Usability testing for serious games: Making informed design decisions with user data

Pablo Moreno-Ger, PhD.[1]
Javier Torrente, Msc[1]

[1] *Universidad Complutense de Madrid, Facultad de Informática, Madrid, España.*
{pablom,jtorrente}@fdi.ucm.es

Yichuan Grace Hsieh, PhD, RN[2]
William T. Lester, MD, MS[2]
[2] *Laboratory of Computer Science, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA*
{ yhsieh1,wlester }@partners.org

Abstract

Usability testing is a key step in the successful design of new technologies and tools, ensuring that heterogeneous populations will be able to interact successfully with innovative applications. While methods for usability testing of productivity tools (e.g., text editors, spreadsheets, or management tools) are varied, widely available and valuable, analyzing the usability of games, especially educational "serious" games, presents unique usability challenges. Games are fundamentally different than general productivity tools, and instruments designed and used to test the usability of productivity tools may not be optimally effective for serious games. In this work we present a methodology especially designed to facilitate usability testing for serious games, taking into account the specific needs of such applications and highlighting the objective of systematically producing a list of suggested improvements from large amounts of recorded game play data. This methodology was applied to a case study for the MasterMed educational game, a teaching tool to empower patients in their medication knowledge. We present the results from evaluating the game, and a summary of the central lessons learned that are likely useful for researchers who aim to tune and improve their own serious games before releasing them for the general public.

1. Introduction

As the complexity of new technologies increase, affecting wider portions of the population, usability testing is gaining even more relevance in the fields of human-computer interaction (HCI) and user interface (UI) design. Brilliant products and fresh ideas may fail completely if users cannot engage with a new tool because they cannot understand how it is meant to be

used. Consequently, product designers are increasingly focusing on end-user usability testing during the prototype phase in order to identify design or implementation issues that might otherwise prevent users from successfully engaging with the final product.

Prototype usability testing is important when a new tool or system is to be used by a heterogeneous population, especially if this target population includes individuals who are not accustomed to interacting with new technologies. In this sense, the field of serious games in general and educational serious games in particular is a good example of these emerging and complex technologies that require special attention to usability issues.

When designing an educational serious game there are different evaluation dimensions. As products that aim to engage players in meaningful learning activities, it is important to cover the evaluation of aspects such as learning effectiveness, engagement, and the appropriateness of the design for a specific context and target audience [1]. However, in addition to those dimensions, serious games target broad audiences, who may or may not play games regularly, and usability issues alone can hinder the gameplay process, affecting negatively the learning experience.

However, measuring the usability of such an interactive system is not a straightforward process. Even though there are different heuristic instruments to measure usability with the help of experts [2], these methods do not always identify all the pitfalls in a design [3]. Furthermore, usability is not an absolute concept per se but is instead relative in nature, dependent on both the task and the user. Consider the issue of complexity or usability across decades in age or across a spectrum of user educational backgrounds – what is usable for a young adult may not be usable for an octogenarian. It is situations like these where deep insight into how the users will interact with the system is required. One of the most common approaches is to allow users to interact with a prototype while developers and designers observe how the user tries to figure out how to use the system, taking notes of the stumbling points and design pitfalls [4].

On the other hand, prototype evaluation for usability testing can be cumbersome and may fail to identify comprehensively all of the stumbling points in a design. When usability testing sessions are recorded with audio and/or video, it can be difficult to process both recorded user feedback and on-screen activity in a systematic way that will assure all pitfalls are identified. Thus usability testing using prototype evaluation can be a time-consuming and error-prone task that is dependent on subjective individual variability.

In addition, many principles for evaluating the usability of a general software tool may not necessarily be applicable to games or more specifically serious games [5]. Games are expected to challenge users, making them explore, try, fail, and reflect. This cycle, along with explicit mechanisms for immediate feedback and perception of progress are key ingredients in game design, necessary for fun and engagement [6]. So the very context that makes a game engaging and powerful as a learning tool may adversely affect the applicability of traditional usability guidelines to serious games.

For example, typical usability guidelines for productivity software indicate that it should be trivial for the user to acquire a high level of competency using the tool, and that hesitation or

finding a user uncertain about how to perform a task are always considered unfortunate events. A serious game connects the pathways of exploration and trial and error loops to help the player acquire new knowledge and skills in the process [7]. This makes it imperative to differentiate hesitations and errors due to a bad UI design from actual trial and errors derived from the exploratory nature of discovering gameplay elements, a nuance typically overlooked using traditional usability testing tools.

In this paper we present a methodology for usability testing for serious games, building on previous instruments and extending them to address the specific traits of educational serious games. The methodology contemplates a process in which the interactions are recorded and then processed by multiple reviewers to produce a set of annotations that can be used to identify required changes and separate UI issues, game design issues, and gameplay exploration as different types of events.

Most importantly, the objective of the methodology is to provide a structured approach for the identification of design issues early in the process, rather than to provide an instrument to validate a product achieving a "usability score".

As a case study, this methodology was developed and employed to evaluate the usability of a serious game developed at the Massachusetts General Hospital's Laboratory of Computer Science. MasterMed is a game designed to help the patients understand more about their prescribed medications and the conditions for which they are intended to treat. The application of this methodology using an actual game has helped us to understand better the strengths and limitations of usability studies in general and of this methodology in particular. From this experience, we have been able to synthesize the lessons learned about the assessment methodology that can be useful for serious games creators to improve their own serious games before releasing them.

2. Usability testing and serious games

Usability is defined in the ISO 9241-11 as "*the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use*" [8]. This broad definition focuses on having products that allow the users to achieve goals and provides a base for measuring usability for different software products. However, digital games are a very specific type of software with unique requirements, and serious games in particular have the additional objective of knowledge discovery through exploratory learning. This presents unique usability challenges that are specific to serious games.

In this Section we provide an overview of the main techniques for usability testing in general, and then focus on the specific challenges posed by serious games.

2.1. Usability testing methods and instruments

Usability represents an important yet often overlooked factor that impacts the use of every software product. While usability is often the intended goal when developing a software package, engineers tend to design following engineering criteria, often resulting in products

that seem obvious in their functioning for the developers, but not for general users, with correspondingly negative results [9].

In order to alleviate this effect, a variety of methods are typically used to assess for usability. As described by McLeod & Ranger [3], these methods can be broadly catalogued as (i) *expert methods*, in which experienced evaluators identify potential pitfalls and usability issues, (ii) *theoretical methods*, in which theoretical models of tools and user behaviors are compared to predict usability issues, and (iii) *user methods*, in which software prototypes are given to end users to interact.

Among user methods, two main approaches exist: observational analysis, in which a user interacts with the system while the developers observe, and survey-based methods, in which the user fills in evaluation questionnaires after interacting with the system. Such questionnaires may also be used when applying expert methods, and they are typically based on heuristic rules that can help identify potential issues [10].

