Co-created Evaluation: identifying how games support police learning

Anne Adams , Jennefer Hart , Ioanna Iacovides , Sian Beavers , Manuel Olivera , Maria Margoudi

PII: S1071-5819(19)30035-7
DOI: https://doi.org/10.1016/j.ijhcs.2019.03.009
Reference: YIJHC 2304

To appear in: *International Journal of Human-Computer Studies*

Received date: 17 February 2018
Revised date: 19 March 2019
Accepted date: 25 March 2019

Please cite this article as: Anne Adams , Jennefer Hart , Ioanna Iacovides , Sian Beavers , Manuel Olivera , Maria Margoudi , Co-created Evaluation: identifying how games support police learning, *International Journal of Human-Computer Studies* (2019), doi: https://doi.org/10.1016/j.ijhcs.2019.03.009

HIGHLIGHTS

- Method to co-create knowledge based evaluation questionnaires
- Compared games against face to face learning for 116 police learners in 3 forces
- Games significantly increased understanding over face to face learning
- Pinpointed type of tacit knowledge games increased and face to face decreased
- Co-created evaluation more valid and relevant for practice application

Co-created Evaluation: identifying how games support police learning

Anne Adams (a*), Jennefer Hart (a), Ioanna Iacovides (b), Sian Beavers (a), Manuel Olivera (c), Maria Margoudi (c)

(a) Institute of Educational Technology, Open University, Milton Keynes, UK
(b) Department of Computer Science, University of York, York, UK
(c) High Skillz, London, UK

Abstract

HCI often produces improved systems through co-creation with practitioners. However, evaluation methods are primarily researcher-led (i.e. not co-created with practitioners). As part of a games-based learning evaluation, we detail a novel co-creation method that produces evaluations on how technology influences learning. Based upon educational threshold concept theories, the Tricky Topic method supported the co-creation of knowledge-based evaluation questionnaires with trainers. The evaluation involved 116 new recruit police officers from three UK police forces who participated in a randomized-control trial. The Tricky Topic method provided insights of how the game significantly increased understanding $p<.001$ (moderate effect size) in comparison with face-to-face training. Tricky topic breakdowns identify increased tacit understanding (e.g. empathy, attention) after games training, and decreased tacit understanding (e.g. respect) after face-to face training. Finally, further research opportunities are discussed concerning co-created evaluation for valid and relevant deconstruction of participants' understanding that allow designers to pinpoint systems-specific learning benefits.

KEYWORDS: EVALUATION; SERIOUS GAMES; TRICKY TOPICS; POLICE TRAINING.

1. Introduction

The domain of HCI has developed nuanced approaches to design and development that aims to include users in the design of technologies and systems. Participatory design (Muller and Kuhn, 1993; Muller, 2003) and co-design (Sanders and Stappers, 2014) approaches have been a cornerstone of HCI's relevance to practice. However, evaluation approaches have not fundamentally changed or moved in line with these approaches. They primarily remain devised and controlled by researchers. Pragmatically, this has been an essential approach for HCI to take as researchers have expertise in evaluation. Some researchers have attempted to address issues of validity within researcher-led evaluation approaches by adopting mixed method approaches to data collection, such as using qualitative and ethnographic data collection methods. However, even when utilizing mixed methods, these evaluations are often created by the researchers who designed the systems being evaluated, creating an inherent subjectivity. We argue that the power of user-centred design and co-created systems also provides HCI with a unique opportunity to advance evaluation approaches. Within practice-based learning systems, this would be particularly valuable since a valid assessment of learning progression is often held within the practice context by trainers and teachers. Nevertheless, within this context, co-created evaluation has been under explored. We argue that educational research into co-created Tricky Topic evaluations (Adams & Clough, 2015), based upon pedagogical threshold concept theories (Meyer and Land, 2006), provides a unique opportunity to test the applicability of this approach to evaluating learning technologies such as serious games.

Serious games, i.e. those that are used for non-entertainment and predominantly educational purposes, have increased in popularity over recent years and have

been advocated as promising technologies for the support of training within sectors such as the military, education, the emergency services, education and healthcare (Susi, et al, 2007., Williams-bell, et al. 2015). An evaluation of a serious game does not only require a consideration of the player experience but also the extent to which the game 'works' i.e. in the case of an educational game, is it able to effectively support learning? While there have been attempts to make the design of serious games more participatory through including stakeholders in the process (e.g. Lukosch, et al. 2012., Khaled, and Vasalou, 2014., Alenljung and Söderholm, 2015) there are very few co-created evaluations that validate if learning has occurred. A potential barrier to doing so may relate to the complexity of examining learning progression, particularly in practice-based learning contexts (such as police training) that involve both tacit and procedural knowledge. However, whilst research-led learning progression is difficult to define and measure, co-creation provides a valuable opportunity to gain valid contextual measures of learning. It is therefore important that this contextual approach to evaluation is utilized to benefit the quality of learning evaluations. Without a comprehensive understanding of how games can support practice-based learning in comparison to other forms of training, there is a significant risk that serious games cannot be claimed to be successful in achieving their aims.

In this paper, we describe a co-created approach that we developed to evaluate the effectiveness of a serious game compared to the current traditional face-to-face learning, within the context of training new police officers. Although not the initial goal of the research, we identified the value of adopting a co-created approach for learning evaluations. The Tricky Topic 'co-created' evaluation method that we employed is the focus of this paper, with the police training context (where games are compared to face-to-face training) serving as an example of its usefulness in practice.

The serious game that was developed focused specifically on the learning points and barriers to learning as part of training in taking initial witness accounts from children. Through presenting the results of the evaluation study, we aim to provide an in-depth exploration of game-based learning for police training and to examine the role of co-created evaluations. We use a 'Tricky Topic process' (detailed below) as a co-created evaluation method to identify barriers to learning, which were then used to test for knowledge acquired from the game.

2. Relevant work

Within this section, we first provide a review of the background to co-created evaluation approaches in HCI for learning, and the learning sciences through to serious games. Next, we present the educational theory underpinning Tricky Topics and its pedagogical relevance for evaluating deeper understanding. Finally, we present the relevance of serious games as a test bed for this co-creation evaluation approach.

2.1 Co-created evaluations for game-based learning

Within HCI and related disciplines like Technology Enhanced Learning there has been a movement towards participatory design (Muller and Kuhn, 1993., Muller, 2003). From research by Carroll (1996) to work by Muller (2003), and Halskov & Hansen (2015) there has been a development of methods, techniques and practices that support a collaboration between the practitioner/ stakeholders and the developer/researcher. In particular, a focus has been placed on the connection spaces between those participating in what has been termed the 'third space' (Muller, 2003). Through facilitating this 'third space', which is neither owned nor directed by only the researcher or the practitioner stakeholder, the goals of both parties could be supported. By working together, a broader understanding of what counts as 'evidence' can be developed and applied to the evaluation process. In particular, practitioners have an understanding of context whilst researchers have an understanding of rigorous evaluations of learning. Within the learning sciences there has been a history of using participatory research approaches to increase the relevance and accuracy of evaluations (Chappell, 2000; Seale, 2010). These approaches go beyond co-creating evaluation tools into co-creating research questions and co-evaluating results. Stoeker (1999) argued almost 2 decades ago that the changing role of academics in participatory research, with participants becoming active researchers and evaluators, produces the question 'are academics irrelevant?' Within this paper we are not taking this stance. However, we would argue that HCI, especially with regard to learning technologies, should not be behind the learning curve in debating and applying these novel approaches to learning technology evaluations.

