Savouring our mistakes: Learning from the FitQuest project

Savouring our mistakes: learning from the FitQuest project

Judy Robertson[1], Andrew Macvean[2], Samantha Fawkner[3], Graham Baker[4], Ruth G Jepson[5]

1. Digital Education Research Centre, University of Edinburgh, Holyrood Rd, Edinburgh, EH8 9JX, Scotland. Judy.Robertson@ed.ac.uk,
2. Computer Science, Heriot-Watt University, Earl Mountbatten Building, Riccarton Campus, Edinburgh, Scotland, EH14 4AS, andrew.macvean@gmail.com
3. Physical Activity Research for Health Research Centre, University of Edinburgh, Holyrood Rd, Edinburgh, EH8 9JX, Scotland. S.Fawkner@ed.ac.uk
4. Physical Activity Research for Health Research Centre, University of Edinburgh, Holyrood Rd, Edinburgh, EH8 9JX, Scotland. Graham.Baker@ed.ac.uk
5. Scottish Collaboration for Public Health Research and Policy, 20 West Richmond Street, Edinburgh, EH8 9DX, Scotland. Ruth.Jepson@ed.ac.uk

Abstract

Although serious games for children can potentially have important social, educational and health benefits, the research process from initial game design to a robust evaluation is lengthy and complex. This paper describes the design and evaluation process of an exergame for children. It reports on the inconclusive results of a cluster randomised controlled trial (RCT) conducted among children aged 10-11 years attending 10 state-funded primary schools in Scotland. One class in each school was randomly allocated to intervention (n=5, 111 children) or control (n=5, 104 children). Intervention schools were given FitQuest, a smartphone game for the Android platform, and were requested to play the game during at least one hour of mandated Physical Education (PE) lessons per week for 5 weeks. Participants in the control arm took part in standard mandated PE lessons. Primary outcome measures were step count, minutes spent in moderate to vigorous physical activity (MVPA) and exercise self-efficacy. None of the children spent the recommended time per week playing FitQuest. There were no significant differences in step count, MVPA or self-efficacy by intervention group.

The paper reflects on possible flaws during the design and evaluation process which could have led to the disappointing results, and presents some proposals for improving the research process for developing serious games for children. These include: deepening the ways in which we interact with domain expert colleagues, developing a shared understanding of the expectations for different phases of evaluation, closing the gap between game design knowledge and domain theories, raising

the standards of evidence for design guidelines, encouraging synthesis across studies by evaluating mid-range theories rather than individual games, and developing guidelines for monitoring intervention fidelity in this domain.

Highlights:

- Designing and evaluating serious games for children is time consuming and complex
- The design and evaluation process of FitQuest, an exergame for children is described
- Results of a cluster randomised controlled trial of FitQuest in 10 schools are reported
- There were no significant differences in physical activity or self-efficacy by intervention group.
- Reflections on the research process for developing and evaluating serious games for children are presented

1. Introduction

Although serious games for children can potentially have important social, educational and health benefits, the research process from initial game design to a robust evaluation is lengthy and complex. Human Computer Interaction (HCI) is an optimistic discipline, in which technological innovation is valued, and researchers are sincerely committed to applying technology to solve social problems[1]. However, as examined in a recent special issue of this journal entitled *Learning from failures in game design for children*, evaluation methodologies used in HCI and interaction design for children (IDC) lack depth [2] – studies often contain only a small number of users, do not employ rigorous methodology, focus on user preferences rather than educational or health outcomes, do not study the same system under repeated use and do not provide a longitudinal insight into how users interact over time. The last criticism is particularly relevant for serious games which aim to facilitate sustained change in social or health behaviours. The review also noted a reduction in the number of papers which reflected on the research process. The special issue editors call for studies which "*provide a deeper understanding of the complex process of the design of games*" [2;73].

The purpose of this paper is to provide insights to other researchers into the complex process of designing and evaluating a serious game for children. The paper describes the research process of an exergame for children which, following four years of user centred design work, was evaluated in a cluster randomised controlled trial with ten primary schools.

The results of the study are disappointing. The exergame (FitQuest) does not improve children's self-efficacy to exercise, nor increase their step counts after using it. The teachers did not include the game as part of their lessons for the length of time to which they initially agreed because of a variety of contextual reasons. However, we have decided not to succumb to gloom. Daniel Dennett advises: "*Try to acquire the weird practice of savoring your mistakes, delighting in uncovering the strange quirks that led you astray. Then, once you have sucked out all the goodness to be gained from having made them, you can cheerfully set them behind you, and go on*" [3;23]. With this in mind, this paper reports not only on the design and evaluation of FitQuest, but ends with the authors' reflections on the process and how we can learn from this in order to improve the methodologies we use for the development and evaluation of serious games.

2. Background: physical activity and exergames

Physical Activity Research

Children's participation in physical activity (PA) is important for their healthy growth and development [4]. Even modest amounts of PA can have health benefits for all, but particularly high-risk youngsters (e.g., those who are overweight or obese) [5]. Encouraging young people to adopt healthy PA habits can help to "prevent chronic conditions including coronary heart disease, stroke, type 2 diabetes, cancer, obesity, mental health problems and musculoskeletal conditions" [6].

A key goal of recent guidelines issued by the Chief Medical Officers of the United Kingdom (UK) is to increase the amount of regular PA undertaken by children [6]. Currently, the target of one hour of moderate to vigorous physical activity (MVPA) per day is often not achieved; for example, a recent study found that only 51% of English children aged between 7 and 8 years old meet this target [7]. UK guidance [4] suggests that the main facilitators for young people being physically active are: social and family influences; enjoyment; socialisation; and intrinsic and extrinsic rewards. A possible way to bring together several of these facilitators is to harness children's enthusiasm for video games. Young people are intrinsically motivated to play video games and spend up to 18 hours per week doing so (although there is variation in usage time according to age and gender) [8]. Exergames, (also known as active video games or AVGs) video games which use the player's bodily movements as input, may be a form of entertainment which encourages young people to be active [9].

Exergames research

Interaction design researchers have been designing and evaluating innovative exergames for around a decade, creating an admirable range of imaginative and engaging games by exploiting emerging technologies (see [10–16] for some particularly good examples). These games have often included children as part of the design process, and embody core values of the IDC research community [17] including social interaction, playfulness, exploration and equity of participation. The design of FitQuest was influenced by early guidelines emerging from the design of exergames and physical games on other research projects [12,21–26]. The specific guidelines, including for example supporting social influence, micro-goals, free play and marginal challenge, are discussed in detail in [27].

While many of the exergame systems developed during research projects have not have extensive evaluation, this section focusses on some high quality evaluations of technology to support physical activity for schools. The studies below were conducted in real world school settings, and involved relatively large numbers of participants or multiple sessions.

The Play Mate active game for children was evaluated in three schools for a single session, with 135 participants in the initial acceptance evaluation and a further 90 players in the evaluation of a revised adaptive version [18]. The results indicated that the children undertook more physical activity while playing the active version of the game, and this did not negatively affect their motivation to play.

In the StepStream project, researchers developed a pedometer based microblog to encourage school students to become more active through a social fitness approach [19]. In a four week study in a school with 42 participants, StepStream users improved their attitudes about fitness and the least active participants increased their daily activity. The study documents how the real life social behaviour (such as class meetings) motivated social usage online.

The American Horsepower Challenge (AHPC) project is notable for the large scale deployment in a real world setting over a year. In this study, a pedometer based video game was conducted in 61 schools across 14 states with a total of 1465 participants [20]. The system encouraged children to become more physically active by accumulating points for a school team based on the number of steps they take. Schools participated in a series of three four week long heats. In a mixed methods evaluation, the researchers documented the important role of the school environment and peer influence in changing the children's behaviour, as well as the crucial role of the teacher in facilitating the usage of the system. A particular challenge was maintaining the enthusiasm of children and teachers in the face of multiple technical glitches. The evaluation showed promising results in the sense that children reported changing their behaviour and making conscious decisions to walk more. Participating in the AHPC did encourage children to take more steps, although there was some novelty effect.

Studies of health outcomes relating to commercial exergames (rather than games produced by research projects) have been more common within the health sciences. Systematic reviews concluded that such games can enable light to moderate PA [28–31], but that the evidence for long-term efficacy is so far inconclusive [30]. Based on the evidence so far, exergames are recommended as an alternative to sedentary behaviour and as a complement to traditional physical activities [32]. Review studies have concluded that rigorously designed studies over longer time periods with better power and comparison of exergames to traditional PA activities are required, and such studies will need to keep up to date with evolution in exergame designs [29].

In summary, the emerging research literature of projects developed immediately prior to and in parallel with FitQuest showed promise in addressing the intractable problem of physical inactivity in schools, suggesting that it was a fruitful area for further research.

3. System Details

We designed and evaluated a location-based exergame called FitQuest to address the problem of physical inactivity in children. This section describes game which was evaluated in the cluster randomised controlled trial documented in Section 5. The design process which led to this game is described in Section 4.

The most recent version of the system runs on the Android operating system, utilizing GPS technology and Google Maps™ in order to provide a series of 8 exercise based mini-games. The initial game was designed and developed by the second author[1]. The software is a research prototype developed at Heriot-Watt University and is not currently commercially available. Figure 1 shows an example of one FitQuest mini-game.