There are a number of survey-based metrics and evaluation methodologies for usability testing. One method most commonly cited is the System Usability Scale (SUS), given that it is simple and relatively straightforward to apply [11]. It focuses on administering a very quick Likert-type questionnaire to users right after they interact with the system, producing a "usability score" for the system. The Software Usability Measurement Inventory (SUMI) is also a very popular and well-supported tool, providing support for detailed evaluations [12] by measuring five different dimensions (Efficiency, Affect, Helpfulness, Control, and Learnability). In turn, the Questionnaire for User Interaction Satisfaction (QUIS) [13] deals in terms more closely related with the technology (such as System capabilities, Screen Factors, Learning Factors, etc.) with attention to demographics for selecting appropriate audiences. Finally, the ISO/IEC 9126 standard is probably the most comprehensive instrument, as described in detail in Jung and colleagues' work [14].

However, many of these metrics suffer from the same weakness in that they can yield disparate results when reapplied to the same tool [15]. In addition, it is very common for such questionnaires and methods to focus on producing a usability score for the system, rather than the identification and remediation of the specific usability issues that are discovered. This focus on identifying remediation actions as well as the prioritization of the issues and the actions, surprisingly is often missing in studies and applications [16].

When the objective is to identify specific issues that may prevent end users from interacting successfully with the system, the most accurate approaches are observational user methods [4], as they provide direct examples of how the end users will use (or struggle to use) the applications. However, observational analysis requires the availability of fully functioning prototypes and can involve large amounts of observational data that requires processing and analysis. The experts may analyze the interaction directly during the session or, more commonly, rely on video recordings of the sessions to study the interaction. This has also led to considerations on the importance of having more than one expert review each interaction session. As discussed by Boring & Gertman [16], a single reviewer watching an interaction session has a small likelihood of identifying the majority of usability issues. The likelihood of discovering usability issues may be increased by having more than one expert review each

session [17], but this increased detection comes at the expense of time and human resources during the reviewing process.

In summary, usability testing is a mature field, with multiple approaches and instruments that have been used in a variety of contexts. All the approaches are valid and useful, although they provide different types of outcomes. In particular, observational user methods seem the most relevant when the objective is to identify design issues that may interfere with the user's experience, which is the focus of this work. However, these methods present issues in terms of costs and the subjectivity of the data collected.

2.2. Measuring usability in serious games

In the last ten years, digital game-based learning has grown from a small niche into a respected branch of technology-enhanced learning [18]. In addition, the next generation of educational technologies consider educational games (or serious games) as a tool to be integrated in different formal and informal learning scenarios [19].

Different authors have discussed the great potential of serious games as learning tools. Games attract and maintain young students' limited attention spans, and provide meaningful learning experiences for both children and adults [20], while offering engaging activities for deeper learning experiences [21].

However, as games gain acceptance as a valid educational resource, game design, UI development, and rigorous usability testing are increasingly becoming an issue. And while there are diverse research initiatives looking at how to evaluate the learning effectiveness of these games (e.g. [1], [22], [23]), the usability of serious games has received less attention in the literature. Designing games for "regular" gamers is reasonably straightforward, because games have their own language, UI conventions and control schemes[1], but serious games are being increasingly applied with broad audiences that include non-gamers, resulting occasionally in bad experiences because the target audience "doesn't get games" [24].

Designing for broad audiences and ensuring that a thorough usability analysis is performed can alleviate these bad experiences. In this context, Eladhary and Ollila conducted a recent survey on prototype evaluation techniques for games [25], acknowledging that the use of *off the shelf* HCI instruments would be possible, but that these instruments should be adapted to the specific characteristics of games as reported in [26]. In this context, there are some existing research efforts in adapting Heuristic Evaluations (with experts looking for specific issues) to the specific elements of commercial videogames [27], [28]. However, usability metrics and instruments for observational methods are not always appropriate or reliable for games. Most metrics were designed for general productivity tools, and thus focus on aspects such as productivity, efficacy, number of errors, etc. But games (both serious or purely entertainment) are completely different, focusing more on the process than on the results, on enjoyment than on productivity, and on providing variety than on providing consistency [5].

[1] More precisely, each game genre has its own established conventions in terms of gameplay, control schemes and UIs.

Games engage users by presenting actual challenges, which demand exploratory thinking, experimentation, and observing outcomes. Ideally, this engagement cycle intends to keep the users just one step beyond their level of skill for compelling gameplay whereas a game that can be easily mastered and played through without making mistakes results in a boring game [6]. Therefore, usability metrics that reflect perfect performance and no mistakes (appropriate for productivity applications) would not be appropriate for (fun) games [29].

A similar effect can be observed with metrics that evaluate frustration. Games should be designed to be "pleasantly frustrating experiences", challenging us beyond our skill, forcing us to fail and therefore providing more satisfaction when we win [6]. In fact, the games that provide this pleasantly frustrating feeling are the games that we find ourselves to be the most addicting and compelling. On the other hand, there are games that frustrate players because of poor design of their user interface. In these cases, while the user is still unable to accomplish the game's objectives, it is the result of bad user interfaces or flawed game concepts. Usability metrics for serious games should distinguish in-game frustration from at-game frustration [30], as well as contemplating that "obstacles for accomplishment" may be desirable, while "obstacles for fun" are not [5].

Unfortunately, as game designers can acknowledge, there is no specific recipe for fun, and as teachers and educators can acknowledge, eliciting active learning is an elusive target. The usability and effectiveness of productivity tools can be measured in terms of production, throughput, efficacy, and efficiency. But other aspects such as learning impact, engagement, or fun are much more subjective and difficult to measure [31].

This subjectivity and elusiveness impacts formal usability testing protocols when applied to games. As White and colleagues found [32], when different experts evaluated the same game experiences (with the same test subjects) the results were greatly disparate, a problem that they attributed to the subjective perception of what made things "work" in a game.

In summary, evaluating the usability of games presents unique challenges, and requires metrics and methodologies that aim to contemplate their variability and subjectivity of interacting with games, as well as their uniqueness as exploratory experiences that should be pleasantly frustrating.

3. General methodology

As discussed in the previous Section, gathering data to evaluate the usability of a serious game is an open-ended task with different possible approaches and several potential pitfalls. Therefore, there is a need for straightforward and reliable methods that help developers identify usability issues for their serious games before releasing them. In our specific case, we focus on facilitating an iterative analysis process based on observational methods, in which users play with early prototypes and researchers gather data with the objective of identifying and resolving design and UI issues that affect the usability of the games.

3.1. Requirements

From the discussion above it is possible to identify some initial requirements to perform usability testing for serious games.