To identify the level of co-created evaluations (defined by Seale, 2010) in HCI game-based learning an initial, yet not exhaustive, review was completed. Each paper's evaluation methods were analysed and compared with regard to: 1) evaluation focus, 2) epistemological design, 3) methodological drivers and 4) the level of researcher and participant input (as outlined by Seale, 2010). Co-creation was defined and operationalised at the most basic level of simply engaging with participants or stakeholders to create or review the evaluation approach before it was applied. An initial review was completed with 45 games-related papers. The papers where drawn from an initial search of all Transactions on Computer-Human Interaction games articles from 2000–2016. This

resulted in 21 papers that were evaluated against the criterion for co-creation (engaging with participant or stakeholders to create or review the evaluation). The analysis was further developed with a random sample of 24 CHI gaming and gaming related papers from 2007–2016. From reviewing a total of 45 papers, only one paper (Costabile. et al. 2008) was identified that used a co-created evaluation approach. The paper used multiple methods including an approach co-created with teachers to develop one of the learning measures used. Two other papers were identified that used a simplistic approach to engaging participants specifically for two different types of co-created research method roles; training participants to implement the systems because of their expertise (Giusti, et al., 2011) or using participants as evaluators as part of a crowdsourcing process (Dontcheva, et al., 2014). However, the evaluation tools employed in these, including surveys and interviews, still appeared to have been developed solely by the researchers.

In the example that did explicitly refer to an evaluation approach that was developed with teachers, Costabile et al. (2008) reported on "Explore!" a mobile game employed in an archaeological park. Teachers co-created one part of the evaluation process in a knowledge test for the game-based learning system. Whilst a clearly invaluable approach, this project only worked with teachers on the evaluation without documenting further educational expert input or theoretical educational underpinning to the evaluation methods. Also, the lack of a pre-test meant that student's initial knowledge was not taken into consideration. However, these limitations are understandable considering the lack of rigorous methods that support effective co-creation of learning evaluations.

The review exercise identified that although the predominant evaluation approach was experimental, there were variations depending on the research questions so the papers included ethnographic and field studies. The papers covered a range of data collection methods from biometric to observational, log and test analysis and questionnaires. The papers also took very varied and creative approaches to the design process and often many of these were participatory. However, within the evaluation design all the presented approaches, apart from one, used evaluation tools that were designed or selected by the researchers. The level of participatory engagement in developing the evaluation methods was therefore extremely limited. This highlights the need for increasing co-created methods to advance evaluation approaches, in the same way that participatory methods have advanced system design approaches.

2.2 Tricky Topics

One possible way to facilitate the co-creation of evaluations concerned with learning could be to use a process based on the concept of 'Tricky Topics' (Adams et al 2018). Tricky Topics are a practice-based application of the theory 'Threshold Concepts' (TCs). The pedagogical theory for TCs (Meyer and Land, 1993., Meyer and Land, 2006) has become a focal point for understanding conceptual barriers learners may encounter when gaining a deeper understanding of a concept. In particular, barriers to understanding TCs have been related to liminality, where the learning process of overcoming troublesome concepts and thus internalizing the understanding, is considered a learning journey rather than an outcome. TCs were originally identified in two founding papers (Meyer and Land, 1993., Meyer and Land, 2006) as a 'portal' to a different way of thinking through internalization of concepts without which the learner finds it difficult to progress (Meyer and Land, 2006). The criteria for Threshold Concepts are that they may be: Transformative (once understood they alter the perception of a subject), Irreversible (unlikely to be unlearned), Integrative (reveal connections in a subject), Bounded (help define a subject), Troublesome (maybe counterintuitive and beyond common sense). They are said to be more than just "key" or "core" concepts (Harlow, et al. 2011., Lucas and Mladenovic, 2007., Adams and Clough 2015), and form the starting point for transformative learning (Meyer and Land, 2006). The barriers presented by TCs can be so great, they may cause a learner to fail or give up on a subject altogether and research has highlighted the need to focus on effective methods for teaching them (Machiocha, 2014).

Although not without their critics, (Rowbottom, 2007., O'Donnell, 2010), TCs have been noted as valuable. In particular through identifying gaps in how academic communities understand teaching and pedagogy in practice (Machiocha, 2014). One key impact of Threshold Concepts is its shift away from learning outcomes and its pedagogical emphasis on the learner and their barriers to understanding within the learning journey. This aligns surprisingly well with HCI's focus on usability. It could be argued, similarly to HCI, it is not simply the end product that is important (learning outcomes) but the interaction journey to that point. Threshold Concepts focus on the learner as HCI focuses upon the user, or in gaming terms the player, or in participatory terms the stakeholder / practitioners and their learning journey. Meyer and Land (Meyer and Land, 2006) argue that designing learning to focus upon the journey can be transformative for the learner. It can then be argued that designing learning systems to focus upon threshold concepts could increase their ability to transform the user. However, we must ensure we are effectively evaluating deep learning, through transforming understanding, rather than purely supporting memorization of facts that are forgotten tomorrow (Adams & Clough, 2015).

Unfortunately, TCs have been poorly related and applied to teaching practices. Tricky Topics are a development of the educational theories underpinning TCs and applied into teaching practice (Adams & Clough, 2015). The Tricky Topic approach has been co-created with teachers and educationalists, in schools, HE and practice-based contexts to identify effective application for this approach (see http://tricky-topics-guide.ac.uk). The process has been broken down into three stages (Identify, Capture and Assess) and a set of terms (stumbling blocks, problem examples, problem distiller) that support the deconstruction of Tricky Topics for practice-based contexts. However, whilst learning design and assessment approaches have developed through the application of Tricky Topics, this has yet to transfer into evaluation approaches.

2.3 Understanding in practice-based serious games

Evaluation is especially important in particular practice-based contexts. In safety-critical domains (e.g. healthcare, emergency services, crisis management), the potential consequences of providing inadequate training could result in significant physical or psychological harm being caused to people in the real-world. However, it is frequently not the accumulation of facts that these systems seek to develop, but effective application of knowledge in practice. Though there are some exceptions (Toups, et al. 2011), the majority of safety-critical games attempt to provide concrete representations of practice due to a desire to provide realistic learning experiences within a 'safe' space (Williams-Bell, et al. 2015) and to increase the chances of transferring learning outside of the game (Whitton, 2014). Game elements are usually used to convey different forms of information to players as a way of indicating progress and providing feedback on performance. These feedback mechanisms then provide a very specific understanding of 'learning points' within a game which is often at odds with pedagogical approaches to learning. For example, a learning point could be a simplistic piece of information that learners have surface understanding of, memorised to get through the game and then forgotten after the game is completed. Games research has not clarified this through pedagogically evaluating the relationship between feedback mechanisms and progression in deep conceptual learning that transforms the users' understanding.

For instance, Di Loreto, Mora & Divitini (2012) provide an overview of serious games for crisis management, highlighting the different examples of how games have been used in this context. However, while the literature suggests that factors such as the level of realism, which is usually interpreted as graphical fidelity (e.g. Linssen, et al. 2015., Toups, et al., 2011, Williams-Bell, et al. 2015) and feedback (e.g. Crookall, 2010., Haferkamp, et al, 2011) are important, it is pedagogically unclear how they relate to developing a depth of understanding nor how they progress insights into how practitioners can support evaluating effective learning for practice.

There are limited examples of games for police training. One is a 3D traffic accident training scenario (Binsubaih, et al., 2006) created for the Dubai Police force. While they did find significant learning effects between those who used the game and the control group (who did not play the game) via the use of pre and post-tests, the depth of police officers' conceptual understanding was very limited in its deconstruction. Researchers led the evaluation by focusing on 'presence as related to learning' by adapting an existing questionnaire (Slater, 1999) measuring subjective 'being there' experiences. (p.340; Binsubaih, et. al. 2006). However, beyond comparing novices with experts, it is unclear why this measure was chosen and its relevance to the police. This research also did not compare the game with other training methods making it hard to establish specific game-based learning value for the police.

Linssen et al. (2015), present another example, of Loiter (LOItering Teenagers, an Emergent Role-play) a game that focused on training Dutch police officers in the interpersonal skills required for street interventions. The game supported learning social interaction skills (including verbal responses and physical stance). The game provided feedback to players in the form of "thought bubbles" (that represent how game characters are reacting to the player) and flashbacks relating to previous actions. Whilst these were creative measures, the pedagogical underpinning for them is not clear with the evaluation identifying that they did not lead to improvement in learning measures. The evaluation measures (rating learning experiences on a likert scale) were researcher-led (not co-created) and poorly linked to conceptual understanding for the police.

