empathy, *Cognitive Technology*, 14(2), (2010), 5-15.
2. Benford, S., Greenhalgh, C., Giannachi, G., Walker, B., Marshall, J, Rodden, T., Uncomfortable interactions, *Proc. CHI 2012*, ACM (2012), 2005-2014.
3. Bogost. I., Persuasive Games: The expressive power of video games. MIT Press, 2007.
4. D'Aprix, R., & Tyler, C. F. Four essential ingredients for transforming culture. Str*ategic Communication Management,* 10, (2006), 22-25.
5. Dunne, A.: Hertzian Tales: Electronic Products, Aesthetic Experience, and Critical Design.The MIT Press, 2006.
6. Dunwell, I., de Freitas, S., Petridis, P., Hendrix, M., Arnab, S., Lameras, P., and Stewart, C. A game-based learning approach to road safety: he code of Everand. P*roc. CHI 2014*, ACM Press (2014), 3389-3398.
7. Flanagan, M. *Critical Play: Radical Game Design.* MIT press, 2009.
8. Flanagan, M., & Nissenbaum, H. A game design methodology to incorporate social activist themes. P*roc. CHI 2007*, ACM Press (2007), 181-190.
9. Gaver, W., Bowers, J., Kerridge, T., Boucher, A., & Jarvis, N. Anatomy of a failure: how we knew when our design went wrong, and what we learned from it. P*roc. CHI 2009*, ACM Press (2009), 2213-2222.
10. Halbert, H., & Nathan, L.P. Designing for negative affect and critical reflection. *Ext. Abstracts CHI 2014*, ACM Press, (2014), 2569-2574.
11. Iacovides, I., & Cox, A.L. Designing persuasive games through competition. Paper presented at *Workshop on Persuasive Participatory Design for Serious Game Design: Truth and Lies*, at CHI Play 2014, Toronto, Canada, October 2014.
12. Iacovides, I., McAndrew, P., Scanlon, E., & Aczel, J.C. The gaming involvement and informal learning framework. *Simulation & Gaming*, Online First, (in press). http://sag.sagepub.com/content/early/2014/11/20/1046878114554191.full.pdf+html
13. Khaled, R., Fischer, R., Noble, J., Biddle R. A qualitative study of culture and persuasion in a smoking cessation game, *Proc of the 3rd International conference on Persuasive Technology*, (2008), 224 – 236.
14. Kirriemuir, J., & McFarlane, A. *Literature review in games and learning*. Futurelab series, Bristol: Futurelab, 2004. http://hal.archives-ouvertes.fr/docs/00/19/04/53/PDF/kirriemuir-j-2004-r8.pdf
15. Lee, S. "I lose, therefore I think": a search for contemplation amid wars of push-button glare. Game Studies, 3, (2003).
16. Lin, J.J., Mamykina, L., Lindtner, S., Delajoux, G., & Strub, H.B., "Fish'n'Steps: encouraging physical activity with an interactive computing game, *Proc. UbiComp 06*, (2006), 261-78.
17. Linehan, C., Kirman, B., Lawson, S. & Doughty, M. Blowtooth: pervasive gaming in unique and challenging environments. *Ext. Abstracts CHI 2010*, ACM Press, (2010), 2695-2704
18. Marsh, T., & Costello, B. Lingering Serious Experience as Trigger to Raise Awareness, Encourage Reflection and Change Behavior. *Persuasive Technology,* (2013), 116-124.
19. Mekler, E. D., Bopp, J. A., Tuch, A. N., & Opwis, K. (2014). A systematic review of quantitative studies on the enjoyment of digital entertainment games. In P*roc. CHI 2014*, ACM Press (2014), 927-936.
20. Montola, M. The positive negative experience in extreme role-playing. *Proc. of 1st Nordic DiGRA 2010*, DiGRA, (2010).
21. Nacke, L. E., Drachen, A. & Goebel, S. Methods for evaluating gameplay experience in a serious gaming context. *International Journal of Computer Science in Sport,* 9 (2), (2010).
22. Ruggiero, D. Spent: changing students' affective learning toward homelessness through persuasive video game play. P*roc. CHI 2014*, ACM Press (2014), 3423-3432.
23. Scott, M. & Wheeless, L. Communication apprehension, student attitudes, and levels of satisfaction. *Western Journal of Speech Communication*, 41, (1975), 188-198.
24. Wilson, D. & Sicart, M. (2010). Now it's personal: on abusive game design. *Proc. of FuturePlay 2010* , (2010), 64-71.