Within the mini-games, the user is presented with a map focused upon their real-world location, with an avatar depicting their physical coordinates. The user must walk or run in the real world to interact with in-game objects and characters. By placing objects and characters at different distances around the character, and imposing various constraints, for example time limits, the players can be encouraged to walk and run. Although FitQuest is a suite of separate mini-games, consistent themes and an overarching points system tie the games together. A player can earn up to 10 points from

[1] Stuart Gray ported the IOS version to Android and carried out minor revisions. The Android version was used in the study reported in this paper.

each mini-game, which is accumulated as a session total, and running total over multiple sessions. The use of multiple mini-games was to intended to give the users choice and variety which we hoped would engage their interest over a sustained period. In addition, the games vary in the required physical intensity so that the children could match their game choices to their energy levels, and the short nature of the mini-games fit enable flexibility to fit around school timetabling constraints. The mini-games include *Collect the Coins* (in which the user must collect virtual objects which have been randomly generated on the map while evading a virtual wolf, figure 2), *Escape the Wolf* (the user must run away from a virtual predator), *Return the Sheep* (a shuttle run style game in which the user must repeatedly collect moving virtual objects and put them within the fixed bound of a virtual pen), *Visit the Fields* (the user must visit different virtual fields in a set sequence, requiring swift directional changes) and *Follow the Chicken* (a lower intensity chasing game where the user chases a virtual character rather than being pursued by one). In addition to these five games, there are three simple "mystery games" in which the user is challenged to run as fast as they can for 20, 30 or 40 seconds. Using an algorithm which consider past player performance and current mini-game difficulty, points are awarded in a way that encourages a level playing field between players of different abilities and fitness. All mini-games can be played and won irrespective of the PA background of the players. The algorithm does not compare the player's performance to that of their peers.

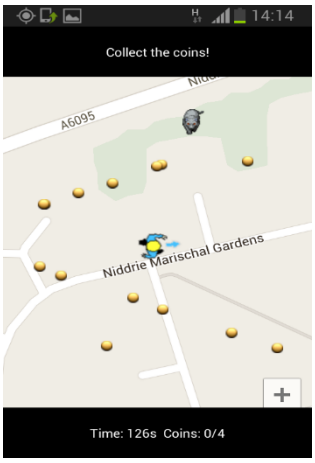

Figure 1. A screen shot from the Collect the Coins Game

The FitQuest system evaluated in this study has formalized support for goal setting, which is a well-documented behaviour change technique (BCT) within health sciences [33]. Users can choose from a range of goal types including achieving a custom set points target for the session, being top of the leader-board and completing all games on the most challenging setting.

The players could choose whether to opt-in to the leaderboard (a version of the social comparison technique for behaviour change [33]) so that their points would be shared and ranked with other class members, or whether their performance would remain private.

4. Design Process

FitQuest was designed by the second author (in conjunction with the first author) during his PhD project using an iterative user-centered design approach over a four year period [34]. The starting point for the project was that exergames applied to promoting physical activity in schools were showing some promise based on initial studies. Due to our interests in pervasive gaming [35], we proposed to use a location based approach. Throughout the process, we attempted to balance the design requirements emerging from the exergame and behaviour change literature, the constraints

of the school environment, the perspectives of the teachers, and the preferences of the young users. We took the CARSS approach to working with stakeholders in a school setting as discussed in detail in [36]; we believe that young people's participation in the design of technology can indeed provide important insights which would otherwise not be possible, but that the design process should be carefully managed to facilitate this. Ehn identifies two important features of the participatory design strategy: a) the political aspect of democracy by which users are empowered through contributing to design, and b) the technical aspect in which the participation of users results in more successful design and higher quality products [37]. In designing for children and young people there can be a trade-off between these two features. We believe that children and young people should be involved to some extent in the design of technology which they use for reasons of empowerment, but we also recognise that this does not automatically result in a higher quality design. We endeavour to be realistic about how time consuming user engagement in the design process can be. We acknowledge that there are circumstances under which the target users do not have enough domain knowledge to contribute useful input, and that in other cases there may be a mismatch between the priorities of a target user group and the aims of the system. For example, in the FitQuest project, while the PE teacher contributed background knowledge about how to develop fitness in young people it was less reasonable to expect this of young people themselves. In contrast, the young people would be better placed to comment on how playable or enjoyable they found different game designs. Potentially the young people might prioritise enjoyment of the game over how much PA is facilitated, which would be inconsistent with the research goals. For these reasons, we carried out a range of consultation activities with different stakeholders who assumed various roles.

In this section, we describe the main stages of the design process in which users participated, and note some limitations which are apparent in hindsight. Throughout this section, sentences which begin with "In retrospect" in bold typeface indicate our reflections on the process. The focus of this paper is not on the design requirements for this particular system, but rather the process for consulting users and integrating user requirements with the literature.

At the beginning of the user consultation process, the initial design idea was to develop a location based pervasive game to encourage 12-15 year olds to undertake physical activity within a formal school setting. A list of design requirements was developed with reference to the literature [27]; user consultation was then required to refine the requirements and develop a suitable game.

Initial user consultation

The user consultation began with an interview with a Physical Education (PE) teacher at a local high school who explained the challenges associated with encouraging high school children to be physically active, described the logistics of a PE lesson and suggested the sorts of physical movements which would be beneficial. The interview, and the researchers' observation of a traditional PE lesson, provided reassurance that the original proposed target audience (12-15 years) was appropriate, and that exergames are a potentially viable option for the PE classroom. The teacher also provided useful feedback on the school context, providing a set of initial design considerations for the exergame. On the advice of the PE teacher, the mini-games focussed on physical activity such as sprinting or agility. Simple game mechanics (collecting items / escaping a NPC) were chosen due to the simple translation from PA to game mechanic.

In order to build upon the initial considerations identified by the teacher, focus groups involving the target demographic were conducted. Two consecutive semi-structured focus groups were conducted with six 13-14 year olds (3 boys, 3 girls) in which participants used commercial Wii Fit

games, performed card sorting activities and discussed their preferences and requirements for exergames. Generally, feedback was positive towards exergames, with participants enjoying the experience and indicating surprise about the high standard of the games. It was important to the users that a game should be simple and thus easy to 'pick up and play'. We drew from this that the learning curve should not be too steep as this could result in demotivation. The 'pick up and play' nature was also important from the perspective that the game was be implemented in a school context which constrains the time available to play the game. The users liked the points system in the example games, in particular as a means for validating performance. However, despite being able to express preferences regarding existing exergames, the participants struggled to conceptualise game design ideas for a new mobile exergame.

Early Prototype Evaluation and Refinement

The findings of this consultation, in combination with design requirements for exergames from the literature, led to the development of a prototype. The prototype contained two example mini-games to illustrate game mechanics (collecting objects and running away from non-player characters) which would be further refined in subsequent iterations. This protoype was evaluated with a class of 25 participants aged 12-15 years old recruited from a local high school who played the game on a single occasion. Feedback from the young people and the PE teacher led to refinements and further development including changing the theme from "Pacman" style ghosts to an animal theme, the introduction of custom difficulty levels, and increased variety for game selection through additional mini-games.

The resulting second prototype was further evaluated with an additional eleven 12-15 year olds during a PE class. A PE teacher observed the session and commented on the intensity of the young people's PA in comparison to other lessons and on safety and logistical issues. Pre and post-test questionnaires, focus groups and log files provided insights into usability and user enjoyment. Overall the participants enjoyed the experience, with a mean enjoyment score of 7.1 (scale 1-10, sd =2.43), and all wished to play the game again.

An analysis of the log-files showed the average speed at which the participants played each of the mini-games (as calculated from phone distance and time data). While there are issues with the accuracy of distance data from phones, it allows for general inferences on how each mini-game facilitated physical activity. As is shown in Table 1, the four mini-games generally supported moderate and light intensity exercise, although both the Collect the Coins and Escape the Wolf game supported vigorous intensity exercise (as indicated by the range). It was positive to note that the mini-games supported a range of intensities, providing allowances for different levels of fitness and the current fatigue of the player. Additionally, the expert PE teacher stated during the post-evaluation interview that she was happy with the level of intensity observed during the session, and that based on her observations, the children would get enough physical activity whilst playing the game to justify its inclusion within a PE class.

Mini-Game	Average Speed (mph)	PA Intensity[2]	Range (mph)
Collect the Coins	3.108	Moderate	0.12 - 7.81
Escape the Wolf	4.519	Moderate	0.21 - 9.57
Visit the Fields	2.77	Light	1.72 - 3.82

[2] Based on the breakdown given by http://www.cdc.gov/nccdphp/dnpa/physical/pdf/PA Intensity table 2 1.pdf

Follow the Chicken	1.04	Light	0.07 - 2.62

Table 1. PA intensity measured estimated by speed (2nd prototype)

To this point, no safety concerns had arisen and user and expert consultation was positive. The evidence so far suggested that users enjoyed playing the game, and that it facilitated MVPA (according to observation by the PE teacher and speed/distance calculations from the phone). However, as we noted that it would be beneficial to encourage more vigorous intensity PA, we introduced three new mini-games in which the children were encouraged to improve their own personal bests in a series of sprint time trials.

In retrospect, it would have been prudent to gather objective accelerometer data to measure PA intensity and use this to optimize the mini-game designs before continuing with the evaluation. The intensity data inferred from distance and time calculated by the phone itself is indicative, but not accurate enough to enable fine grained analysis of the games.