These studies indicate that a games learning evaluation and its co-creation is often not given a significant amount of attention in the context of practice-based serious games. Questions remain about how to evaluate learning for serious games and what role practitioners can play in developing these evaluations. In order to further explore these issues, we present a co-created learning evaluation of a serious game for training new UK police officers in taking an initial account from a child witness.

3. Evaluation trial background

The Child Interview Simulator (CIS) was co-developed with the police as a serious game to support the training of UK police recruits in collecting initial witness accounts from children. In addition to conceptual understanding, the trainees developed a confidence for interacting with

children, which previously was only refined through experience. The CIS provides an interactive scenario where one takes on the role of an officer that needs to interview a nine-year-old boy, who allegedly witnessed a woman being attacked on his way home from school. The first episode requires the trainee to take an 'initial response' account from the child at their home whilst the second episode requires the trainee to conduct a full ABE (Achieving Best Evidence) interview for the purposes of gathering evidence.

Figure 1 illustrates a screenshot from the final game displaying the following interface elements: two parallel horizontal bars representing the Tricky Topic rapport bar (indicated by a green feedback bar that moves up or down depending on the players' interactions with the child and parent) and several procedural and exploratory icons (ABE form, Notepad). When entering conversation mode, a menu partially covers the scene, enabling players to ask questions about various topics via text-based multiple-choice options. Within the scene, various interactive objects are highlighted with a white border, inviting players to observe these by clicking on them and thus unlocking further dialogue options in conversation mode.

Figure 1. Screenshot of the kitchen scene (episode one)

Police training is traditionally based on experiential knowledge (HMIC 2015, HMCPSI/HMIC 2014). In this project we extended this to include gaming and pedagogical literature and processes.

Co-design approaches were included within the game's development and the learning evaluation. Designing and evaluating effective deeper learning is critical since the game must lead to embedded understanding for the police, and ultimately support them in life and death decision making. To achieve effective real-world understanding the research and design team had to work closely with the police to develop a game and evaluation criterion that supported and assessed the learners in developing both tacit and procedural knowledge. Within police force practice, procedural knowledge, is closely connected to police procedures (e.g. note taking), whilst tacit is closely connected to implicit understanding of a situation (e.g. responding to body language).

A co-design process developed a story structure, which consisted of two distinct episodes (see Figure 2). While further details of this design process can be found in Margoudi et al (2016), the co-design involved a multidisciplinary team of experts in; child interviewing, police trainers, experienced police officers, game developers and academic researchers. An agile development approach was adopted, with iterative versions of the game that used storyboards, interactive mock-ups and subsequent software prototypes until the final version was produced. Evaluation approaches began early in this process where they were designed according to 'problems' (linked to Tricky Topics) that police learners encountered in nationally standardised face-to-face training sessions (benchmarked by the Ofqual qualifications credit framework and aligned to College of Policing standards for interviews and initial incident response), which served as an effective reference condition. The co-design process not only shaped the design but was also found to support co-evaluation approaches (i.e. the Tricky Topic process). For example, whilst initial research questions were established to fulfil funding obligations, a second complimentary set of research questions were established that fulfilled practitioners' needs e.g. rapport building

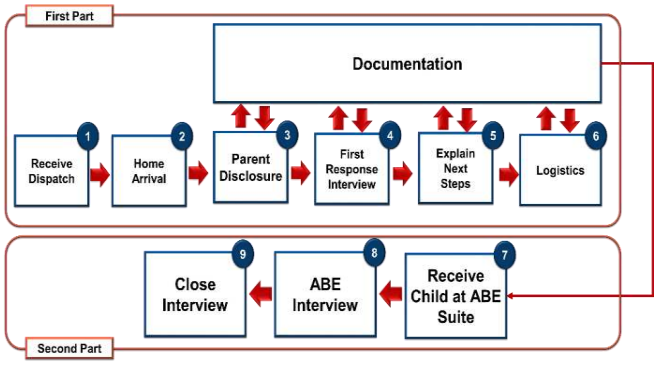

Figure 2. Story structure overview

During the design process, the whole direction for the game-based training project changed. Initially the project's objective was academically driven by research into training the police in tactics for securing a location. However, through discussions with the police it became clear that these current practice needs were being met, as situational role-play appropriately covered these training areas. Collaborations and discussions within the police domain identified specific gaps in the types of training needed with a number of areas being highlighted, such as taking first accounts from a child. Thus, the focus of the game changed. In the following section, we highlight our

approach to co-creating an evaluation for the final version of CIS.

3.1 Tricky Topic process

The stages in the Tricky Topic process are outlined in detail on: http://tricky-topics-guide.ac.uk and within a free badged open course (Adams et al. 2018). The Tricky Topic process can be used to help design learning points and facilitate outcomes that are learner centred. As such, they would be developed in parallel to any participatory or agile learning design process. They can also be used to support the development of co-created evaluations of learning. In the case of the current study, they were used only to develop the knowledge-based tests.

There are three main stages in the Tricky Topic process (see http://tricky-topics-guide.ac.uk, Adams et al. 2018); identify, capture and assess. The key target of Tricky Topics is to focus on barriers to learning i.e. 'problem examples' rather than learning outcomes. Within the initial 'identify' stage of the process there are three main concepts that need to be identified in conjunction with practitioners:

1) Tricky Topics: Specific topics containing challenging concepts that learners find difficult to grasp, and teachers and trainers find difficult to teach.
2) Stumbling Blocks: Identifiable and assessable component parts of a Tricky Topic that are common to a variety of learners' problems. You would expect to find at least 3 or 4 key Stumbling Blocks in a Tricky Topic but there may be as many as 6. This number is based upon evaluated practice-based application and psychological memory retention literature which suggests 4 (+ or -) 2.
3) Problem Examples: Examples of the problems that learners have, which display their misunderstanding of the Tricky Topic and are symptoms of one or more Stumbling Blocks within that Tricky Topic.

A mind mapping exercise that moves on to a more structured mapping exercise helps to deconstruct the Tricky Topics and their related stumbling blocks and problem examples (see http://tricky-topics-guide.ac.uk and Adams et al. 2018). Once the Tricky Topic concepts are identified with practitioners, they are mapped together into a one-to-many relationship and then classified by the 'problem distiller'. This is a classification table (theoretically underpinned from pedagogical literature; see Adams & Clough, 2015) which supports identifying why learners have specific problems in Tricky Topics.

Tricky Topic concepts can be 'captured' through an online Tricky Topic Tool (TTT). This tool was previously co-created with teachers, subject matter and educational experts to facilitate developing appropriate evaluation questions (Adams & Clough, 2015). An important part of the TTT is the construction of quizzes, pedagogically underpinned throughout to deconstruct and determine the effective acquisition of knowledge by an individual, in what is termed 'deep learning'. This process turns quizzes into an effective tool for assessing deep learning. There are further questions that need to be researched into how well they identify long-term transformative learning, but this is not the focus of the current paper.

4. Evaluation method

Whilst there is procedural guidance for taking a first account from a child (based on evidence-based practice), there is no direct specific training provided. In addition, evaluations of the training that exist only broadly relate to police practice e.g. training in taking accounts from vulnerable witnesses. These limitations were reported nationally as an 'area of concern' (HMIC. 2015), that identified the need for all police officers to improve their ability to listen and communicate with children, especially when taking an 'initial response' witness account when first arriving on a scene. The report (HMIC. 2015) went on to highlight the current reliance on simplistic online training that was deemed 'ineffective', as it 'does not provide any opportunity for reflection' (p.67, HMIC. 2015). Further insights for the evaluation of this project were provided by internal police reports, such as, the 'Achieving Best Practice' (ABP) guidelines, which provided procedural direction on safeguarding the welfare of children whilst collecting high quality evidence (Binsubaih, et al. 2006., Blandford, 2013, HMIC. 2015, HMCPSI/HMIC. 2014).