1. **Test Users.** First, it is necessary to have a set of test users to evaluate the prototype. These test users should ideally reflect the serous game's target audience in terms of age, gender, education, and any other demographic characteristics that might be unique or pertinent to the educational objective of the serious game. In terms of number of test users, according to Virzi [33], five users should be enough to detect 80% of the usability problems, with additional testers discovering few additional problems. In turn, Nielsen & Landauer [34] suggested that, for a "medium" sized project, up to 16 test users would be worth some extra cost, but any additional test users would yield no new information. They also suggested that the maximum benefit/cost ratio would be achieved with four testers. We suggest selecting at least as many users that would span the range of your target audience, but not so many users that hinder the team performing the usability data analysis.
2. **Prototype Session Evaluators.** Another important requirement is consideration of the numbers of evaluators or raters to analyze the play session of each test user. Having multiple evaluators significantly increases the cost, making it tempting to use a single evaluator. However, while some analyses are performed with a single evaluator observing and reviewing a test user's play data, Kessner and colleagues suggested that it is necessary to have more than one evaluator to increase the reliability of the analysis, because different evaluators identified different issues [3]. This effect is even stronger when evaluating a game, because their high complexity results in evaluators interpreting different causes (and therefore possible solutions) for the problems [32]. Therefore, we suggest having more than one evaluator analyze each play session and a process of conciliation to aggregate the results.
3. **Instrument for Serious Game Usability Evaluation.** For an evaluator who is analyzing a play session and trying to identify issues and stumbling points, a structured method for annotating events with appropriate categories is a necessity [17]. Because serious games differ from traditional software packages in many ways, we suggest using an instrument that is dedicated to the evaluation of serious game usability. Section 3.2 below is dedicated to the development of a Serious Game Usability Evaluator (SeGUE)
4. **Data Recording Set-Up**. Nuanced user interactions can often be subtle, non-verbal, fast-paced, and unpredictable. A real-time annotation process can be burdensome, or perhaps even physically impossible if the user is interacting with the system rapidly. In addition, any simultaneous annotation process could be distracting to the user's game interactions and detract from the evaluative process. For these reasons, we recommend screen-casting of the test play sessions along with audio and video recordings of the user with minimal, if any coaching, from the evaluation staff. These recordings can be viewed and annotated later at an appropriate pace.
5. **Ready-for-Play Prototype.** The ready-to-play prototype should be as close to the final product as possible for the test-users to evaluate. The prototype should allow the test-users to experience the interface as well as all intended functionalities so that the interactions could mimic the real play session, therefore, maximizing the benefits of conducting a usability test. When it is not feasible or cost-effective to provide a full

prototype, using an early incomplete prototype may fail to reflect the usability of the final product once it has been polished. White and colleagues [32] conducted their usability studies using a “vertical slice quality” approach, in which a specific portion of the game (a level) was developed to a level of quality and polish equivalent to the final version.

6. **Goal-oriented Play Session Script**. Lastly, prior to the initiation of the study, a play-session script should be determined. The script for the evaluation session should be relatively brief and have clear objectives. The designers should prepare a script indicating which tasks the tester is expected to perform. In the case of a serious game, this script should be driven by specific learning goals, as well as cover all the relevant game-play elements within the design. There may be a need for more than one play session to be exposed to each user so that all the key game objectives could be included.

3.2. Development of the Serious Game Usability Evaluator (SeGUE)

Evaluators who analyze a prototype play session will need a structured method to annotate events as they try to identify issues and stumbling points. This pre-defined set of event types is necessary to facilitate the annotation process as well as to provide structure for the posterior data analysis. This evaluation method should reflect the fact that the objective is to evaluate a serious game, rather than a productivity tool. As described in Section 2.2, serious games are distinct from other types of software in many ways. Importantly, serious games are useful educational resources because they engage the players on a path of knowledge discovery. This implies that the evaluation should focus on identifying not only those features representing a usability issue, but also the ones that really engage the user.

Since the objectives of evaluating a serious game not only focus on the prototype itself but also the process of interacting with the game and the user’s experience, our research team developed a tool, the Serious Game Usability Evaluator (SeGUE), for the evaluation of serious game usability. The SeGUE was derived and refined from two randomly selected serious game evaluation sessions, in which a team comprising game programmers, educational game designers and interaction experts watched and discussed videos of users interacting with an educational serious game. Two dimensions (system-related and user-related) of categories were created for annotation purposes. Within each dimension, several categories and terms were defined to annotate events.

Within the system-related dimension, there are six different event categories. Two event categories are related to the game design, including game-flow and functionality. Events of these categories are expected to require deep changes in the game, perhaps even the core gameplay design. Three event categories are related to the game interface and implementation, including: content, layout / UI and technical errors, where solutions are expected to be rather superficial and have less impact on the game. A non-applicable category is also considered for events not directly related to the system, but still deemed relevant for improving the user experience.

In the user-related dimension, there are ten event categories across a spectrum of emotions: negative (frustrated, confused, annoyed, unable to continue), positive (learning, reflecting, satisfied/excited, pleasantly frustrated) or neutral (non-applicable and suggestion / comment). For researchers' convenience an additional category named "Other" was included in both

dimensions for those events that were hard to categorize. Such events may be an indication that a new category is required due to specific traits of a specific game. More details about the categories and their meanings are presented in Table 1 and Table 2.

Table 1. Event categories for the system dimension

System-Related Event	
Functionality	An event is related to prototype's *functionality* when it is the result of the user activating a control item and it is related to one specific action.
Layout/UI	An event is related to *layout/UI* when the user makes a wrong assumption about what a control does, or when the user does not know how to do something (negative events). It is also a *layout/UI* positive event when a user appreciates the design (figures, attempts, colors, etc.) or having specific information displayed.
Game-flow	An event that is caused not by a single specific interaction, but as a consequence of how the game sequences interactions and outputs, and the specific gameplay design of the game.
Content	A *content* event is related to text blurbs and other forms of textual information provided by the game.
Technical error	A *technical error* event is related to a non-intentional glitch in the system that must be corrected.
Non-applicable	When the event is not related to the system and/or not prompted by a system behavior.
Other	An event that is related to the system, but does not match any of the above (this suggests that a new category is needed).

Table 2. Event categories for the user dimension

User-Related Event	
Learning	The user figures out how to perform an action that was unclear before (learn-to-play), or when the user is actively engaging in consuming content (learn-content).
Reflecting	The user pauses or wonders what to do next. Unlike when the user is *confused* and does not know what to do, *reflecting* events indicate pause to create action plans within the game space.
Satisfied/Excited	The user displays a remarkably positive reaction.
Pleasantly frustrated	The user expresses frustration in a positive manner. A *pleasantly frustrating* moment urges the user to try to overcome the obstacle again.
Frustrated	The user voices or displays negative feelings at not being able to complete the game or not knowing how to do something. A *frustrating* moment urges the player to stop playing.
Confused	The user does not know how to perform an action, misinterprets instructions, and/or does not know what he/she is supposed to do.
Annoyed	The user performs properly a task in the game (knows how to do it), but feels negatively about having to do it.
Unable to continue (fatal)	This is usually the consequence of one or more of the above, or of a fatal technical error. An event is related to when the user becomes definitely stuck and/or cannot continue without the help of the researcher. Such events are highlighted because the origin of these events must always be resolved.

Non-applicable	An event is not related to the user (e.g. it is a remark by researcher, or a glitch appeared but the user did not notice it).
Suggestion / Comment	The user verbalizes a comment or a suggestion that is not related to a specific interaction or event.
Other	An event is related to the user, but does not match any of the above (this suggests that a new category is needed).