Evaluation over a longer time frame

The next stage in the development process was to evaluate the game over a longer time period, as previous literature suggested a novelty effect might occur [38]. Two school based evaluations were conducted: a) a five week study (4 sessions) in a high school PE class with fourteen participants (9 girls, 5 boys) aged 14-15 years and b) a seven week study (12 sessions) in a primary school with twelve 11 year olds. An-depth qualitative case study analysis from both evaluations indicated promising results in terms of the children's enjoyment and PA during sessions, although there were some indications of a novelty effect after several sessions in the longer study. The older children rated FitQuest with a mean of 6.3 out of 10 for enjoyment, and the younger children 6.8 out of 10. At the end of the studies 10 of the 14 older children and 8 of the 12 younger children stated that they would like to play the game again. Across both settings, the average PA intensity (as measured by speed/distance data from the log files) varied across mini-games, but resulted in at least light, and in some cases, moderate intensity activity. As the GPS signal fluctuations cause inaccuracies in distance calculations, we also gathered objective Actigraph tri-axial accelerometer data from eight children during one FitQuest session in the primary school, with the assistance of the third author who is a pediatric physiologist (for analysis details see [27]). This data demonstrated that the participants participated in a range of physical activity intensities from light to vigorous, as shown in Table 2 .

Participant	*Sedentary*	*Light*	*Moderate*	*Vigorous*
1	15.7	49.4	10.1	24.7
2	6.1	26.8	11.0	56.1
3	12.4	72.2	6.2	9.3
4	20.0	63.8	5.0	11.3
5	11.1	69.4	6.9	12.5
6	14.8	47.7	13.6	23.9
7	15.6	56.7	13.3	14.4
8	15.5	69.1	7.3	8.2

Table 2. PA intensity breakdown: % of time spent in each intensity from Actigraph data

In retrospect, it would have been beneficial to refine the game design to reduce the proportion of time spent sedentary, as this accounted for up to 20% of the users' time. While it is appropriate for the children to rest between burst of high intensity PA, they also rested between the lower intensity games, spending a lot of time chatting to each other and selecting new games. Requiring the users to select a "running order" of mini-games to play in advance with pre-specified rest times between

them would have increased the pace of the game and reduced sedentary time. However, this could potentially have detracted from the social experience of playing the game and reduced enjoyment for children who were less fit initially. It was also not clear at this point that schools would find it difficult to devote much time to FitQuest, and so we did not anticipate a pressing need to reduce extraneous time spent in sedentary rather than higher intensity PA.

From comparisons of the contexts in which the two studies took place, it became clear that the flexible primary school environment was a more suitable context for FitQuest. The class teacher could decide to run sessions both in scheduled PE slots, and at her discretion at other times. Sessions could be of variable length to fit on with other commitments. At the high school, FitQuest use was constrained within the 50 minute lesson period for PE, of which a considerable proportion was taken up with changing and walking to the sports field. As a result, when continuing the work, we chose to focus on the younger age group (11 year olds) within a primary school, as there seemed to be more opportunity for flexibility around the frequency and duration of sessions.

Extensive qualitative case study data from the two studies drew on Bandura's theory of self-efficacy to identify interesting patterns of in-game behaviour which could be related to the case study participants' self-efficacy for PA. One key finding was the way in which the self-setting of informal in-game goals was an effective motivator of players. This was further developed for the final version of FitQuest evaluated in the present paper in which users may formally select between various goal types.

In order to understand more about the primary school setting, and behavioural change theory in relation to physical activity, we met with a local authority education advisor who specializes in PE in primary schools, and the fifth author who is a senior research fellow in public health. After the experts played the game, we discussed how best to implement goal setting. The education advisor recommended that it would be appropriate for children to set and monitor their own goals, as this was a key skill which teachers try to develop in PE classes, and the public health expert confirmed this was an important aspect of behavioural change. This input was integrated as part of goal setting feature game which was used in the study reported here.

In retrospect, the goal setting feature was less straightforward than we initially realized and it would have been useful to have evaluated and refined this in a pilot session before going further. On the other hand, some aspects of the goal setting feature can only be used over the course of multiple sessions and resources to run an additional such study were not available. Additionally, it would have been beneficial to plan more thorough training and discussion related to goal setting with FitQuest for teachers involved in the next stage of the research.

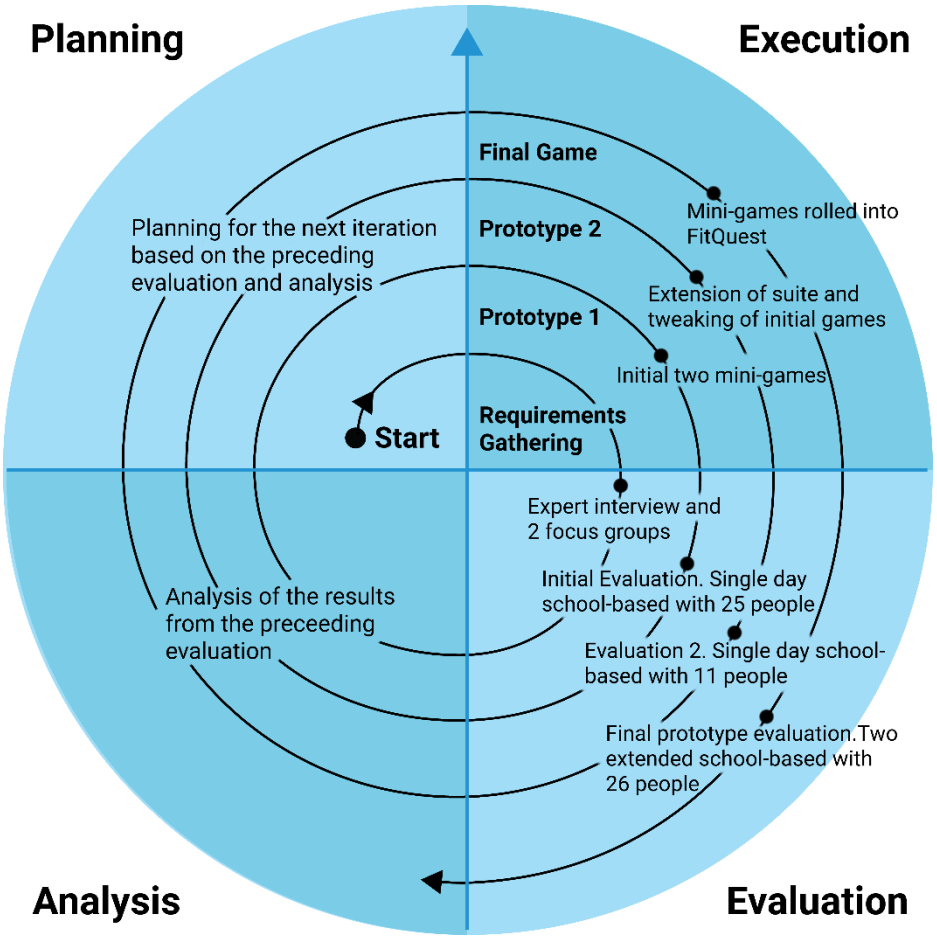

Figure 2. FitQuest design process

5. A cluster randomised controlled trial

After an extensive user centered design process for FitQuest, a more robust study was required. Previous studies had examined the effects of FitQuest in depth and over multiple weeks but it was necessary to increase generalizability by involving a larger number of schools and to introduce a comparison group. A cluster randomized controlled trial (RCT) design was chosen to avoid treatment group contamination and reduce classroom disruption hence randomisation was conducted at the school level.

The outcome variables were self-efficacy for PA and PA. Self-efficacy is a mediator variable which drives behavioural change, and so improvements in self-efficacy between pre and post-test would be encouraging. As Hekler et al write: "*if theory suggests that an application for encouraging physical activity works in part by strengthening self-efficacy, an evaluation that finds improved self-efficacy would provide preliminary evidence that the application is functioning as intended, even if the study is not able to detect behavioral changes due resource constraints on the study*." [1]. Changes in PA between pre and post-test for the intervention group would suggest that FitQuest had been successful in changing PA behavior even after the intervention was withdrawn.

The aim of this study was to evaluate whether a theory based location-based exergame (FitQuest) could increase self-efficacy and PA at school compared to standard provision in physical education (PE) classes. The research objectives were to: a) evaluate whether the intervention would lead to

increases in self-efficacy for PA compared to control and b) evaluate whether the intervention would lead to increases in PA at school (step count and time spent in MVPA) compared to control. The study did not aim to examine the potential impact of the game on physical activity patterns beyond the school setting.

This cluster randomised controlled trial[3] was performed between October 2013 and April 2014 in ten state funded primary schools within a city local authority in Scotland, UK. Complete details of the study design including the CONSORT flowchart and checklist for the trial can be made available in the supplementary materials for this paper.

5.1. Participants

Participants were 10 to 11 year olds from state funded primary schools. All of the schools are in suburban areas of the same city in Scotland.

School ID	SIMD quintile	Arm	Wave[4]	Boys	Girls	Total
1	2	FitQuest	2	9	9	18
2	2	FitQuest	1	13	14	27
4	5	FitQuest	1	14	10	24
7	4	FitQuest	2	9	12	21
10	5	FitQuest	2	10	11	21
3	2	Control	2	4	9	13
5	5	Control	1	5	11	16
6	5	Control	2	13	17	30
8	3	Control	1	9	9	18
9	2	Control	2	14	13	27

Table 3. Information about schools. SIMD (Scottish Index of Multiple Deprivation)

5.2. Sample size

Sample size was calculated according to Hutchison and Style's recommendation for cluster designs (Hutchison & Styles, 2010). An alpha value of .05 and power of 0.8, and intra-cluster correlations of 0.03 for self-efficacy and 0.018 for PA were assumed based on intra-cluster correlations reported in a recent study of children's PA (Salmon et al., 2011). A sample of 10 clusters, with 30 individuals in each cluster (which was possible given the resource constraints), gives a minimum detectable effect of 0.44 for self-efficacy relating to PA and 0.40 for PA. Both are interpreted by Cohen to be in the medium effect range (Cohen, 1992).