It is important to note that the main aim of the study was to evaluate the CIS intervention (i.e. games-based learning for policing), but in order to do so, a novel co-created method was developed, which became an additional contribution (i.e. the value of Tricky Topics as a co-created evaluation method for game-based learning in a practice contexts). There were two levels to the evaluation taken in this project. The first was a traditional researcher-led (not co-created) approach to games evaluation, which focused purely on previous games research techniques such as:

- Learner attitudes towards game-based learning (see Appendix 3)
- Learner engagement when interacting with the game e.g. through a short version of the Game Experience Questionnaire (see Hart et al. 2017)

However, for the second level of evaluation, the participatory approach involved co-created evaluation methods focused on practice-based learning. Through working with police trainers, the College of Policing, police officers and interviewing research experts, the co-created evaluation process produced collaborative research questions that there were both generic and specific:

- Generic: How effective, compared to existing face-to-face training, is game-based learning for training new police recruits in taking initial accounts from children?
- Specific: How does game-based learning support new police recruits' conceptual understanding when interacting with a child (i.e. when taking an initial account), in relation to the specific Tricky Topics of "rapport" and "interview techniques" (see figures 4 and 5). These topics emerged from applying the Tricky Topic process (described below).

As a comparison to CIS, the existing face-to-face training (benchmarked by Ofqual & the College of Policing) provided the following list of current UK police face-to-face training;

- Provide an initial response to incidents
- Conduct Priority and Volume Investigations
- Interview Victims and Witnesses

These were verified by trainers and the College of Policing to match the game learning points, providing effective face-to-face reference conditions.
To address both the researcher-led (i.e. not co-created) and co-created 'generic' and 'specific' research questions, a mixed methods approach was used to evaluate the effectiveness of the game-based learning system. This combined a Randomised Control Trial (RCT) that used a mixture of existing game evaluation questionnaires along with a co-created Tricky Topic knowledge-based quiz (see Appendix 1 & 2). Researcher-led evaluation questionnaires consisted of Player Experience (PX) and User Experience (UX) questionnaires (see Hart et al. 2017) that captured player engagement and the overall user experience of interacting with the game-based tool. Learners' attitudes towards different types of training were also captured through researcher-created questionnaires. Finally, researcher-led focus groups and in-depth interview proformas elicited further feedback from participants about their learning experiences. The findings relating to engagement are presented in Hart, et al. (2017), in this paper we focus on the role of co-created learning evaluations. The co-created and researcher-led questions produced learning evaluation data through a triangulated understanding gained from combining quantitative with qualitative data. The distinction of co-creation was implemented to increase research rigor and was not designed as a variable to study in itself.

4.1 Application of the Tricky Topic process

Through the participatory process, several barriers to police understanding were identified. As part of co-creating the evaluation, further consultation with the police trainers, experienced police officers, expert child interviewers and a review of practice and research literature, led to the development of two Tricky Topics (Rapport and Interview Technique – see Figures 4 and 5). Tricky Topic concepts were then 'captured' through the Tricky Topic Tool (as described in Section 3.1).

The pedagogical design underpinning the tool imposed a minimum of three stumbling blocks per Tricky Topic. In a similar manner, a maximum of six stumbling blocks was imposed as a constraint. When dealing with the various stumbling blocks, the Tricky Topic Tool supports the development of a well-formed quiz that ensures all the stumbling blocks are covered with at least three questions. Finally, when collating the results from the quizzes, the tool facilitates intuitive spider graph visualisations of student answers that relate to the different stumbling blocks (see example in Figure 3). The real power of the visualization emerges when cross-referencing results amongst peers, and between a single individual and the aggregate analysis of the entire learner cohort. A police expert provided input throughout the process of developing the Tricky Topic quiz. The evaluation measures are further described below.

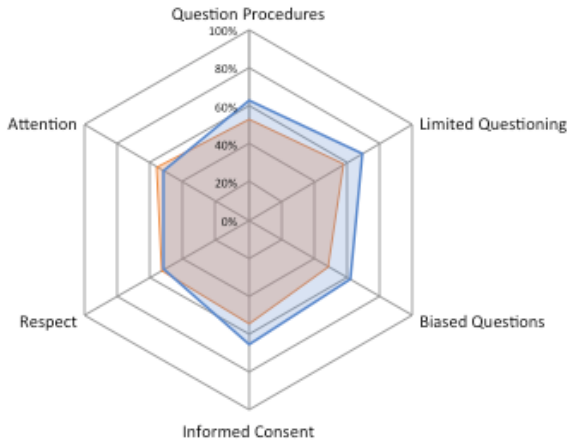

Figure 3. Rapport and Interview Technique Tricky Topics

4.2 Participants

Data was collected from a total of 116 participants from the target population. These were new recruit police

officers across three different UK police forces that were currently on their 13-20 week 'Initial Police Learning and Development Programme' (IPLDP). The different police forces covered different sized organisations, organisational structures and cultural backgrounds to ensure that findings could be generalised beyond specific police force practices or police officer biases.

There were slightly more male (56%) to female participants, with 80% falling below the age of 35. The majority (90%) had studied beyond secondary school, either obtaining a Further Education qualification (42%), an Undergraduate Degree (41%) or a Masters Degree (6%). Just under half (42%) of the participants stated they play games regularly. Around half (48%) of the participants *'would not call themselves gamers'*, and only 4% of the participants described themselves as 'expert *hardcore gamers'*, with the remaining being split between *'casual gamers'* (28%) or *'moderate gamers* (20%). A total of 21% felt that games can 'develop problem-solving *skills'*, and 18% felt they were 'good at promoting teamwork and communication', with the majority having a neutral position.

4.3 Procedure

Tricky Topics (TT) quizzes were used to capture specific knowledge acquisition around learners' conceptual understanding specific to collecting first accounts from a child. The TT quizzes captured data at three key points during the IPLDP, once before and once just after a 3-5 day 'vulnerable witness interview training' course (the length depended on the police force), and then again just after interacting with the game. Although the face-to-face interview training did not focus specifically on collecting first accounts from children, it has been deemed as providing appropriate reference conditions with general skills developed for initial witness interviewing following an incident, by the Ofqual qualifications credit framework and the College of Policing Diploma in Policing (in which the IPLDP sits). The game and the face-to-face training were counterbalanced to overcome practice effects (game then f2f; f2f then game). The TT quiz data was collected via an online survey tool (see Appendix 1).

The focus group data was captured post-game interaction. A series of fourteen focus groups (which varied in number N=5-19 total: 116) were conducted using a semi-structured interview guide. These typically lasted between 20-25 minutes and were constructed and led by one or two researchers. The qualitative data (focus groups) was audio recorded for later transcription, and completed within the three police force training centres.

4.4 Measures

Each TT was split into either three or four stumbling blocks each, so for Rapport (Attention, Empathy, Respect and Informed Consent), and Interview Technique (Question Procedure, Limited Questioning and Biased Questioning, see Figures 4 and 5).

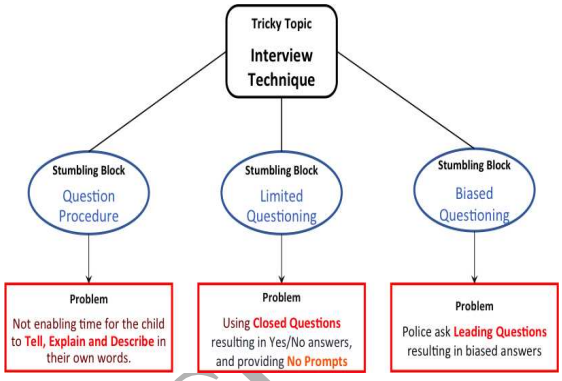

Figure 4. 'Interview Technique' Tricky Topic

A total of 17 questions (see Appendix 2) were constructed focusing on 7 stumbling blocks that were specifically identified by the police as important for evaluation (a further 12 reviewed the stumbling blocks as part of a scenario). An important part of these quizzes that helped to unpick the depth of understanding was the linkage of the questions to stumbling blocks, which in turn were linked to problem examples that the police had identified as barriers to retaining this knowledge. One problem example could relate to several stumbling blocks and so a single question could link several stumbling blocks.