3.3. Evaluation Process

We present here a step-by-step methodology to assess for usability events in serious games. Additionally we will show as a case study how we employed this methodology to assess for usability while accounting for the MasterMed game's specific learning objectives. According to the requirements described above, the methodology is organized in discrete stages, from the performance of the tests to the final preparation of a list of required changes. The stages of the methodology follow:

1. **Design of the play session.** The evaluation session should be brief and have clear objectives. The designers should prepare a detailed script indicating which tasks the tester is expected to perform. This script should be driven by specific learning goals, as well as include all the relevant gameplay and UI elements within the design. There may be a need for more than one scripted play session to cover all the key objectives.
2. **Selection of the testers.** As noted above, invited testers' characteristics should closely represent the intended users and mimic the context for which the serious game is designed.
3. **Performance and recording of the play sessions.** The testers are given brief instructions about the context of the game and the learning objectives, and prompted to play the game on their own, without any further directions or instructions. The testers are instructed to speak out loud while they play, voicing out their thoughts. During the play session, the evaluator does not provide any instructions unless the user is fatally stuck or unable to continue. Ideally, the session is recorded on video, simultaneously capturing both the screen and the user's verbal and non-verbal reactions.
4. **Application of the Instrument and annotation of the results.** In this stage, the evaluators review the play sessions identifying and annotating all significant events. An event is a significant moment in the game where the user found an issue or reacted visibly to the game. Events are most commonly negative events, reflecting a usability problem, although remarkably positive user reactions should also be tagged, as they indicate game design aspects that are engaging the user and should be enforced. Each event is tagged according to the two dimensions proposed in the SeGUE annotation instrument (Section 3.2). Ideally each play session should be annotated by at least two evaluators separately.
5. **Reconciliation of the results.** Since multiple reviewers should annotate the videos independently, the annotations and classifications likely will end up being different. Therefore, it is necessary for all of the reviewers to confer for reconciliation of the results. There are several possibilities that result from initial discrepant event assessments: (1) An observed event may be equally recognized by multiple reviewers with identical tagging; (2) A single event might be interpreted and tagged differently by each reviewer; or (3) an event could be recognized and tagged by one observer and overlooked by another. In the

latter two cases, it is important to have the all reviewers verify and agree on the significance of the event and have subjective agreement on the proper tag. Most importantly, the objective of this task is not to increase the interrater reliability, but to study collaboratively the event in order to better understand its interpretation, causes and potential remediation actions.

6. **Preparation of a task list of changes.** Finally, the eventual product from this evaluation process should be a list of potential improvements for the game, with an indication of their importance in terms of how often the problem appeared and how severely it affected the user or interfered with the game's educational mission. For each observed negative event, a remediation action is proposed. Changes proposed should avoid interfering with the design and game-play elements that originate positive events to maintain engagement. Users' comments and suggestions may also be taken into account. Quite possibly, some of the encountered issues will occur across multiple users, and some events might occur multiple times for the same user during the same play session (e.g. a user may fail repeatedly to activate the same control). For each action point there will be a frequency value (how many events were recorded that suggest this action point) and a spread value (how many users were affected by this issue).

Finally after reconciliation, the evaluation team should have an exhaustive list of potential changes. For each modification, the frequency, the spread and a list of descriptions of when the event happened for each user all contribute to the estimating of importance and urgency for each action, as it may not be feasible to implement every single remediation action.

It must be noted that although a predefined set of tagging categories facilitate the annotation and reconciliation process, the work performed in stages 4 and 5 can be labor intensive and time consuming depending on the nature and quantity of the test user's verbal and non-verbal interactions with the prototype.

Finally, depending on the scope and budget of the project, it may be appropriate to iterate this process. This is especially important if the changes in the design were major, as these changes may have introduced further usability issues that had not been previously detected.

4. Case study: Evaluating MasterMed

This SeGUE methodology, including the specific annotation categories, has been put to the test with a specific serious game (MasterMed) (see Figure 1), currently being developed at Massachusetts General Hospital's Laboratory or Computer Science. The goal of MasterMed is to educate patients about the medications they are taking by asking patients to match each medication with the condition it is intended to treat. The game will be made available to patients via an online patient portal, iHealthSpace[2], for patients who regularly take more than three medications. The target audience for this game is therefore a broad and somewhat older population that will be able to use computers, but not necessarily technically savvy. This makes it very important to conduct extensive usability studies with users similar to the target audience, to ensure that patients will be able to interact adequately with the game.

[2] https://www.ihealthspace.org/portal_web/public/about.ihs

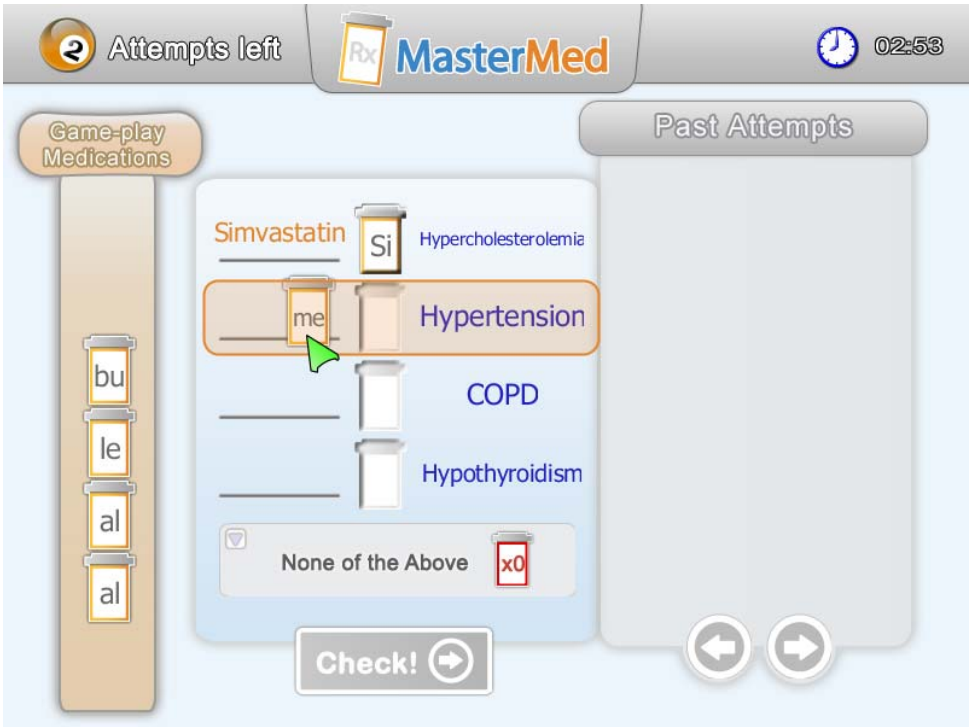

Figure 1. A screenshot of the MasterMed game, version 0.4.5: user dragging a medication to a condition.

Performing an in-depth evaluation of the MasterMed game helped us refine and improve the evaluation methodology, gaining insight into the importance of multiple reviewers, the effect of different user types in the evaluation or how many users and reviewers are required. In addition, the experience helped improve the definitions of the categories in the SeGUE instrument.

In this Section we describe this case study, including the study setup, the decisions made during the process and the results gathered. From these results, we have extracted the key lessons learned on serious game usability testing, and those are described in Section 5.