[3] Registered as ISRCTN11693550

[4] Wave 1 took place between October and December 2013. Wave 2 took place between January and April 2014.

5.3. Randomisation

After schools agreed to take part in the study, they were randomised to treatment by a staff member of the Scottish Collaboration for Public Health Research and Policy who was not involved in the project, using a computerised random number generator. No stratification was applied. Consent was sought from the local authority and head teachers before the randomisation and from parents and children afterwards. It was not possible to blind the participants to the intervention. Children and teachers were informed of the aims of the study. Data collectors were not blinded to the study aims and hypotheses as they were core members of the research team who assisted in the design of the study.

5.4. Ethics

The study was approved by the institution's Research Ethics Committee. Written permission to conduct a study in the schools was granted by the City Council prior to randomisation. Written informed consent was granted by at least one parent/guardian and all child participants after randomisation.

5.5. Delivery of the intervention in the school settings

The FitQuest project loaned each school allocated to the intervention group one suite of fifteen Samsung Galaxy Ace II phones for the duration of the project. Due to resource constraints, the children shared these phones with a partner, but each child had a personal user name and password to ensure that only their own performance data was recorded. The project paid for all data costs on the mobile phones. Each school has access to outdoor play space for FitQuest use in the form of a tarmac playground, a grassy field or an all-weather surface. All teachers involved in the project were briefed and given a demonstration of FitQuest in person before the project started. All teachers agreed to take part and were positive about the software.

Schools in the intervention arm used FitQuest for 5 weeks during PE lessons. Pupils in Scotland have a mandatory total of two hours of PE classes per week (usually split into two 1-hour lessons, as was the case in all the schools participating). Schools undertook to use FitQuest during both of these classes for the duration of the study. However, due to the limitation in the number of mobile devices, only half of class could use FitQuest at any given time. The one-hour lessons were split into two 30-minute play sessions. Each participant, therefore, played the game for approximately one hour per week within his or her PE class, over two 30-minute sessions. Mobile phones were handed out to the children at the beginning of the session and returned to the teacher at the end of the session. The children were given a demonstration of how to use the game on the first session, and researchers subsequently provided help to individual children when necessary. During the PE lesson, when the children were not using the FitQuest system, they participated in a traditional PE activity. This differed from session to session and was at the discretion of the PE teacher. The schools were invited to allow the children to play the game during play time and lunchtime but as this was at the discretion of the head teacher, only one school did so. Participants in the control arm took part in the normal PE classes provided by the school for 5 weeks.

5.6. Data gathering

Quantitative data

Pre and post intervention data (one week before the intervention and one week after the 5 week intervention period) was collected for all of the outcomes.

A paper version of the Pender questionnaire [39] on exercise self-efficacy was administered pre and post-test. These were completed during class time. Participants were asked to tick a box (very true, quite true, not very true, not true at all) which best represented their view on a series of eight statements (see Table 4). Following the instructions on the Pender questionnaire, responses were reverse coded (so that very true = 4 and not at all true = 1). The total score was computed by adding responses to all eight items. The highest possible score was 32.

I could exercise even if I was tired
I could exercise even I had other things I wanted to do
I could exercise even if I had to exercise on my own
I could exercise even if I had a bad day at school
I could exercise even if I was feeling lazy
I could exercise even if I was not very good at it
I could exercise even if I was sore from exercising the day before
I could exercise even if I was not in the mood

Table 4. Pender exercise self-efficacy statements

All participants wore an NL 1000 piezoelectric accelerometer (New Lifestyles Inc, Lee's Summit, Missouri, USA) which provides data on steps and MVPA. Accelerometers were set to record MVPA using the manufacturer default of 3.6 METs or above (level 4) during school hours for the first four days of the pre- and post-test weeks. Accelerometers were given to the pupils at the beginning of the school day and removed at the end of the school day.

Qualitative data

Qualitative data from observations and interviews with children and teachers was also collected in order to provide potential explanations for the quantitative results. A full analysis of this data can be found in [40], but the main findings are summarised here to give the reader further context for the reflections presented in this paper.

The purpose of the observations was to gain an overall impression of how the game was used in the playground setting, to document social interactions, and informally track participants' changes in attitudes particularly with respect to the potential novelty effect. At least one researcher from the data collection team observed all PE based usage of FitQuest. Interview data was also collect post-intervention during 20 minute semi-structured interviews with a total of 6 pairs of children (from two of the intervention schools–schools 2 and 7), and with three teachers. The children who took part in the interviews were selected using critical case sampling by the research team based on their session observations – the aim was to learn more about behaviours relating to goal setting and self-efficacy and hear from children with a range of views on the game.

6. Analysis

Separate analyses were conducted to examine the impact of treatment on the outcome measures of step count, MVPA and self-efficacy. A multilevel linear regression model with school as a level 2 variable and treatment (intervention or control) and pre-test scores as covariates was used in each case. Given that the participants were clustered within schools, initially null two-level (pupil>school) models were estimated in order to examine the proportion of the variation in each outcome attributable to differences between schools (the intraclass correlation coefficient, ICC) [41]. Single level models were also estimated in order to test whether accounting for school clustering significantly improved model fit [42]. Models which did and did not account for baseline (pre-test score) were tested, and adjusting for baseline was found to significantly improve the fit of the

models for all three outcomes (step count, MVPA and self-efficacy). All the analyses were undertaken in R [43]. The alpha level was set at .05 throughout (two sided). Details of the process used to address missing data are given in the supplementary materials.

Transcripts of interviews and researchers' notes from observations were analysed using thematic analysis using the software Dedoose. Details of the analysis method can be found in [40].

7. Results

Quantitative measures

The average time logged playing FitQuest for each school in the intervention arm is reported in Table 5. Descriptive statistics by school for step count, MVPA and self-efficacy are shown in tables Table 6, Table 7, and Table 8 respectively. Note that the recommended usage time was two, thirty minute sessions per week for 5 weeks (300 minutes).

		Mean time spent using FitQuest in minutes per child	Standard Deviation	Number of sessions	Proportion of overall recommended usage time
School ID	1	38	9	2	13%
	2	121	51	8	40 %
	4	125	15	6	42%
	7	151	20	8	50%
	10	73	20	5	24%

Table 5. Treatment fidelity (time spent using FitQuest in treatment schools)

It can be seen from Table 5 that none of the schools spent the recommended time using FitQuest; in the best case School 7 used it for 50% of the recommended time, whereas School 1 used it for only 13% of the time.

	Pre-test Mean	Pre-test SD	Post-test Mean	Post-test SD
Control (N=57)	24.9	4.4	24.8	4.1
FitQuest (N=79)	24.7	4.1	25.8	4.1

Table 6. Descriptive statistics for self-efficacy (total score on Pender scale). Note that 1 FitQuest and 2 control schools did not return self-efficacy data

Table 6 illustrates that the control and FitQuest groups had similar relatively high levels of self-efficacy at pre-test (around 25 out of a possible 32 points). Although the self-efficacy of the FitQuest

group increased slightly, this was not statistically significant as shown in the multilevel linear regression model.

	Pre-test Mean	Pre-test SD	Post-test Mean	Post-test SD
Control (N=87)	6287	1414	6081	1945
FitQuest (N=70)	6033	1809	5213	1692

Table 7. Descriptive statistics for step count.

Table 7 and Table 8 respectively show the descriptive statistics for step count and minutes spent in MVPA pre-test and post-test. Once again the pre-test step count and minutes are comparable between control and FitQuest groups. However, at post-test, the step count and minutes spent in MVPA is lower in the FitQuest group. It is worth noting that there was very poor weather during the post-test week for two of the schools which meant that the children were unable to play outside during some of the lunch and playtime activities. This will have had some impact on the time the children were able to be active which is unrelated to FitQuest participation.

	Pre-test Mean	Pre-test SD	Post-test Mean	Post-test SD
Control (N=87)	30.2	8.2	29.9	11.5
FitQuest (N=70)	29.9	10.8	24.4	9.6

Table 8. Descriptive statistics for minutes spent in MVPA

Undertaking likelihood ratio tests identified that accounting for school clustering improved the fit of the models for each outcome (Self-efficacy: X2 (2 d.f.)=11.5, $p < 0.01$; Step count: X2 (1 d.f.)=36.84, $p<0.01$, MVPA: X2 (2 d.f.)=37.866, $p<0.01$). There was *no significant effect* of treatment on self-efficacy, (b= 1.05, 95% CI [-1.08, 3.19], t(5) = 1.25, p = 0.26), step count (b= -715.78, 95% CI [-1957, 526] , t(7) = -1.34, p = 0.21), or time spent in MVPA (b= -4.96, 95% CI [-12.24, 2.32] , t(7) = -1.59, p = 0.155).

Summary of qualitative findings

Evidence from interviews and observations indicates that the children enjoyed playing the game, particularly in the early sessions. There was mixed evidence with respect to the novelty effect – some children became bored with the game after the initial few sessions (particularly in the school where the children used the game intensively of the course of the first week). This was not the case in all schools, as some children explained that they became more interested and motivated in the game as the sessions progressed, particularly after they mastered the game mechanics. Goal setting in general did had a positive effect on motivation as intended, and the children were able to adjust their goals in response to success or failure. However, the leader-board goal type created an over-competitive, demotivating dynamic in one school, and in fact this goal type, although initially popular, tended to be used less at all schools as the sessions progressed.