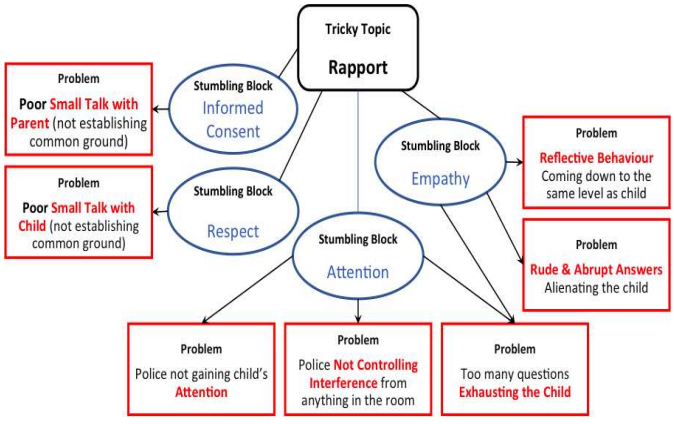

Figure 5. 'Rapport' Tricky Topic

For instance, the problem example of, 'asking too many questions and exhausting the child' was reflected by the game's design, where the child will respond negatively if the player asks too many questions. This learning point is linked to two stumbling blocks within the Rapport TT: 'attention' and 'empathy'. A single question (see below) was created to test the players understanding

of this learning point, connected to two stumbling blocks, thus making a more difficult question and uncovering more depth of understanding than if there had just a separate question for each stumbling block.

Question: When collecting the initial account from a child witness, do you: (Correct answer is C).

A: Repeat every question back to the child to ensure they have answered it correctly.
B: Make sure you fully cover every detail required for an accurate account to be drawn.
C: Try to ask only a limited number of questions that cover the key points.
D: If the child starts to look tired offer them a drink and continue the questioning.

These difficult questions for multiple stumbling blocks, more accurately reflect practice knowledge, which is often complex.

4.5 Analysis methods

A three-stage quantitative analysis was conducted on the TT quizzes to assess learning. The first stage involved the combined data from across the three police forces to identify overall findings: in the second the data was split across the three police forces to identify any specific trends within and between police forces. Lastly, the data was split into the separate stumbling blocks across the police forces to identify specific areas of learning progression after interacting with the game. This further identified how the game facilitated learning progression e.g. regarding aspects of attention or biased questioning.

An in-depth analysis was conducted on the qualitative interview data with a focus on engagement (Hart, et al. 2017). This paper includes a further analysis in relation to triangulating and verifying the Tricky Topics of Rapport and Interview Technique. The qualitative analysis coding was guided by the frequency and fundamentality approach (Adams, et al. 2008) with an emphasis on those concepts that occurred frequently or those that were deemed in the police context as of fundamental importance. This approach followed quality guidelines for research (Henwood and Pidgeon, 1992).

5. Findings

A one-way repeated measures ANOVA was conducted to compare students TT quiz scores gained from Interview Training (Pre & Post), with Games Training (Pre & Post). For the police, the term ‘game’ was controversial (as it implied triviality) and so ‘simulation’ was used instead, although the system contains both game and simulation elements. Two significant differences were identified from this analysis. There was a significant improvement ($p < .001$, M=9.2 to M=10.5) in understanding following the games training of the Tricky Topics with a moderate effect size (Cohen’s d=0.5). This was in comparison to the face-to-face training which had a significant decrease ($p < .05$, M=9.9 to M=9.3) in understanding, with a small effect size (Cohen’s d=0.2). These results are unlikely to be due to practice effects as the game was counter balanced with the face-to-face training condition, i.e., AB / BA split testing for game then f2f and then f2f then games. Variations in particular police forces, although evident, where minimised through conducting the trials across three police forces.

A significant difference was also found between **Pre-Training** (Wilks’ Lambda = .93, $F(1, 115) = 8.7$, $p < .01$, $\eta^2 = .07$), **Post-Training** (Wilks’ Lambda= .96, $F(1, 115) = 4.8$, $p < .05$, $\eta^2 = .04$), and **Interaction** (Wilks’ Lambda= .84, $F(1,115) = 16.5$, $p < .001$, $\eta^2 = .13$). Post hoc comparisons using Bonferroni tests indicate that the mean score for Pre-game (M=9.2) was significantly lower than Pre-Interview (M=9.9), yet the direction changed for Post-game (10,5) with Post-Interview (9.3), both showing moderate effects (partial eta squared - $\eta^2 = .04$-7).

The significant interaction effect can be clearly seen in Figure 6, clearly indicating that the games training increased (Pre-Post Training), compared to the Interview training that decreased (Pre-Post Training).

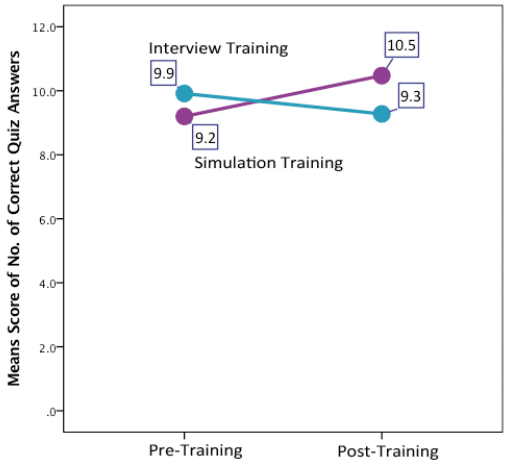

Figure 6. Pre/post F2F and games (simulation) training

5.1 Stumbling block analysis

To gain an understanding of the reasons for the conceptual changes seen in the general findings, further analysis was conducted to deconstruct the data according to the 7 stumbling blocks (directly linked and balanced with the stumbling blocks in the quiz questions) across all the data.

T-tests on the pre-post quiz questions for each stumbling block were analyzed for games and face-to-face training. The results were used to identify exactly where the statistical differences in understanding were for the game and for face-to-face training. This then helped to identify what areas of conceptual understanding the game or the face-to-face training were supporting or not supporting.

Four of the seven stumbling blocks showed a statistically significant improvement in understanding but only for the game-based training. One of the stumbling blocks showed a statistically significant decrease in understanding for the face-to-face training (see Table 1).

Stumbling Blocks	Pre M	Post M	t-test	Cohen's d
(G) Empathy	M= 3.1	M=3.7	5.2***	Mod 0.5
(G) Attention	M= 2.0	M= 2.5	6.0***	Mod 0.5
(G) Informed Consent	M= 1.9	M= 2.4	4.7***	Large 0.8
(G) Question Procedure	M= 2.2	M= 2.5	3.2**	Small 0.3
(F2F) Respect	M= 1.7	M= 1.5	2.5*	Small 0.3

***p<.001, **p<.01, *p<.05. (G)=Game, (F2F)=Face-to-Face

Table 1. Stumbling block Quiz breakdown

The analysis was conducted across three different police forces that employed variations within their delivery and content of their face-to-face training programmes.

5.2 Qualitative findings

The conceptual understanding resulting from the game compared to the face-to-face training were triangulated with the qualitative data. A key point the quantitative data identified for the Tricky Topic **rapport** was around the stumbling block of empathy with the child. For example, one problem example focused on issues that the police had in asking too many questions and exhausting the child. The qualitative data verified this issue as police learners seemed to value how realistic narratives and character responses could transfer into real-life activities:

"I liked you could ask too many questions … they [the child] *just get a bit bored of you… it's* like realistic in that *sense" [P9, G5]*[1]

The interview data also verified the specific stumbling block of gaining the child's **attention,** in particular through establishing a relationship with the child and controlling interferences. The police learners noted in their own words how this was tacit knowledge that the game helped them to focus upon:

"Gaining your rapport. It's common sense, if you think about it but you don't always think about it." [P2, F7]

In addition, game mechanics, such as the feedback mechanisms like the rapport bar, increased player interest and learning motivation for the Rapport Tricky Topics:

"You were conscious of that green bar, so it kept you alert the whole time" [P10, F7]

The qualitative data also expanded on perceptions of rapport, focusing on the problem examples of establishing common ground with the parent (overcoming the stumbling block of **informed consent**) and with the child (overcoming the stumbling block of **respect** from the child). They noted the value of realistic interaction through the gameplay that triangulated with overcoming these stumbling blocks.