4.1. Case Study Setup

1. **Design of the play session.** The session followed a script, in which each participant was presented three increasingly difficult scenarios with a selection of medications and problems to be matched. The scenarios covered simple cases, where all the medicines were to be matched, and complex cases in which some medicines did not correspond with any of the displayed problems. In addition, we focused on common medication for chronic problems, and included in the list potentially problematic medications and problems, including those with difficult or uncommon names. As a user progressed through the script, new UI elements were introduced sequentially across sessions. The total playing time was estimated to be around 30 minutes.
2. **Selection of the testers.** Human subject approval was obtained from the Institutional Review Board of Partners Human Research Committee, Massachusetts General Hospital's parent institution. The usability testing used a convenience sampling method to recruit ten patient-like participants from the Laboratory of Computer Science, Massachusetts General Hospital. An invitation email message contained a brief description of the study, eligibility criteria, and contact information was sent out to all potential participants. Eligible

participants were at least 18 years old and not working as medical providers (physicians or nurses). Based on a database query, our expected patient-gamer population should be balanced in terms of gender with roughly 54% of participants female). Patient age ranges from 26 to 103 with a mean of 69.3 years (SD = 12.5) for men and a mean of 70.14 years (SD = 12.75) for women. We recruited five men and five women with their age ranged from mid-30s to 60s to evaluate the game.

3. **Performance and recording of the play sessions.** Each participant was asked to interact with the game using a think-aloud technique during the session. The screen and participant's voice and face were recorded using screen/webcam capture software. The duration of the play sessions ranged between 40 and 90 minutes.
4. **Application of the Instrument and annotation of the results.** After conducting the sessions, a team of evaluators was gathered to annotate the videos identifying all potentially significant events. There were four researchers available, two from the medical team and two from the technical team. Five videos were randomly assigned to each researcher to review, thus each video was processed by two different researchers. In order to avoid any biasing factors due to the backgrounds of each researcher, the assignment was made so that each researcher was matched to each of the other three researchers at least once. The annotations used the matrix described in Section 3.2. Two more fields were added to include a user quote when available and comments describing the event in more detail.
5. **Reconciliation of the results.** The reconciliation was performed in a meeting with all four researchers, where (i) each unique event was identified and agreed upon, (ii) each matched event classified differently was reconciled, and (iii) each matched event with the same tags was reviewed for completeness. This process was crucial in determining the nature of overlooked events and facilitated the discussion on the possible causes for those events that had been tagged differently by the reviewers.
6. **Preparation of a task list of changes.** For each observed negative event, a remediation action was proposed and prioritized.

4.2. Case Study Results

The first outcome of the case study was a set of 10 video files resulting from the screen/webcam capture software. Since the evaluation method was experimental, two randomly selected videos were used for a first collaborative annotation process. This step helped refine and improve the tags described in Section 3.2. Therefore, the final evaluation was performed only on the eight remaining play sessions.

The average play session had been estimated to be around 30 minutes, although most users took between 40 and 60 minutes (and only one user as much as 90 minutes). A total of 290 events were logged. We summarize the events identified for each user (see Figure 2). A *unique* event is defined as when the event was only tagged by one of the two researchers reviewing the video (and overlooked by the other). A *matched* event is defined as when the event was tagged by both researchers and classified equally with the same tags and interpretation. Finally, a *reconciled* event is defined as when the event was identified by two researchers, but tagged differently and then agreed upon during the reconciliation process.

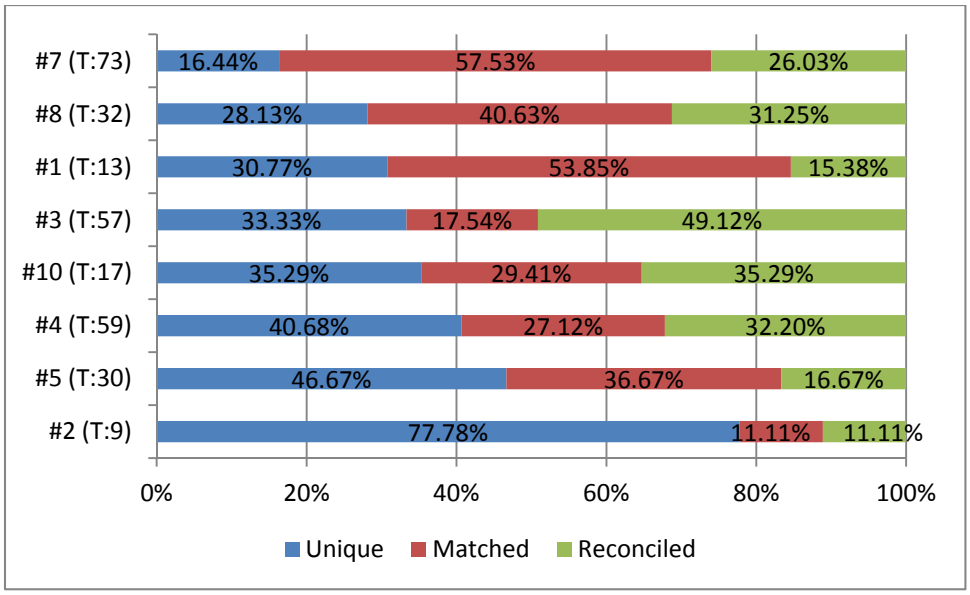

Figure 2. Event statistics. Each bar shows percentage of unique, matched and reconciled events for one of the users. Total number of events of each user is shown in brackets.

In Table 3, we summarize the number of appearances of each tag and the relative frequencies for each event type. The number of negative events (138) was much higher than positive events (46). Also the number of interface and implementation events (179) is greater than events related to design (91).

Table 3. Tag statistics

		Interface			Design		N/A	TOTAL	
		Content	Layout / UI	Technical error	Game-flow	Functionality			
Negative	Annoyed	12	3	4	12		1	32	138
	Confused	9	36	14	19	5	3	86	
	Frustrated	1	3	5	3	2		14	
	Unable to continue (fatal)		1	4		1		6	
Positive	Pleasantly frustrated				2			2	46
	Reflecting	1	3		2	3		9	
	Satisfied / Excited	1	2		9	7		19	
	Learning		2		3	7	4	16	
Neutral	N/A			7			1	8	106
	Suggestion / Comment	28	41	2	7	9	11	98	
TOTAL		52	91	36	57	34	20	290	
		179			91				

Finally, in Table 4 we provide an excerpt of the action points that were derived from the analysis of the results. For each action, we also indicate the frequency (number of events that would be solved by this action) and the spread (number of users that encountered an event that would be solved by this action). Both numbers were used to determine the priority of each action.

Table 4. Excerpt of the prioritized action points list. It shows the type (D:design / I:interface), the frequency (number of occurrences), the spread (number of users affected) and the priority they were given according to these two numbers.