Interviews with the teachers indicated that they thought the children enjoyed the game and appropriately engaged with PA while playing it. However, they did have some scepticism towards the use of technology in a PE setting. They offered various suggestions for improving the game, particularly in terms of improving team work which one teacher considered to be an important aspect of the PE curriculum.

Various contextual barriers emerged during the study which prevented the children from using the game for the recommended time. A common reason for cancelled FitQuest sessions was poor weather because the GPS signal was not accurate enough for usage indoors in a sports hall, and at the time indoor location beacons were not available on the market. There are many pressures on the school timetable and teachers have limited time to engage in additional projects due to their commitment to core aspects of their jobs. On several occasions, FitQuest sessions were cancelled because the children had the opportunity to take part in other sporting or cultural events – this could be considered a loss to the research project but of benefit to the individuals. There was an unanticipated barrier to the acceptability of the project in the school environment which did not arise during the pilot projects: many of the schools had policies which banned or restricted the general usage of mobile phones in school. Although approval was given for the phones to be used in this specific project, it was thought by some that it would be inconsistent to allow phone usage in break times for the use of FitQuest only. The local authority area in which the study took place has subsequently invested in large scale iPad provision for learners; it is possible that an exergame implemented on school iPads would have been considered more acceptable and therefore may have been used more. A full description of the importance of context in evaluating this exergame, including a logic model documenting the theory of change can be found in [40].

8. Reflection on the FitQuest research process

The trial results showed no statistically significant impact of FitQuest on self-efficacy, step count or time spent in MVPA. While all the intervention schools did use FitQuest on at least two sessions, on average the schools used FitQuest for only 35% of the recommended time (103 minutes over 5 weeks). The reasons for this included poor weather but also motivational and contextual factors [40]. Therefore, it is not possible to draw conclusions from this study about the efficacy of the FitQuest intervention as originally designed. Concluding that this exergame intervention does not increase self-efficacy ,step count or MVPA from these results would be what Dobson and Cook [44]consider as a Type III error. It is not possible to distinguish between the possibilities that the results were a function of a) the inefficacy of using this particular exergame during PE or b) the failure of the intervention to be delivered as intended. In addition, as due to resource constraints we did not collect data on exercise intensity during either FitQuest or control sessions it is difficult to assess the relative efficacy of FitQuest to PE lessons.

As a reviewer of this paper pointed out, it is possible that an exergame was not a particularly useful or appropriate response to the reality of the situation, and that the research process would be incapable of learning this lesson. While we do not believe this to be the case, it is worth consideration. In the area of HCI for sustainability, Baumer and Silbmernman [45] argue that there are conditions under which the implication might be not to design a technology product at all: a) when the technological approach could be replaced by a low tech product, or no technology at all; b) when the technology causes more harm than good; and c) when the technology solves a version of the problem which is computational tractable rather than the problem itself. Certainly, we have come across examples of all three of these conditions within educational technology research in general, but the conditions do not clearly apply to FitQuest. With respect to a), the exergame

solution could be replaced by lower tech pedometers, as in the case with AHPC [20]. It could also be replaced by traditional playground games, or advances in physical education pedagogy. However, there are tradeoffs involved in all these solutions, and it is our job as interaction designers to understand them. There is reason to believe that the mechanics of the pervasive exergame in which players interact dynamically with game objects would have a different impact on user motivation than the gamification of steps using pedometers. Furthermore, the status quo of non-technological solutions has not prevented the global epidemic of physical inactivity [46].Physical activity researchers are turning to technology as part of wider attempts to solve the problem; interaction designers can and should assist. Considering b), there is no evidence that FitQuest caused harm from the objective PA data, the self-efficacy measure or from qualitative evidence from multiple sources. Moreover, the professionals working in the schools did not raise any concerns. Neither does issue c) apply in the case of FitQuest, where the real world problem is to increase physical activity and the game itself requires physical activity in order to work.

Would the research process we used be capable of identifying that an exergame was inappropriate for this setting? Yes: in our paper which documents the qualitative findings of this work with a realist evaluation methodology we identify contexts in which the exergame was not suitable or less effective, for example in schools which have strict rules against mobile phone usage in the playground or in schools where there is not an adult to champion its usage [40]. Our methodology did enable us to discover these negative cases (as well as positive cases). The methodology could also have provided evidence that the exergame was completely unsuitable: the local authority, head teachers, class teachers, parents and children could have all denied consent or withdrawn within the standard ethical research procedure we used. This did not happen. The comments from the multiple interviews with young people, children and teachers could have been uniformly negative, but they were not. They were positive for the most part, and the criticisms which were offered were useful to inform this design and designs in our future research.

We endeavoured to follow good research practices during the FitQuest project: an iterative user centred design process with children and teachers; applying game design guidelines from the literature; drawing on theories of behaviour change; evaluating in two pilot schools over a number of weeks before moving to a more robust trial design. In spite of this, we made some mistakes from which we hope others may learn, specifics of which were noted in the “in retrospect” statements in the design process section. Consideration of these specific issues led us to identify two themes in our missteps: balancing risks and resources and the involvement of users and experts. We also review progress within exergame research since we designed FitQuest.

Balancing risk and resources

One intention of user centred design is to reduce the risk that a technological product will not be suitable or effective for the target user group. In this project, the risk was not mitigated and the software did not address the problem to be solved in the school context. It could be argued that the misstep was not entirely related to the software itself, but also to an unrealistic expectation of how it might be used in the real world. In order to gain benefit from PA, children need to spend enough time exercising. Creating sufficient regular opportunities for PA in children’s lives (including at school) has proved to be a significant societal problem. From a certain point of view, the lack of treatment fidelity in this study is symptomatic of an underlying problem: PA opportunities are difficult to schedule and maintain in a crowded school curriculum. Technological innovation by itself was never going to completely solve such an intractable problem, even if the game had been perfect.

We are not, however, claiming that FitQuest is perfect. It does have design flaws, such as the way custom goal setting is implemented [40] and there are certainly opportunities to enhance the social play aspect. The question at issue here is whether it would have been possible reduce the risk of design flaws by making different decisions at each iteration of the process.

Consider user enjoyment of the game: after each prototype iteration, we reviewed users' numerical ratings of the game as well as their qualitative comments. For example, in a 5 week pilot study in the secondary school, the children made positive remarks such as "it was a good game and I enjoyed it" and "the games were fun to play and they keep you fit". The average numerical rating at this point was 6.3 on a scale between 1 and 10 (range = 5 - 8, sd =1.14). At the time, it seemed reasonable evidence that the game was enjoyable. But were there warning signs in the numerical rating? Were ratings high enough? Should we have interviewed the children who rated it less favourably to discover why and then refined it? Similarly, the PA intensity data was acceptable, but probably not optimal in that a shift from light to moderate intensity on more of the games would have produced more health benefits. Our approach has been to continue with the next stage of the design process unless there is strong evidence that a design is *not* suitable. Indeed, on other projects, we have rejected designs and started again when faced with evident user confusion and technological limitations. An alternative, more stringent approach could be to iterate and refine in early stages of prototype design until a certain numerical threshold on a user satisfaction scale has been met. For example, if there was a standard user satisfaction scale used across the IDC community, it would be possible to set a threshold using published benchmarks for similar software. Our impression is that such a numerical threshold approach would be unwelcome in the IDC community which (rightly) places considerable value on qualitative data and design judgement.

The involvement of users and experts

There is a general enthusiasm for involving users in the design process within the IDC community but it also acknowledged that user centred design is time consuming and resource intensive. Setting up a user study takes time and can be difficult to schedule, particularly when schools are involved. This is particularly true when trying to arrange appointments with busy teachers - in Yarosh et al's review of IDC papers, only 5% involved teachers in the design process [17] . There are a number of points during the FitQuest project where in hindsight it might have been prudent to have a) postponed the next user study until further refinement of the game was complete b) carried out an additional user study to ensure that refinements were suitable c) consulted with a wider range of experts/stakeholders or d) carried out more formal objective assessments of our outcome variable. The reasons for not doing so related to lack of budget to buy enough research grade accelerometers or to pay to extend the contract of researchers to do more field work, or the lack of availability of schools to reschedule or introduce new sessions. We suspect that other IDC researchers have also confronted the problem of deciding how best to spend a limited budget for user consultation. How many users should be consulted and how often? Which expert groups are relevant to the domain, and is it possible to gain access to representatives of these groups? There is also the ethical question of how much time it is reasonable to ask of participants to commit to a project which is not necessarily core to their educational or professional goals.

We question in hindsight whether it was necessary to run focus groups with young people during early requirements gathering phases of FitQuest. As Davis et al. point out, when designing games, "*Focus groups can be useful for concept generation in the initial stages of a project or for obtaining a better general understanding of a problem space in some circumstances. However, they are poor at providing specific, actionable data*" [45]. While the focus groups were useful practice for working with an adolescent audience, and for highlighting some of the behavioural and social issues inherent

in the demographic, it was left primarily to the developers to design new content and establish the themes and content for the early prototypes.

It is also worth noting that over the course of the project we uncovered a range of users' opinions about the game. Not all the children enjoyed or were motivated by the same features. This highlights that while small groups of pilot users may give encouraging feedback on a design, this is not a guarantee that the design will be well received by all members of the target user group in a real life setting.