For example, the placing of interactive objects within the gameplay environment was intended to inspire realistic curiosity for trainee police officers and provide prompts for discussion with the characters. This relates well to the Tricky Topics of establishing common ground with the child and with the parent in real-life.

*"Havi*ng a *look around the room… that's what you do* when you normally go *into a room… you look around" [P* 7, F2]

The quiz also identified that the face-to-face training decreased an understanding of **respect**, and how it is supposed to emerge through establishing common ground with the child. When triangulating respect through the qualitative data, there was particular emphasis on this as a focus of effective learning in the game. Part of the reason for this emphasis, could be how (in contrast to face-to-face training) the game presented interactive objects that inspired the players to see themselves as investigative police officers looking for 'clues' within the environment that could help them establish common ground.

"Clues in like the trophies, football, you could click on it and it tells [you/trainees] *what they're interested in…"* [P6, F2]

The qualitative data also triangulated with the Tricky Topic of **interview techniques,** which has stumbling blocks of biased and leading questions as well as needing to give the child time to respond. The police learners verified this Tricky Topic through reflecting on how they would apply specific learning points in future situations:

[1] Quotes are anonymously related to participants via a number e.g. P10 and to the focus group they participated in e.g F5

"I was thinking as I was going through, if I was to speak to a child now, I would change *my type of questioning"* [P8, F5]

In particular, the findings identified how the learners understanding of the Tricky Topics helped them to develop deeper understanding for future application; as one trainee explains:

"You learn to play the game, and you also develop a sort of skill base you can take to reality [P8, F3]

6. Discussion

The project's participatory approach helped to support co-created evaluations from early on in the project. The approach made it easier to identify the Tricky Topics and develop the conceptual understanding quizzes. The project methods sought to increase research rigour by triangulating different methods and data sources, both co-created and researcher-led (i.e. not co-created). Within this approach, a range of information was gathered from procedural guidance and findings from empirical studies to personal experiences and practice-based interpretive reports. The breadth of resources and knowledge exchange made the development of effective practice-based evaluations with Tricky Topic research questions possible.

As can be highlighted from the findings, it was the co-created quiz questions around conceptual understanding that produced the key insights that could be used for wider impact. For example, a report on the findings from this research project has been produced and circulated to the police. This report has received positive responses from senior representatives in the police forces and has since been showcased to the UK College of Policing. Upon receiving the conceptual understanding results several UK police forces and the College have expressed a desire to acquire the game for their internal training. The game has now been turned into a commercial product for use by the police.

6.1 Tricky Topic evaluation approach

The co-created Tricky Topic quiz helped identify a more in-depth understanding of how learning progression resulted from game-play. A statistically significant finding was identified with moderate effect sizes for game-based learning, indicating that the Child Interview Simulator led to increased understanding when compared to face-to-face training. In fact, the face-to-face training produced a decrease in understanding, though this was not statistically significant at the broader level. However, one of the strengths of the co-created Tricky Topic evaluation is that it allowed a further breakdown of learning concepts for a more in-depth analysis. Once broken down into stumbling blocks, it appeared that this decreased understanding was related to the concept of 'respect' in face-to-face training. A possible explanation for this effect may relate to the lack of specificity in the current face-to-face training which currently only refers to 'vulnerable witnesses', where recruits may then grow confused when having to deal with the specific issues of interviewing a child. However, it must be highlighted that the current IPLDP training has, until this research, been considered by the national Ofquals Qualification Credit Framework as adequately providing skills development for first response child witness interviewing. Our findings have therefore had a major impact on changing the current training since it was previously assumed that generic training was effective at meeting specific needs and not potentially detrimental, as identified here. The tricky topic evaluation also allowed for a deeper analysis of the statistically significant increase in understanding via game-based learning. The stumbling block analysis also identified that the game particularly helped improve tacit understanding.

Through triangulating the Tricky Topic findings with the qualitative findings, we are able to provide not only a deeper understanding of the learner but also the statistical importance of those comments. Through relating qualitative themes to stumbling blocks (e.g. empathy and attention) we identified the design implications for factors that influenced the tacit and procedural learning. The qualitative data also helped to unpack how the game supported the development of deeper understanding concerning the connected stumbling blocks. This in turn illustrated how the game has the potential to feed into ongoing practice. For example, 'feedback' is seen as an important way to support learning in games (Haferkamp, et al. 2011), where the Tricky Topic findings gave more detail to exactly how they supported learning. While the rapport bar is obviously not something that exists outside of the game, it was able to provide relevant real-time feedback on player actions that they could use to progress. It might be obvious that a rapport bar impacts on understanding, but the Tricky Topic approach provides more detailed evidence of this relationship and illustrates how the approach can be used to evaluate learning technologies.

It could also be argued that the stumbling block findings, triangulated with the qualitative analysis, indicate how the narrative helped narrow the gap between the trainee recruits virtual and real-world identities. Players were able to adopt what Gee (2004) refers to as a 'projective identity', where they could reflect on their own learning. Games have been found to support different levels of reflection (Mekler et al. 2018) but our findings

suggest that one mechanism for doing so relates to the relevance of the learning point, rather than providing an environment that is completely realistic. The qualitative findings triangulated and verified the Tricky Topic quiz results to highlight what relevant learning can result (e.g. in relation to rapport). However, it is important to note how these issues were identified (i.e. through co-created knowledge tests) and the need to experimentally review and compare co-created and researcher-led variables. While the potential to operationalise and control variables in a practice-based setting is very limited, it could be valuable to experimentally review these variables, without the added complexity of comparing a game to face-to-face delivery.

The learning within the game was clearly mapped by the co-created Tricky Topic evaluation providing a link to design features and game mechanics (such as feedback mechanisms and learning objects). The Tricky Topic approach also provided guidance for learning challenges in the game e.g. don't tire out the child with questions. Challenge is generally seen as important for facilitating learning, where, for example, Iacovides, et al. (2015) illustrate the ways in which breakdowns provide opportunities for players to develop deeper understanding. Triangulation of the data (i.e. the comparison between the Tricky Topic quizzes and the qualitative findings) identified that the game not only challenged players, but that the challenge led to a more rewarding learning experience. The qualitative data has also indicated the potential for longer term benefits where the police have learned not only how to improve their performance in the game, but in real-world settings. Further Tricky Topic and qualitative evaluations with the police are required to identify how this benefit impacts upon day-to-day practices.

From triangulating the quantitative and qualitative it appears that learning resulted from the merging of both procedural tasks and tacit in-game feedback-mechanisms (e.g., active objects, rapport bar), which were interwoven with decision-making within the storyline (e.g. selecting interview techniques, paying attention to information). In relation to evaluating a game for impact, we argue it is particularly important to ensure co-creation of evaluation measures. Within this project multiple evaluation metrics were used. However, it was the co-created Tricky Topic measures of 'rapport' and 'interview technique' that were verified by the qualitative data as having significance for practice outcomes and further informing future training processes. The approach also shows how co-created evaluations might be applied to other learning technologies or even other types of evaluations outside of education.

6.2 Co-Creating evaluations

Grand, et al. (2015) highlight that whilst two-way engagement is often described in research, the level of genuine reciprocity across the research process is debatable (e.g. see Iacovides et al. 2019 for a discussion). Part of the reason for this could be due to misconceptions around the terminology used within participatory research. Cornwall and Jewkes (1995) propose that there have been some discrepancies in how the term is applied. For example, different notions of participatory, participation and participant emphasize tensions in power structures and drivers between domains and processes thus changing the level of equity. It could be argued that the challenge is not participation per se but in supporting genuine equity throughout the research process. Whilst this is not appropriate for all forms of research, this paper argues that for a practice-based HCI research project, equity should be the focus for methodologies throughout the research cycle. However, defining co-creation methods in practice is very different to laboratory-based co-creation, where variables can be more effectively controlled. Variations in practice contexts, such as norms and biases, can confound results. Nevertheless, equity between researcher and practitioner in a practice context has been the premise behind the development of the Tricky Topic process. The practice context was particularly helpful in identifying the trainers' (and national benchmarking standards) inaccurate assumptions about the value of current face-to-face training providing adequate knowledge development for child witness interviewing.