Priority	Action	Type	Freq	Spread
1	Rearrange the tutorials (Shortening and skipping)	D	28	8
2	Remove "None of the above" feature	D	23	8
3	Unify "close dialog" interactions	I	37	5
4	UI tweaking (color schemes, minor layout changes, etc.)	I	22	6
5	Review wording	I	13	6
6	Improve mouse clicking accuracy	I	11	4
7	Improve handout contents (remove unnecessary sections)	I	11	4

4.3 Case Study Discussion

An interesting aspect for discussion is the variability of event statistics across users. Figure 2 is sorted according to the number of unique events, as this category requires special attention. Indeed, while a reconciled event indicates an event that was perceived different by each researcher, a unique event indicates that one of the researchers overlooked the event. In a scenario with only one reviewer per play session, such events may have gone unnoticed. The annotations for some users presented very high numbers of unique events. It is possible that this is related to the total number of events, affecting the subjective thresholds of the reviewers when the frequency of events is high. However, the results do not suggest that a correlation between the total number of events and the proportion of unique, matched and reconciled events. For example, results from users with small total number of events vary, as user #2 presents 77.78% unique events while user #1 has only 30.77% unique events.

Regarding the tag statistics, the number of negative events in the user dimension is clearly predominant. This result may be considered normal, as evaluators are actively looking for issues and pitfalls, while regular play working as intended may not be considered an event. However, the identification of specific positive events was still helpful to identify specific game moments or interactions that really engaged the users in a visible way.

In the game element dimension, the number of events related to the design of the game was significantly less than the number of events related to the interface and implementation (91 vs 179). This data suggests that users were more satisfied with the flow and mechanics of the MasterMed game than with its look and feel. Nonetheless, this difference seems reasonable, as it is easier for users to identify pitfalls in superficial elements like the UI (e.g. font size is too small) then in the design (e.g. the pacing is not appropriate). The correlation between user and system dimensions is also interesting, as positive events are usually related to aspects of the game design. Since the gameplay design is the key element for engagement, this result may be considered an indication that the design was, in fact, successful.

The process to determine the remediation actions and a heuristic assessment of their importance deserves also some discussion. The prioritization of the list is not fully

automatable. While the frequency was an important aspect to consider (an event that happened many times), so was the spread (an event that affected many users). These variables allowed researchers to limit the impact of having multiple occurrences of the same event for a single user. A specific example: the action "remove none of the above feature" was regarded as more important than "unify close dialog interactions" because it affected all users, even though the total number of occurrences was significantly lower (23 vs 37).

Other factors such as the cost of implementing a change or its potential return were not considered, but large projects with limited budget or time constraints may need to consider these aspects when prioritizing the remediation actions.

5. Lessons learned

The result of the case study not only helped to identify improvement points, but also served as a test to improve and refine the SeGUE instrument for annotation. Some design decisions, taken on base of the existing literature, were put to the test in a real study, which allowed us to draw important conclusions. And these conclusions are helpful for researchers using this methodology (or other variations) to evaluate and improve their own serious games. The main lessons learned are summarized below.

Multiple evaluators

As discussed in Section 3.1, different studies have taken different stances when it comes to how many researchers should review and annotate each play session. The key aspect is to make sure that all usability issues are accounted for (or as many as possible).

The inter-rater reliability displayed by the results for our case study is, in fact, very low (Figure 2). Both matched and reconciled events were identified by both reviewers, but unique events were only registered by one of the reviewers. For most users, the number of unique events is between 33% and 50%, giving a rough estimate of how many events may have been lost if only one reviewer had been focusing on one play session (user #2 has an unusually high number of unique events).

This result is consistent with the concerns expressed by White and colleagues [32], and confirms the importance of having multiple evaluators for each play session in order to maximize the identification of potential issues. While it might be very tempting for small-sized teams to use only one annotator per gameplay session to reduce costs, our experience shows that even after joint training the number of recorded unique events is high. Thus, multiple evaluators should be considered a priority when planning for usability testing.

Importance of think-aloud methods

Most observational methods do not explicitly require users to verbalize their thoughts as they navigate the software, as it is considered that the careful analysis of the recordings will suffice to identify usability issues, even with only one expert reviewing each recording.

However, the results from the case study indicate the importance of requesting (and reinforcing) users to think aloud while they play. For our case study MasterMed evaluation

there was a direct correlation between the number of unique events tagged and the amount of comments verbalized by users. While all users were instructed to verbalize their thoughts, not all users responded equally. On one extreme, user #7 was loquacious, providing a continuous stream of thoughts and comments. On the other extreme, user #2 was stoic, apparently uncomfortable expressing hesitations out loud, rarely speaking during the experiment, despite of being reminded by the researcher about the importance of commenting. This had a direct impact in the number of unique events (16.44% unique events registered for user #7 and 77.78% unique events for user #2), as it made difficult for the researchers to distinguish between hesitations caused by a usability issue from actual pauses to think about the next move in the game.

Length of the play sessions

The length of the play sessions was estimated to be around 30 minutes, although the range was 40-90 minutes. During the play session, familiarity with the tool and its expected behaviors may improve, and this may mean that most usability issues would be detected in the first minutes of a play session. To get a better insight about this issue, we produced the event timestamp frequency histogram provided below in Figure 3. Most of the events were tagged during the first 13 minutes of the session (44.06%) after which the rate decreases, with only 24.95% of the events tagged in the following 13 minutes. Beyond this point, the rate slowed even further, even though new, more complex gameplay scenarios were being tested.

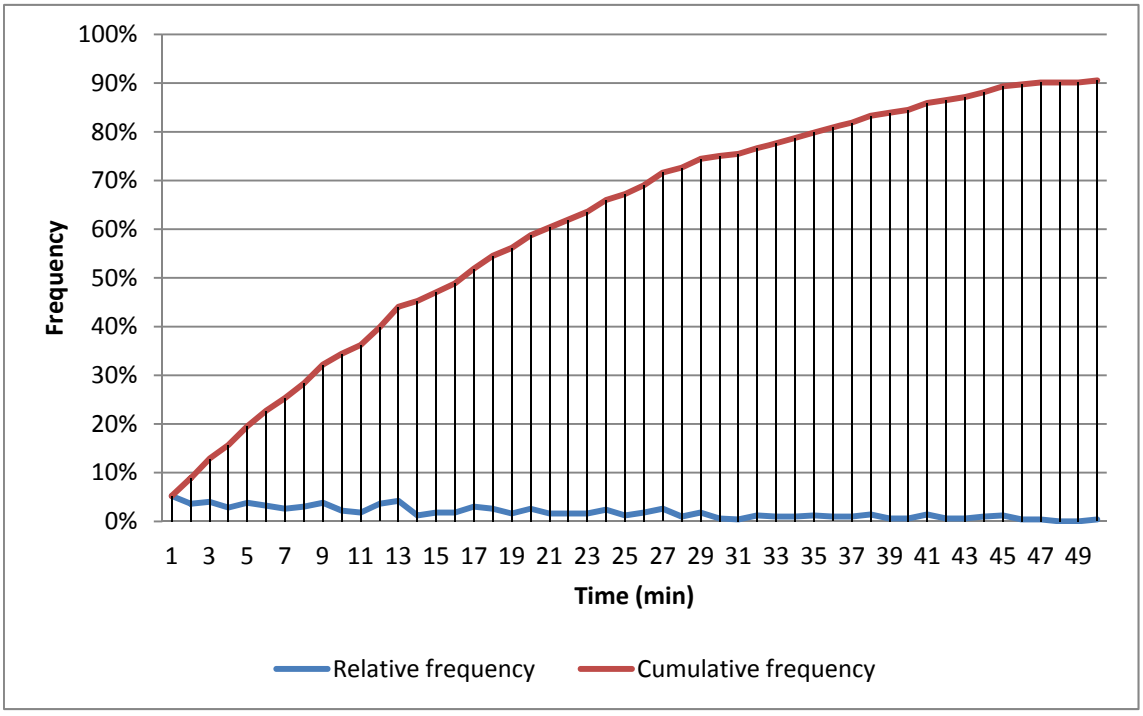

Figure 3. Frequency Histogram of tagged events during MasterMed evaluation. Cumulative frequency graph shows that 40% of the events were tagged within the first 12 minutes.