Throughout, we consulted with secondary school PE teachers, a primary school PE specialist and various academic experts in physical activity and public health. There was also an opportunity to discuss the game with the teachers who were involved in the intervention study. Perhaps the mistake we made here was to interview only single experts in the early stages. Focus groups of experts would have given us a variety of opinions, a higher chance of identifying potential problems and perhaps a more realistic expectation of the time which schools may be able to devote to such activities. However, in the early stages of the project we did not yet have the network of contacts or social capital necessary to form a focus group.

Stimulated by review comments of this paper, we have also reflected on the overlap between contributions from user centred design and contributions from previous research in other disciplines. A reviewer suggested that it would have been beneficial to engage in open discussion and more participatory co-design activities with the young people and their teachers about their barriers to inactivity; the criticism is that our decision to design an exergame before listening to young people was premature. The reviewer advocated that power should have been "given to participants to establish what the actual problems are that limit children in participating in exercise in school." Had we done this, the design of FitQuest would no doubt be different, but it would also have been fulfilling a different research agenda. The problem space of physical inactivity in children and adolescents has been extensively explored by physical activity researchers previously, including qualitative studies with young people [47–49]. At the time the project started, physical activity researchers and HCI researchers alike were exploring the potential benefits of exergames as one point in a wide solution space. Our design process started from the aim to further research exergames in the school context, and the involvement of young people and teachers was therefore focussed on their views about exergames rather than their wider experiences of barriers to physical activity. There is clearly a spectrum of views on the role and purpose of user involvement in HCI; our approach has been to respect both the time of the participants and the existing literature by constraining the problem before starting user consultation.

The approach taken by future designers of serious games will be guided by their personal philosophies and beliefs about participatory design, but we recommend that the maturity of the literature in related fields should also be a consideration. In the light of Marshall and Linehan's findings with respect to the misinterpretations of healthcare research by exergame researchers, it would be prudent to collaborate closely with public health researchers when examining the literature (see recommendation 1) [50]. Open ended, more exploratory co-design work which includes problem finding and establishing the design constraints may be suitable for contexts which are under-researched.

Game design

Scholarship in game design has moved on since 2009 when the FitQuest project started: for example, Rigby and Ryan's "Player experience of needs and satisfactions" model, which was published in 2011 [51], has informed our subsequent work in serious game design for children. In

this model, the concept of "fun" can be explained through three strands of intrinsic engagement: competence, autonomy, and relatedness. FitQuest does promote these strands to some extent: *competence* is supported by the progression of difficulty of the games which is connected to the previous performance of the player, the FitQuest design guideline of free play [27] promotes *autonomy*, and the leader-board supports some aspects of *relatedness* although it is clear there is room for improvement here. Specifically related to exergames, Marshall and colleagues recently proposed that design strategies which emphasise the richness of experience from sports participation and interactive entertainment would be helpful, and that exergames need not be limited to simply promoting healthy outcomes like increased energy expenditure. They detail strategies by which exergames could take into account how exertion changes of time, consider the pain of exercise, and embrace highly social interactions [52]. This work may be of interest to future designers of games for children where the focus is on the sheer physical enjoyment of sport and activity for its own sake. Indeed, the physical education teachers involved in the FitQuest work shared the view that sport is inherently worth learning irrespective of health benefits. This is a different part of the design space of exertion games from the original FitQuest project; both research strands can thrive within the creative and interdisciplinary community of interaction design.

Methodologies in games user research have also been developing [53], embracing a distinctive set of approaches to user engagement in collaboration with industry such as calibrated questionnaires and automatic video analysis through face recognition. Such developments are potentially beneficial for the designers of serious games in the future.

The quality of the literature review and underlying assumptions of exergame research within HCI have been heavily criticised by Marshall and Linehan [54], who scrutinised unwarranted claims about the link between obesity, physical inactivity and exergames in published papers. Their analysis is based around citations of an influential paper in health research which has been consistently misrepresented in literature reviews. As well as suggesting that exergaming should be focussed on areas where it can be realistically useful and advocating longer term studies, they advise authors to develop their understanding of health care research more deeply. We agree and suggest that can be achieved through deeper collaboration with domain experts (see recommendation 1).

9. Reflections on the design and evaluation process of serious games

Beyond the specific flaws in the study described here, we present our wider reflections on how the field in general could improve research processes for developing serious games for children.

1. *Deeper collaboration between interaction design researchers and domain experts.* During the FitQuest project, the HCI researchers' perspectives on the scope and role of technology in public health interventions has shifted. Becoming associate members of an established research centre in physical activity for health gradually revealed to us the extent and complexity of physical inactivity and how it plays out in social and physical environments. In the beginning of the design process, we were straightforwardly optimistic about the promise of an exergame to increase PA. Now, after years of learning with our PA colleagues, we have a deeper appreciation that by itself, an exergame could never be the whole solution to physical inactivity, particularly if only used within the school setting. We see FitQuest as part of a wider ecosystem of solutions which are being developed within public health such as encouraging active commuting to school; making streets and public spaces safer for play; structuring the school day to incorporate less sitting and more physical activity; and mandating more PE time within the curriculum.

We have learned that fruitful interdisciplinary collaboration involves learning new terminology, challenging the weaknesses and embracing the strengths of the partner discipline(s), and integrating methodologies from each discipline while seeking the highest standard of evidence. Interaction designers typically have strengths in participatory design and user centred design techniques [55] and knowledge of how to gather requirements from stakeholder groups [56] . Teams of health researchers are very knowledgeable about the development of traditional interventions, and have expertise in designing robust complex evaluations for real world settings [57] with the statistical knowledge required for appropriate quantitative analysis. One difficulty which we encountered was confusing terminology between public health and HCI about the different phases in the evaluation process. Indeed, the "development" phase in the influential Medical Research Council (MRC) guidance framework for the evaluation of complex interventions does not cover the complexities required in software development.

2. *Develop a shared understanding of the expectations for different phases of evaluation.* Klasnja and colleagues [58] make the point that reviewer expectations might be too high when evaluating whether technology designed to facilitate behavourial change is effective because behavioural change is a complex process which unfolds over a long time period, and is influenced by a series of internal and external factors which are unrelated to technology. They suggest that "HCI contributions should focus on efficacy evaluations that are tailored to the specific behavior-change intervention strategies … embodied in the system and studies that help gain a deep understanding of people's experiences with the technology" [p3063]. We agree with this point; indeed such an analysis of the FitQuest strategies and user motivations may be found in [40]. Klasjna and colleagues believe that HCI researchers should work with healthcare researchers to conduct more robust evaluations in the longer term, although they argue that the resource requirements make such studies prohibitive in the early stages. But what should the process be for moving from initial focussed efficacy studies to more robust and larger scale effectiveness evaluations? While the process for HCI evaluation and evaluations of healthcare interventions are separately well understood, how these processes relate to each other is underexplored. Developing a framework for complex technological interventions which integrates a user centred process with the MRC evaluation framework would be a beneficial first step for future collaborations between game designers and health researchers.
3. *Integrate knowledge of game design techniques with domain-level theories*. Studies reported in health journals are often of commercial exergames, and lack game design details (see for example [59] or [31]); there is a tendency to treat an "exergame" as a black box. The taxonomy of behaviour change [33] is a very useful starting point for communication about the active ingredients in a serious game for behaviour change, but there is a need to collectively catalogue and study the ways in which these techniques could and should be incorporated as design elements in a technological intervention. While attempts have been made to catalogue *which* BCT are used in apps [60], the details of the ways in which they are implemented are at least as important. Within serious games for education, a framework which maps game mechanics to learning mechanics has already been developed [61]; similar efforts could be made for other domains of serious games.
4. *Challenge and evaluate design knowledge*. The game design literature in HCI contains many design recommendations or guidelines which are based on the development and evaluation of a prototype with a single user study. Indeed, FitQuest was partly based on the amalgamation of such design guidance. Designs are also influenced during the user centred design process, but there is little attempt to establish the extent to which the preferences of a small sample of users might be a useful guide to a design which can have beneficial effects to a larger population groups. In short, the prevailing standards of evidence are not high. What is required is a way for

HCI researchers to document, synthesise, challenge and evaluate design knowledge gained over a series of studies. This would result in a more coherent body of interaction design knowledge which could be the foundation for future serious games work. We acknowledge that some interaction design researchers or practitioners may disagree with this recommendation from a philosophical perspective, perhaps because it represents a cultural bias towards 'scientism'[62]. One means of sharing research and design knowledge which avoids this bias is *research through design* in which the designed artefact itself, or an annotated design portfolio, is the means for sharing learning with design practitioners [62,63].

5. *Shift the emphasis from evaluating systems to evaluating mid-range theories*. Understanding *why* an intervention works (or does not work) in particular user groups is necessary for designers to improve system design, as Klasjna and colleague point out in their critique of traditional RCT designs in HCI behavioural change research[64]. We suggest that the theoretical paradigm of realist evaluation is appropriate here. Realist evaluation [65] responds to the messiness of the real world by acknowledging that it is highly likely that different groups will have different reactions to interventions, and attempts to document this so that mid-range theories (e.g. relating to behavioural change) may be synthesised from findings over a range of studies. Rather than focussing on "what works?" (as would be the aim of a traditional RCT), the realist aim is to discover "what works for whom, and under what circumstance and why?". This does not necessarily mean that we give up on the rigour of designs such as RCTs, but it does require that we focus on systematically documenting and comparing the context of interventions [66]. For example, in serious games for behavioural change, an active design ingredient (referred to as a behavioural change strategy by Kasjna and colleagues) may be social comparison. It would be useful if individual research teams reported the way in which social comparison techniques were designed into the game, along with details of the intended user group and the context of use. This would facilitate the integration of findings between studies so that we could collectively establish the design approaches which are likely to work under particular sets of circumstances.
6. *Develop approaches to monitoring intervention fidelity appropriate to the use of serious games with children*. If serious game research within IDC moves towards larger, more rigorous longitudinal studies with focus on evaluating the design intentions against real world outcome measures as advocated by [2], it would be beneficial to develop guidelines to help IDC researchers monitor and enhance the fidelity of an intervention. There are frameworks available to guide health behaviour researchers plan and evaluate trials [67], although such guidance is lacking within K-12 educational studies [68]. A synthesis of previous guidance from relevant disciplines, adapted to suit the specifics of technological interventions would be helpful.