Many within participatory research have focused on only 'participation' as requiring equity between different participants (Haferkamp, et al. 2011., Machiocha, 2014., Khaled and Vasalou, 2014). As noted earlier we have seen this impacting strongly on participatory design (Meyer and Land, 2006., Carroll, 1996., Muller and Kuhn, 1993., Hall, et al. 1982) as well as in games research, and is starting to impact upon evaluations for games (Costabile, 2008., Dontcheva, 2014). However, participation is still not feeding into the fundamentals of the research, such as the design of evaluations where control tends to remain with researchers and academics. It could be argued that limited co-created evaluations seen in the games-based learning literature reviewed are due to poor support for implementing co-creation in evaluations. There is also a limited number of papers that review evaluation approaches especially through experimentally comparing different methods. Whilst experimental laboratory work may provide valuable insights, there may be additional differences when an evaluation approach is applied within practice contexts. Within HCI, as has been noted, there has been a tradition of participatory design in practice contexts that has framed an equitable

collaboration between researchers and practitioners. It seems at odds then that this has not transferred more effectively into evaluation procedures.

6.3 Wider HCI implications

Within HCI the balance between theory and practice, compared to research and implementation is an evident issue. HCI has never sat in an Ivory tower isolated from practice, through in design and in particular participatory design there has been a close connection between theory and practice. Co-creation in evaluation approaches have been slower to change, with ethnographic and qualitative methods assumed to be making the largest impact on participatory evaluation in HCI (Adams & Clough, 2015).

With the Tricky Topic approach in this paper we take a co-created, practice-based approach applied to a mixed method procedure using a Randomised Controlled Trial (RCT) evaluation triangulated with qualitative data. It must be recognised that within the use of a Tricky Topic evaluation approach there is the potential to increase practice relevance for evaluation methods, regardless of their epistemological underpinnings.

With respect to the HCI literature on game-related evaluations, we were unable to find many examples of co-created approaches. While a variety of methods have been employed to assess learning in games, such as game analytics and in-game assessments (e.g. Culbertson, 2016) and the use of pre and post-tests (e.g. Chen and Chen, 2013), it is not always clear how an evaluation was developed. For instance, Dunwell, et al. (2011) present the evaluation of 'Everand' (a game that teaches about road safety) which included surveys, game logs and interviews. However, little information is provided about who was involved in the design of the methods and it seems implied that the research team was responsible for making most decisions about what to focus the evaluation on. This highlights that for game-based learning research in particular that there is a real potential to advance co-created evaluations.

7. Conclusion

This article presents the Tricky Topic approach as way to support co-created evaluations with stakeholders in a practice-based context. The Tricky Topic approach led to a co-created evaluation method that was applied along with the researcher-led (i.e. not co-created) evaluation methods to evaluate a game-based learning system for training police officers in child interviewing. While the main aim of the study was to evaluate CIS, the insights uncovered would not have been identified without the method we developed. The particular effectiveness of the Tricky Topic approach was identified when triangulating findings with researcher-led evaluation approaches. The approach allowed for a tailored evaluation that not only identified a statistically significant increase in police understanding but also allowed for a detailed breakdown of how the system supported this understanding. This breakdown is also of scientific interest since the process identified how tacit knowledge in particular was supported by the game, in a way that the existing face-to-face training did not.

As an additional outcome from this research we identified that the researcher-led (i.e. not co-created) evaluation methods did not on their own provide an effective level of clarity around the gaming intervention. However, as already noted, the comparison between co-created and researcher-led methods needs to be more specifically and experimentally evaluated in future research using extended timeframes and resources. Ultimately then, this project has identified the advantage that HCI has over others disciplines as co-created design provides a natural move into co-created evaluation through extending current approaches. It has been argued that when co-creating game design, we provide a more effective solution for practice needs and for our customers. Similarly, when co-creating an evaluation method, we can provide a more effective evaluation for practice needs and our customers. This is not to say that the objectivity and rigor of research evaluation and laboratory-based approaches is devalued. Just as the creativity and expertise in design is not devalued by co-created design. If anything, the value for these skills increases as those who participate and co-create increase their understanding of the depth of expertise required within HCI. What has changed is researchers' understanding and value of practice, and practitioners' expertise, regardless of the field.

Acknowledgments

We would like to acknowledge the funding received from HEFCE and the College of Policing for the initial participatory design and evaluation work. The detailed evaluation work was then further funded by the 'Centre for Policing Research and Learning'. We would also like to thank the police forces (Lancashire police, Thames valley police and Bedfordshire police) who helped to co-design the system and the evaluation processes and later supported the randomized controlled trial.

References

Adams, A., Pike, A., McFarlane, R., Clough, G., Hart, J., Sargeant, J., Anastasopi. P., Hartnett, E. (2018) Teaching and learning Tricky Topics. badged open course. http://www.open.edu/openlearn/education-

development/learning/teaching-and-learning-tricky-topics/content-section-0

Adams, A., Lunt, P. and Cairns. P. (2008). A qualitative approach to HCI research. In: Cairns, P and Cox, A eds. Research Methods for Human-Computer Interaction. Cambridge, UK: Cambridge University Press, 138–157.

Adams, A., and Clough, G. (2015). The e-assessment burger: supporting the before and after in e-assessment systems. Interaction Design and Architecture(s)(25) 39–57. ISSN: 2283-2998

Alenljung, B. and Söderholm, H. (2015). Designing Simulation-Based Training for Prehospital Emergency Care: Participation from a Participants Perspective. In: Masaaki Kurosu (ed.), Human-Computer Interaction: Designing and Evaluation: 17th International Conference, HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings, Part I (pp. 297-306). http://dx.doi.org/10.1007/978-3-319-20901-2_27

Binsubaih, A., Maddock, S. and Romano.D (2006). A serious game for traffic accident investigators. Interactive Technology and Smart Education, 3(4), 329-346. doi: 10.1108/17415650680000071

Blandford. A (2013). Semi-structured qualitative studies. In: Soegaard M, Dam RF (eds) The encyclopedia of human-computer interaction, 2nd edn. The Interaction Design Foundation, Aarhus. Retrieved April 14th 2017: http://www.interaction-design.org/encyclopedia/semi-structured_qualitative_studies.html

Carroll., J. M. (1996). Encountering others: reciprocal openings in participatory design and user-centered design. Human–Computer Interaction, 11(3), pp.285-290.

Chappell, A. (2000) Emergence of participatory methodology in learning difficulty research: understanding the context, British Journal of Learning Disabilities, 28, 38–43.

Chen, Z. H. & Chen. S. Y. (2013). A surrogate competition approach to enhancing game-based learning. ACM Transactions on Computer-Human Interaction (TOCHI), 20(6), 35.

Cornwall, A. and Jewkes. R.(1995). What is participatory research?. Social science & medicine, 41(12), 1667-1676.

Costabile, M. F. Angeli, A. De Lanzilotti, R. Ardito, C. Buono, P.and. Pederson. T (2008), April. Explore! possibilities and challenges of mobile learning. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 145-154). ACM.

Crookall. D (2010). Serious games, debriefing, and simulation/gaming as a discipline. Simulat Gaming, 41, 898–920, doi: 10.1177/1046878110390784

Culbertson, G. Wang, S. Jung, M., and Andersen, E. (2016), May. Social Situational Language Learning through an Online 3D Game. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 957-968). ACM.

Di Loreto, I., Mora, S., Divitini., M. (2012). Collaborative serious games for crisis management: an overview. In Proc. IEEE 21st International WETICE, doi: 10.1109/WETICE.2012.25

Dontcheva., M. Morris, R. R., Brandt, J. R. and Gerber. E. M. (2014), April. Combining crowdsourcing and learning to improve engagement and performance. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems (pp. 3379-3388). ACM.

Dunwell, I. Christmas, S. and de Freitas, S. (2011). Code of Everand: Evaluation of the Game. London: Department of Transport.

Gee., J. P. (2004). What video games have to teach us about literacy and learning. New York: Palgrave Macmillan.