Users are also encouraged to verbalize their impressions and explain their reasoning when deciding the next move or interaction, but as the play session becomes longer, the users also grow tired. This suggests that play sessions should be kept short and focused. It should also be

noted that researchers observing recorded play sessions thoroughly needed to stop, rewind and re-review video footage frequently to tag the issues encountered, thereby requiring lengthy evaluation sessions. When more than 30 minutes are required to explore all the concepts, different sessions with breaks may be desirable.

Evaluator profile

Even though the proposed methodology called for multiple experts evaluating each play session, we have found differences between the annotations depending on the researcher's profile. The foremost difference was between technical experts (developers) and field experts (clinicians).

Technical issues were one of the main sources of events that had to be reconciled (cases in which both researchers tagged the same event, but assigned different categories). Developers would spot subtle technical issues and tag them accordingly, while clinicians often attributed those events to usability problems related to the UI. This does not necessarily mean that an effort should be made to assign field experts and technicians to review each play session (although it may be desirable). However, it does reflect the importance of having experts from all sides participating in the reconciliation stage. In particular, the goal of the reconciliation stage is not necessarily to agree on the specific category of the event, but on its origin, impact on the user experience and significance, so that appropriate remediation actions can be pursued based on the data gathered.

Limitations

The methodology has a very specific objective: to facilitate the identification of design pitfalls in order to improve the usability of the game. As such, it does not deal with other very important dimensions of user assessment in serious games. In particular, it cannot be used to guarantee that the game will be effective in engaging the target audience or to assess the learning effectiveness of the final product. While the methodology takes care of identifying those elements that are especially engaging, this is done in order to help the designers preserve the elements with good value when other design or UI issues are addressed. Before the final version of the game is released for the general public, further assessment of engagement and learning effectiveness should be conducted.

Another limitation that this methodology shares with typical observational methods (and in particular with think-aloud methods), is that the results are subjective and dependant on both the specific users and the subjective interpretations from the evaluators. The subjectivity of the process was highlighted in the case study in the number of events overlooked by at least one reviewer (number of *unique* events) and the discrepancies when annotating the perceived root cause of each event. While this subjectivity could be reduced by increasing the number of users and evaluators, this increases the cost of the evaluation process. This problem is further aggravated when the process is applied iteratively.

Small and medium sized development projects will need to carefully balance the number of users, evaluators and iterations depending on their budget, although we consider that having more than one evaluator for each session is essential. Similarly, multiple iterations may be

required if the changes performed affect the design or UI significantly, potentially generating new usability issues. In turn, bigger projects with enough budget may want to complement the observational methods by tracking physiological signals (e.g. eye-tracking, EKG, brain activity) to gather additional insight into engagement. However, such advanced measurements fall beyond the scope of this work, which targets smaller game development projects with limited budgets.

6. Conclusions

The design of serious games for education is a complex task in which designers need to create products that engage the audience and provide a learning experience, weaving gameplay features with educational materials. In addition, as with any software product targeting a broad audience, the usability of the resulting games is important. In this work we have discussed the unique challenges that appear when we try to evaluate the usability of a serious game before its distribution to a wide, non-gamer audience. The key challenge is that typical usability testing methods focus on measurements that are not necessarily appropriate for games, focusing on aspects such as high productivity, efficacy and efficiency as well as low variability, number of errors and pauses. However, games contemplate reflection, exploration, variety and trial, and error activities.

While generic heuristic evaluative methods can be adapted to contemplate the specificities of games, observational instruments that generate metrics and scores are not directly applicable to serious games. In addition, observational data is by definition subjective, making it difficult to translate a handful of recorded play sessions into a prioritized list of required changes.

For these reasons, we have proposed a step-by-step methodology to evaluate the usability of serious games that focuses on obtaining a list of action points, rather than a single score that can be used to validate a specific game. Observational methods can be useful in determining design pitfalls but, as we have described in the article, the process is subjective and sometimes cumbersome. The methodology provides a structured workflow to analyze observational data, process it with an instrument designed specifically for serious games, and derive a list of action points with indicators of the priority for each change, thus reducing the subjectivity of the evaluative process.

The Serious Games Usability Evaluator (SeGUE) instrument contemplates tagging events in the recorded play sessions according to two dimensions: the system and the user. Each observed event has an identifiable cause from a certain interaction or UI element, and effect on the user (confusion, frustration, excitement, etc.). The categories for each dimension contemplate aspects specifically related to serious games, distinguishing for example between in-game frustration (a positive effect within the description of games as “pleasantly frustrating experiences”) and at-game frustration (a negative event when the game interface, rather than the game design, becomes a barrier for achieving objectives).

The inclusion of positive events is relevant when studying the usability of serious games. These games need to engage users by both presenting challenges and variability and achieving a learning objective. The events in which the users are engaging intensively with the game

(displaying excitement or pleasant frustration) are important parts of the game-flow, and the action points to improve usability should be designed such that they do not dilute the engagement.

The application of the SeGUE methodology in the MasterMed case study allowed us to draw some conclusions and summarize important lessons learned during the process, as summarized in Section 5. Among them, the experience provided answers to typically open questions regarding observational methods such as (a) the appropriate number of test subjects, (b) number of experts to review each play session, and (c) the importance of the think-aloud technique.

We expect the methodology, the SeGUE tagging instrument, and the summary of lessons learned to be useful for researchers who aim to improve the usability of their own serious games before releasing them. Small and medium-sized projects can use this methodology to test the usability of their games, record data that is typically subjective and difficult to process, and then follow a structured methodology to process the data. The number of evaluation cycles, the specific designs and the aspects of the games that need be evaluated may vary across development projects. Therefore, these steps and the SeGUE instrument might be adapted and/or refined to incorporate any particular elements required by specific serious game developments.

Acknowledgements

This project was funded by the Partners Community Healthcare, Inc System Improvement Grant program as well as the European Commission, through the 7th Framework Programme (project "GALA - Network of Excellence in Serious Games" - FP7-ICT-2009-5-258169), and the Lifelong Learning Programme (projects SEGAN-519332-LLP-1-2011-1-PT-KA3-KA3NW, and CHERMUG 519023-LLP-1-2011-1-UK-KA3-KA3MP).