7. Conclusions

A cluster RCT in 10 primary schools found that the FitQuest exergame was not successful in increasing 10-11 year olds' self-efficacy, post-test step counts or MVPA. No adverse events or important negative unintended effects were found during the study.

The lengthy design and evaluation process of FitQuest highlights the complexity and challenges involved in designing games for behaviour change in real world settings. Given the potential social benefits of such technology, we recommend that as a community we persist and overcome these challenges by deepening the ways in which we interact with domain expert colleagues, developing a shared understanding of the expectations for different phases of evaluation, closing the gap between game design knowledge and domain theories, raising the standards of evidence for design guidelines, encouraging synthesis across studies by evaluating mid-range theories rather than

individual games, and developing guidelines for monitoring end enhancing intervention fidelity of serious games for children evaluations.

8. Acknowledgments

The authors would like to thank all the schools and teachers who took part, Jan McIntyre of Edinburgh City Council, Stuart Gray and Andrew Williams for assistance with this work. This study was funded by an EPSRC Impact Acceleration Account scheme at Heriot-Watt University, Scotland (F12R10074).

9. References

[1] E.B. Hekler, P. Klasnja, J.E. Froehlich, M.P. Buman, Mind the Theoretical Gap: Interpreting, Using, and Developing Behavioral Theory in HCI Research, Proc. CHI 2013. (2013) 3307–3316. doi:10.1145/2470654.2466452.

[2] C. Moser, M. Tscheligi, B. Zaman, V. Vanden Abeele, L. Geurts, M. Vandewaetere, P. Markopoulos, P. Wyeth, Editorial: Learning from failures in game design for children, Int. J. Child-Computer Interact. 2 (2014) 73–75. doi:10.1016/j.ijcci.2014.10.001.

[3] D. Dennet, Intuition Pumps and Other Tools for Thinking, Reprint ed, Allen Lane, n.d.

[4] National Institute for Health and Care Excellence (NICE), Promoting physical activity for children and young people [PH17], NICE, 2009. https://www.nice.org.uk/guidance/ph17 (accessed February 16, 2015).

[5] I. Janssen, A.G. Leblanc, Systematic review of the health benefits of physical activity and fitness in school-aged children and youth., Int. J. Behav. Nutr. Phys. Act. 7 (2010) 40. doi:10.1186/1479-5868-7-40.

[6] Department of Health Physical Activity and Health Improvement, Start Active , Stay Active: A report on physical activity for health from the four home countries' Chief Medical Officers, London, 2011. http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_128209.

[7] L.J. Griffiths, M. Cortina-Borja, F. Sera, T. Pouliou, M. Geraci, C. Rich, T.J. Cole, C. Law, H. Joshi, A.R. Ness, S. a Jebb, C. Dezateux, How active are our children? Findings from the Millennium Cohort Study., BMJ Open. 3 (2013) e002893. doi:10.1136/bmjopen-2013-002893.

[8] B.S. Greenberg, J. Sherry, K. Lachlan, K. Lucas, A. Holmstrom, Orientations to Video Games Among Gender and Age Groups, Simul. Gaming. 41 (2010) 238–259. doi:10.1177/1046878108319930.

[9] M.O. Lwin, S. Malik, The efficacy of exergames-incorporated physical education lessons in influencing drivers of physical activity: A comparison of children and pre-adolescents, Psychol. Sport Exerc. 13 (2012) 756–760. doi:10.1016/j.psychsport.2012.04.013.

[10] S. Berkovsky, J. Freyne, M. Coombe, Recommender algorithms in activity motivating games, in: RecSys 10, ACM Press, Barcelona, 2010: p. 175. doi:10.1145/1864708.1864742.

[11] Y. Xu, E.S. Poole, A.D. Miller, E. Eiriksdottir, R. Catrambone, E.D. Mynatt, Designing pervasive health games for sustainability, adaptability and sociability, FDG '12 Proc. Int. Conf. Found. Digit. Games. (2012) 49. doi:10.1145/2282338.2282352.

[12] S. Munson, S. Consolvo, Exploring Goal-setting, Rewards, Self-monitoring, and Sharing to Motivate Physical Activity, in: Int. Conf. Pervasive Comput. Technol. Healthc., IEEE, 2012: pp. 25–32. http://www.scopus.com/inward/record.url?eid=2-s2.0-84865047997&partnerID=tZOtx3y1.

[13] K. Stanley, I. Livingston, A. Bandurka, Gemini: A Pervasive Accumulated Context Exergame, in: 10th Int. Conf. Entertain. Comput., 2011: pp. 65–76. http://www.springerlink.com/index/GX20773282604Q71.pdf (accessed May 24, 2012).

[14] I. Soute, P. Markopoulos, R. Magielse, Head Up Games: Combining the best of both worlds by merging traditional and digital play, Pers. Ubiquitous Comput. 14 (2010) 435–444. doi:10.1007/s00779-009-0265-0.

[15] F. Buttussi, L. Chittaro, R. Ranon, A. Verona, Adaptation of Graphics and Gameplay in Fitness Games by Exploiting Motion and Physiological Sensors, in: A. Butz (Ed.), LNCS 4569, Springer-Verlag, Berlin, 2007: pp. 85–96.

[16] H.A. Hernandez, T.C.N. Graham, D. Fehlings, L. Switzer, Z. Ye, Q. Bellay, M.A. Hamza, C. Savery, T. Stach, Design of an Exergaming Station for Children with Cerebral Palsy, CHI '12 Proc. 2012 ACM Annu. Conf. Hum. Factors Comput. Syst. (2012) 2619–2628. doi:10.1145/2207676.2208652.

[17] S. Yarosh, I. Radu, S. Hunter, E. Rosenbaum, Examining Values : An Analysis of Nine Years of IDC, in: T. Moher, C. Quintana, S. Price (Eds.), Proc. 10th Int. Conf. Interact. Des. Child., Ann Arbor, MI, 2011: pp. 136–144.

[18] S. Berkovsky, J. Freyne, M. Coombe, Physical activity motivating games, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 6206 LNAI (2010) 283–284. doi:10.1007/978-3-642-17080-5_30.

[19] A.D. Miller, E.D. Mynatt, StepStream, Proc. 32nd Annu. ACM Conf. Hum. Factors Comput. Syst. - CHI '14. (2014) 2823–2832. doi:10.1145/2556288.2557190.

[20] E. Eiriksdottir, Y. Xu, A. Miller, E. Poole, R. Catrombone, D. Kestranek, E. Mynatt, Assessing Health Games in Secondary Schools : Technical Report. An Investigation of the American Horsepower Challenge 2009--2010, Atlanta, 2011. http://smartech.gatech.edu/jspui/bitstream/1853/42173/1/AHPC_tech-report_v01.pdf.

[21] S. Consolvo, K. Everitt, I. Smith, J. a. Landay, Design requirements for technologies that encourage physical activity, Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI '06. (2006) 457–466. doi:10.1145/1124772.1124840.

[22] T. Campbell, B. Ngo, J. Fogarty, Game design principles in everyday fitness applications, Proc. 2008 ACM Conf. Comput. Support. Coop. Work. San Diego, (2008) 249–252. doi:10.1145/1460563.1460603.

[23] K. Suhonen, H. Väätäjä, T. Virtanen, Seriously fun – Exploring how to combine promoting health awareness and engaging gameplay, MindTrek'08. (2008) 18–22. doi:10.1145/1457199.1457204.

[24] S.M. Arteaga, V.M. González, S. Kurniawan, R. a. Benavides, Mobile games and design requirements to increase teenagers' physical activity, Pervasive Mob. Comput. 8 (2012) 900–908. doi:10.1016/j.pmcj.2012.08.002.

[25] J. Yim, T.C.N. Graham, Using games to increase exercise motivation, in: Proc. 2007 Conf. Futur. Play - Futur. Play '07, ACM Press, 2007: pp. 166–173. http://dl.acm.org/citation.cfm?id=1328202.1328232.

[26] T. Bekker, J. Sturm, B. Eggen, Designing playful interactions for social interaction and physical play, Pers. Ubiquitous Comput. 14 (2010) 385–396. doi:10.1007/s00779-009-0264-1.

[27] A. Macvean, Understanding the Exergame User Experience: Users' Motivation, Attitude and Behaviour in a Location-Aware Pervasive Exergame for Adolescent Children, Heriot-Watt University, 2013. http://www.ros.hw.ac.uk/handle/10399/2825.

[28] A.J. Daley, Can exergaming contribute to improving physical activity levels and health outcomes in children?, Pediatrics. 124 (2009) 763–71. doi:10.1542/peds.2008-2357.