Grand, A., Davies, G., Holliman, R. and Adams, A., (2015). Mapping Public Engagement with Research in a UK University. PLoS ONE, 10(4) pp. 1–19.Guisti. et al, 2011

Giusti, L., Zancanaro, M., Gal, E. and Weiss. P. (2011). Dimensions of collaboration on a tabletop interface for children with autism disorder, Proceedings of SIGCHI Conference on Human Factors in Computing Systems. May 07-12, Vancouver, BC. Canada doi>10.1145/1978942.1979431

Haferkamp., N. Kraemer., N. C. Linehan., C. and Schembri., M (2011). Training disaster communication by means of serious games in virtual environments. Entertainment Computing, 2 (2), 81-88, doi: http://doi.org/10.1016/j.entcom.2010.12.009

Hall, B. Gillette A., and Tandom. R. (1982). Creating Knowledge a Monopoly? Participatory Research in Development. International Council for Adult Education, Toronto.

Halskov., K. and Brodersen-Hansen., N. (2015). The diversity of participatory design research practice at PDC 2002–2012, 74, 81-92. https://doi-org.libezproxy.open.ac.uk/10.1016/j.ijhcs.2014.09.003

Harlow., A. J. Scott, P. M. and Cowie. B. (2011). Getting stuck" in Analogue Electronics: Threshold Concepts as an Explanatory Model. European Journal of Engineering Education, 36 (5), pp. 435-447

Hart., J. Iacovides., I, Adams., A. Oliveira., M. and Magroudi. M (2017). Understanding Engagement within the Context of a Safety Critical Game. In: The ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play (CHIPLAY 2017), October 15-18 2017, Amsterdam, Netherlands.

Henwood., K. and Pidgeon., N. (1992). Qualitative research and psychological theorizing. Brit J Psychol,

83(1), 97-111, doi: 10.1111/j.2044-8295.1992.tb02426.x

HMIC Her Majesty's Inspectorate of Constabulary (2015). In harm's way: the role of the police in keeping children safe HM Crown Prosecution Service Inspectorate (HMCPSI) and HM Inspectorate of Constabulary (HMIC) https://www.justiceinspectorates.gov.uk/hmic/wp-content/uploads/in-harms-way.pdf

HMCPSI Her Majesty's Crown Prosecution Service Inspectorate / HMIC. (2014). Achieving Best Evidence in Child Sexual Abuse Cases - A Joint Inspection, December 2014. Available from: www.justiceinspectorates.gov.uk/cjji/wp-content/uploads/sites/2/2014/12/CJJI_ABE_Dec14_rpt.pdf

Iacovides, I., Cox, A., Furniss, D., Stawarz, K., Jennett, C., & Adams, A. (2019). Supporting engagement in research through a game design competition. Research for All, 3(1), 25-41.

Iacovides., J. Cox., A. L. McAndrew, P. Aczel, J. and Scanlon., E. (2015). Game-play breakdowns and breakthroughs: exploring the relationship between action, understanding, and involvement. Human–computer interaction, 30, 3-4, 202–231, doi:10.1080/07370024.2014.987347

Khaled., R. and Vasalou., A. (2014). Bridging serious games and participatory design. International Journal of Child-Computer Interaction 2 (2), 93-100.

Law., E, L and Sun., X (2012) Evaluating user experience of adaptive digital educational games with activity theory, Int. J. Hum-Comput St, 70, 478-497 doi: 10.1016/j.ijhcs.2012.01.007

Lavie., T and Tractinsky., N. (2004) Assessing dimentions of perceived visual aesthetics of web sites. Int. J. Hum-Comput St studies, 60(3), 269-298, doi:10.1016/j.ijhcs.2003.09.002

Lenz., S (2012). Participatory Research in Argentina: Three Experiences in the Educational Field within the Context of Reinstatement of Democracy. Forum: Qualitative Social Research, 13(1), Art. 3, http://nbn-resolving.de/urn:nbn:de:0114-fqs120133.

Linssen., J. Theune., M. de Groot., T. and Heylen., D. (2015). Improving social awareness through thought bubbles and flashbacks of virtual characters. In Proc. International Conference on Intelligent Virtual Agents, 250-259. Springer International Publishing, doi: 10.1007/978-3-319-21996-7_25

Lucas., U. and Mladenovic., R. (2007). The potential of threshold concepts: an emerging framework for educational research and practice. London Review of Education, 5 (3)**,** pp. 237-248

Lukosch., H. van Ruijven., T. and Verbraeck., A. (2012). The participatory design of a simulation training game. Proceedings of the winter simulation conference. Winter Simulation Conference.

Margoudi, Maria; Hart, Jennefer; Adams, Anne and Oliveira, Manuel (2016). Exploring Emotion Representation to Support Dialogue in Police Training on Child Interviewing. In: Serious Games JCSG 2016, Lecture Notes in Computer Science 9894, Springer, pp. 73–86.

Machiocha., A. (2014). Teaching research methods: threshold concept. In 13th European conference on research methods for business and management, London

Mekler, E., Iacovides, I., & Bopp, J. (2018). A Game that Makes You Question... Exploring the Role of Reflection for the Player Experience. In Proceedings of the annual ACM Conference CHI Play 2018, pp. 315-327, New York, NY, USA, ACM.

Mergler., D. (1987). Worker participation in occupational health research: theory and practice. International Journal of Health Services, 17, 151.

Meyer., J. and Land., R (2003) Threshold concepts and troublesome knowledge: linkages to ways of thinking and practising within the disciplines. In Rust, C. (Ed.) Improving student Learning - Theory and Practice Ten Years on. Oxford, Oxford Centre for Staff and Learning Development (OCSLD), pp.412-424,

Meyer., J. and Land., R (2006). Overcoming barriers to student understanding: Threshold concepts and Troublesome Knowledge. In Meyer, J. and Land, R. (Eds.) Overcoming Barriers to Student Understanding: Threshold concepts and Toublesome Knowledge. London and New York, Routledge, pp.19-32,

Muller., M. J. and Kuhn., S. (1993). Participatory design. Communications of the ACM, 36(6), 24-28.

Muller., M. J. (2003). Participatory design: the third space in HCI. Human-computer interaction: Development process, 4235, pp.165-185.

O'Donnell., R. (2010). A Critique of the Threshold Concept Hypothesis and an Application in Economics. Working Paper Series, 164, Available at http://www.finance.uts.edu.au/research/wpapers/wp164.pdf [Accessed 31-3-17],

Rowbottom., D. (2007). Demystifying Threshold Concepts. Journal of Philosophy of Education, **41** (2)**,** 236-270.

Sanders., E. B-N. and Stappers., P. J. (2014). "Probes, toolkits and prototypes: three approaches to making in codesigning." CoDesign 10.1. 5-14.

Seale, J. (2010). Doing student voice work in higher education: an exploration of the value of participatory methods. British Educational Research Journal Vol. 36, No. 6, December 2010, pp. 995–1015

Slater., M (1999). Measuring Presence: A Response to the Witmer and Singer Questionnaire, Presence: Teleop Virt, 8(5), 560–566, doi: 10.1162/105474699566477

Stoecker, R. (1999) 'Are academics irrelevant? Roles for scholars in participatory research' in American Behavioral Scientist, 42(5) pp. 840-854

Susi., T. Johannesson., M. and Backlund., P. (2007). Serious Games - An overview. Technical Report HS-IKI -TR-07-001, School of Humanities and Informatics, University of Skövde, Sweden. Retrieved April 14th 2017: http://www.diva-portal.org/smash/get/diva2:2416/FULLTEXT01.pdf

Toups., Z. O. Kerne., A. Hamilton., W. A. and Shahzad., N. (2011). Zero-fidelity simulation of fire emergency response: improving team coordination learning. In *Proc. CHI '11,* ACM Press. 1959-1968, doi: 10.1145/1978942.1979226

Whitton., N. (2014). Digital Games and Learning: Research and Theory. London, UK: Routledge.

Williams-Bell., M. F. Murphy., B Kapralos., B. Hogue., A and Weckman., E. J. (2015). Using Serious Games and Virtual Simulation for Training in the Fire Service: A Review. Fire Technol. 51(3), 553-584. doi:10.1007/s10694-014-0398-