References

[1] S. de Freitas and M. Oliver, “How can exploratory learning with games and simulations within the curriculum be most effectively evaluated?,” *Computers & Education*, vol. 46, no. 3, pp. 249–264, Apr. 2006.

[2] J. Nielsen, “Heuristic evaluation,” in *Usability Inspection Methods*, vol. 17, no. 1, J. Nielsen and R. L. Mack, Eds. John Wiley & Sons, 1994, pp. 25–62.

[3] M. Kessner, J. Wood, R. F. Dillon, and R. L. West, “On the reliability of usability testing,” in *CHI ’01 extended abstracts on Human factors in computing systems - CHI '01*, 2001, p. 97.

[4] M. Macleod and R. Rengger, “The development of DRUM: A software tool for video-assisted usability evaluation,” in *HCI’93*, 1993, pp. 293–309.

[5] R. J. Pagulayan, K. Keeker, D. Wixon, R. L. Romero, and T. Fuller, “User-centered design in games,” in *Design*, vol. 28, no. 4, J. A. Jacko and A. Sears, Eds. Lawrence Erlbaum Associates, 2003, pp. 883–906.

[6] R. Koster, *Theory of Fun for Game Design*. Scottsdale, Arizona: Paraglyph, 2004.

[7] E. Ju and C. Wagner, “Personal computer adventure games: Their structure, principles and applicability for training.,” *The Database for Advances in Information Systems*, vol. 28, no. 2, pp. 78–92, 1997.

[8] International Organization For Standardization, “ISO 9241-11: Guidance on Usability,” *Ergonomic requirements for office work with visual display terminals*, 1998. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=16883. [Accessed: Oct-9-2012].

[9] A. R. Cooper, *The Inmates Are Running the Asylum : Why High Tech Products Drive Us Crazy and How To Restore The Sanity*. Indianapolis, IN: Macmillan Publishing Co., Inc., 1999.

[10] J. Nielsen and R. Molich, “Heuristic evaluation of user interfaces,” in *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people - CHI ’90*, 1990, pp. 249–256.

[11] J. Brooke, “SUS: A ‘quick and dirty’ usability scale,” in *Usability Evaluation in Industry*, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland, Eds. London: Taylor & Francis, 1996, pp. 189– 194.

[12] J. Kirakowski and M. Corbett, “SUMI: the Software Usability Measurement Inventory,” *British Journal of Educational Technology*, vol. 24, no. 3, pp. 210–212, Sep. 1993.

[13] B. D. Harper and K. L. Norman, “Improving user satisfaction: The questionnaire for user interaction satisfaction version 5.5,” in *1st Annual Mid-Atlantic Human Factors Conference*, 1993, pp. 224–228.

[14] H.-W. Jung, Kim, Seung-Gweon, and C.-S. Chung, “Measuring Software Product Quality: A Survey of ISO/IEC 9126,” *IEEE Software*, vol. 21, no. 05, pp. 88–92, Sep. 2004.

[15] I. Wechsung and A. B. Naumann, “Evaluation Methods for Multimodal Systems: A Comparison of Standardized Usability Questionnaires,” *Lecture Notes in Computer Science*, vol. 5078, pp. 276–284, 2008.

[16] R. L. Boring and D. I. Gertman, “Advancing Usability Evaluation Through Human Reliability Analysis,” in *Human Computer Interaction International 2005*, 2005.

[17] E. L.-C. Law and E. T. Hvannberg, “Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation,” in *Proceedings of the third Nordic conference on Human-computer interaction - NordiCHI ’04*, 2004, pp. 241–250.

[18] P. Moreno-Ger, D. Burgos, and J. Torrente, “Digital Games in eLearning Environments: Current Uses and Emerging Trends,” *Simulation & Gaming*, vol. 40, no. 5, pp. 669–687, Jul. 2009.

[19] J. Kirriemuir and A. McFarlane, “Literature review in games and learning.,” NESTA Futurelab., Bristol, 2004.

[20] R. Van Eck, “Digital game-based learning: It’s not just the digital natives who are restless,” *EDUCAUSE Review*, vol. 41, no. 2, pp. 16–30, 2006.

[21] J. P. Gee, *Good videogames and good learning: collected essays on video games*. New York: Peter Lang Publishing, 2007.

[22] V. J. Shute, I. Masduki, and O. Donmez, “Conceptual framework for modeling, assessing, and supporting competencies within game environments,” *Technology, Instruction, Cognition, and Learning*, vol. 8, no. 2, pp. 137–161, 2010.

[23] C. S. Loh, “Designing Online Games Assessment as Information Trails,” in *Games and Simulations in Online Learning: Research and Development Frameworks*, D. Gibson, C. Aldrich, and M. Prensky, Eds. Hershey, PA: Information Science Publishing, 2007, pp. 323–348.

[24] K. Squire, “Changing the game: What happens when video games enter the classroom,” *Innovate, Journal of Online Education*, vol. 1, no. 6, 2005.

[25] M. P. Eladhari and E. M. I. Ollila, “Design for Research Results: Experimental Prototyping and Play Testing,” *Simulation & Gaming*, vol. 43, no. 3, pp. 391–412, Apr. 2012.

[26] E. Ollila, “Using prototyping and evaluation methods in iterative design of innovative mobile games,” Tampere University of Technology, Tampere, Finland, 2009.

[27] J. A. Garcia Marin, E. Lawrence, K. Felix Navarro, and C. Sax, “Heuristic Evaluation for Interactive Games within Elderly Users,” in *eTELEMED 2011 : The Third International Conference on eHealth, Telemedicine, and Social Medicine*, 2011, pp. 130–133.

[28] D. Pinelle and N. Wong, “Heuristic Evaluation of Games,” in *Game Usability Advice from the Experts for Advancing the Player Experience*, K. Isbister and N. Schaffer, Eds. ACM Press, 2008, pp. 79–89.

[29] W. Ijsselsteijn, Y. De Kort, K. Poels, A. Jurgelionis, and F. Bellotti, “Characterising and Measuring User Experiences in Digital Games,” in *Avances in Computer Entertainment (ACE)*, 2007, pp. June 13–15.

[30] K. M. Gilleade and A. Dix, “Using frustration in the design of adaptive videogames,” in *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology - ACE ’04*, 2004, pp. 228–232.

[31] G. Sim, S. MacFarlane, and J. Read, “All work and no play: Measuring fun, usability, and learning in software for children,” *Computers & Education*, vol. 46, no. 3, pp. 235–248, 2006.

[32] G. R. White, P. Mirza-babaei, G. McAllister, and J. Good, “Weak inter-rater reliability in heuristic evaluation of video games,” in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA ’11*, 2011, p. 1441.

[33] R. A. Virzi, “Refining the test phase of usability evaluation: how many subjects is enough?,” *Human Factors*, vol. 34, no. 4, pp. 457–468, 1992.

[34] J. Nielsen and T. K. Landauer, “A mathematical model of the finding of usability problems,” in *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’93*, 1993, pp. 206–213.