[29] A.G. LeBlanc, J.-P. Chaput, A. McFarlane, R.C. Colley, D. Thivel, S.J.H. Biddle, R. Maddison, S.T. Leatherdale, M.S. Tremblay, Active video games and health indicators in children and youth: a systematic review., PLoS One. 8 (2013) e65351. doi:10.1371/journal.pone.0065351.

[30] E. Biddiss, J. Irwin, Active Video Games to Promote Physical Activity in Children and Youth, ARCH PEDIATR MED. 164 (2010) 664–672.

[31] W. Peng, J.C. Crouse, J.-H. Lin, Using Active Video Games for Physical Activity Promotion: A Systematic Review of the Current State of Research., Health Educ. Behav. (2012). doi:10.1177/1090198112444956.

[32] Z. Gao, S. Chen, D. Pasco, Z. Pope, A meta-analysis of active video games on health outcomes among children and adolescents, Obes. Rev. (2015) n/a-n/a. doi:10.1111/obr.12287.

[33] C. Abraham, S. Michie, A taxonomy of behavior change techniques used in interventions., Health Psychol. 27 (2008) 379–387. doi:10.1037/0278-6133.27.3.379.

[34] A. Macvean, Understanding the Exergame User Experience: Users' Motivation, Attitude and Behaviour in a Location-Aware Pervasive Exergame for Adolescent Children, Heriot-Watt University, 2013.

[35] M. Montola, Exploring the edge of the magic circle: Defining pervasive games, Proc. DAC. 1966 (2005) 16–19. doi:10.1.1.125.8421.

[36] A.M. Judy Robertson , Judith Good , Katy Howland, Issues and Methods for Involving Young People in Design, in: J.U. Rosemary Luckin , Sadhana Puntambekar , Peter Goodyear , Barbara Grabowski, N. Winters (Eds.), Handb. Des. Educ. Technol., Routledge, 2013.

[37] P. Ehn, Scandinavian design: on participation and skill, in: D. Schuler, A. Namioka (Eds.), Particip. Des. Princ. Pract., Lawrence Erlbaum, Hillsdale, N.J., 1993: pp. 41–77.

[38] J. Lin, L. Mamykina, S. Lindtner, Fish'n'steps: Encouraging physical activity with an interactive computer game, in: LNCS 4206, Springer-Verlag, Berlin, 2006: pp. 261–278. http://www.springerlink.com/index/vtw615807853xn0v.pdf (accessed July 14, 2012).

[39] N. Pender, C. Murdaugh, M. Parsons, The health promotion model, in: Heal. Promot. Nurs. Pract., 4th ed., Prentice Hall, Upper Saddle River, N.J., 2002: pp. 59–79.

[40] J. Robertson, R. Jepson, A. Macvean, S. Gray, Understanding the Importance of Context: A Qualitative Study of a Location-Based Exergame to Enhance School Childrens Physical Activity, PLoS One. 11 (2016) 1–15. doi:10.1371/journal.pone.0160927.

[41] P.D. Bliese, Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis, in: K.J. Klein, S.W. Kozlowski (Eds.), Multilevel Theory Res. Methods

Organ., Jossey-Bass, 2000: pp. 349–381. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Within-group+agreement,+non-independence,+and+reliability:+Implications+for+data+aggregation+and+analysis#0.

[42] A. Field, J. Miles, Z. Field, Discovering Statistics Using R, Sage Publications, London, UK, 2012.

[43] R Development Core Team, R: a language and environment for statistical computing, (2011). http://www.r-project.org/.

[44] D. Dobson, T. Cook, Avoiding type III error in program evaluation: Results from a field experiment, Eval. Program Plann. 3 (1980) 269–276. http://www.sciencedirect.com/science/article/pii/0149718980900427 (accessed September 17, 2014).

[45] E.P.S. Baumer, M.S. Silberman, When the implication is not to design (technology), Proc. 2011 Annu. Conf. Hum. Factors Comput. Syst. - CHI '11. (2011) 2271. doi:10.1145/1978942.1979275.

[46] S.N. Blair, Physical inactivity: the biggest public health problem of the 21st century., Br. J. Sports Med. 43 (2009) 1–2. http://www.ncbi.nlm.nih.gov/pubmed/19136507.

[47] M.M. Casey, R.M. Eime, W.R. Payne, J.T. Harvey, Using a socioecological approach to examine participation in sport and physical activity among rural adolescent girls, Qual. Health Res. 19 (2009) 881–893. doi:10.1177/1049732309338198.

[48] M. Dobbins, K. DeCorby, P. Robeson, H. Husson, D. Tirilis, Cochrane review: School-based physical activity programs for promoting physical activity and fitness in children and adolescents aged 6-18, Evidence-Based Child Heal. A Cochrane Rev. J. 4 (2009) 1452–1561. doi:10.1002/ebch.461.

[49] M.J. Babic, P.J. Morgan, R.C. Plotnikoff, C. Lonsdale, R.L. White, D.R. Lubans, Physical Activity and Physical Self-Concept in Youth: Systematic Review and Meta-Analysis, Sport. Med. 44 (2014) 1589–1601. doi:10.1007/s40279-014-0229-z.

[50] E. Prepared, Digimap Evaluation : Impact , Barriers , and Teaching Practice, 2017.

[51] S. Rigby, R.. Ryan, Glued to games: How video games draw us in and hold us spellbound, Praeger, Santa Barbara, CA, 2011.

[52] J. Marshall, F. Mueller, S. Benford, S. Pijnappel, Expanding exertion gaming, Int. J. Hum. Comput. Stud. 90 (2016) 1–13.

[53] P. Mirza-Babaei, V. Zammitto, J. Niesenhaus, M. Sangin, L. Nacke, Games user research, CHI '13 Ext. Abstr. Hum. Factors Comput. Syst. - CHI EA '13. (2013) 3219. doi:10.1145/2468356.2479651.

[54] J. Marshall, C. Linehan, Misrepresentation of Health Research in Exertion Games Literature, Proc. 2017 CHI Conf. Hum. Factors Comput. Syst. - CHI '17. (2017) 4899–4910. doi:10.1145/3025453.3025691.

[55] R. Khaled, A. Vasalou, Bridging serious games and participatory design, Int. J. Child-Computer Interact. 2 (2014) 93–100. doi:10.1016/j.ijcci.2014.03.001.

[56] O. De Troyer, E. Janssens, Supporting the requirement analysis phase for the development of serious games for children, Int. J. Child-Computer Interact. 2 (2014) 76–84. doi:10.1016/j.ijcci.2014.05.001.

[57] P. Craig, P. Dieppe, S. Macintyre, P. Health, S. Unit, S. Michie, I. Nazareth, M. Petticrew, Developing and evaluating complex interventions: the new Medical Research Council guidance, Br. Med. J. 337 (2008) 1655. doi:10.1136/bmj.a1655.

[58] P. Klasnja, S. Consolvo, W. Pratt, How to evaluate technologies for health behavior change in HCI research, in: Proceeding CHI '11 Proc. 29th Int. Conf. Hum. Factors Comput. Syst., 2011. http://dl.acm.org/citation.cfm?id=1979396 (accessed October 6, 2014).

[59] D. Schoene, S.R. Lord, K. Delbaere, C. Severino, T.A. Davies, S.T. Smith, A Randomized Controlled Pilot Study of Home-Based Step Training in Older People Using Videogame Technology, PLoS One. 8 (2013) e57734. http://dx.doi.org/10.1371%2Fjournal.pone.0057734.

[60] L.T. Cowan, S. a Van Wagenen, B. a Brown, R.J. Hedin, Y. Seino-Stephan, P.C. Hall, J.H. West, Apps of steel: are exercise apps providing consumers with realistic expectations?: a content analysis of exercise apps for presence of behavior change theory., Health Educ. Behav. 40 (2013) 133–9. doi:10.1177/1090198112452126.

[61] S. Arnab, T. Lim, M.B. Carvalho, F. Bellotti, S. De Freitas, S. Louchart, N. Suttie, R. Berta, A. De Gloria, Mapping learning and game mechanics for serious games analysis, Br. J. Educ. Technol. 46 (2015) 391–411. doi:10.1111/bjet.12113.

[62] W. Gaver, What should we expect from research through design?, Proc. 2012 ACM Annu. Conf. Hum. Factors Comput. Syst. - CHI '12. (2012) 937. doi:10.1145/2207676.2208538.

[63] J. Zimmerman, J. Forlizzi, S. Evenson, Research Through Design as a Method for Interaction Design Research in HCI design research in HCI, (2007) 493–502.

[64] P. Klasnja, S. Consolvo, W. Pratt, How to Evaluate Technologies for Health Behavior Change in HCI Research, in: CHI 2011, Vancouver, 2011.

[65] R. Pawson, The Science of Evaluation : A Realist Manifesto, 1st ed., SAGE Publications Ltd, London, 2013.

[66] C. Bonell, A. Fletcher, M. Morton, Realist randomised controlled trials: a new approach to evaluating complex public health interventions, Soc. Sci. …. 75 (2012) 2299–2306. http://www.sciencedirect.com/science/article/pii/S0277953612006399 (accessed October 7, 2014).

[67] B. Borrelli, D. Sepinwall, D. Ernst, A.J. Bellg, S. Czajkowski, R. Breger, C. DeFrancesco, C. Levesque, D.L. Sharp, G. Ogedegbe, B. Resnick, D. Orwig, A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research., J. Consult. Clin. Psychol. 73 (2005) 852–860. doi:10.1037/0022-006X.73.5.852.

[68] C.L. O'Donnell, Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research., Rev. Educ. Res. 78 (2008) 33–84. doi:10.3102/0034654307313793.