SCOTT WARREN, GREG JONES AND LIN LIN

8. USABILITY AND PLAY TESTING

The Often Missed Assessment

INTRODUCTION

There is little room in today’s educational climate for technologies that do not either accelerate or greatly increase learning (Roblyer, 2005). While 3-D environments, like their game cousins, are motivating and engaging to students (Jenkins, Squire, & Tan, 2003; Tuzun, 2004), there are other educationally sound mechanisms that fit into current time and learning constraints that also achieve the same or better learning outcomes for students. The fact that students spend a lot of time playing games does not mean that the games are based on a sound, efficient and effective instructional design. An examination of several documented games and environments used for learning indicate that many learning games do not demonstrate a sound, efficient educational or instructional design (Dondlinger, 2007).

One major problem with the design of many games is not their ultimate effectiveness; instead, the problem stems from the time required by learners interacting within 3-D environments and games to show improvement in achievement related to formal learning outcomes (Jones & Warren, 2008). A student learning in an immersive multi-user environment often requires an increased amount of time and contiguity in the virtual space to achieve increases in formal learning outcomes when compared with more traditional, face-to-face learning environments.

However, if reported at all, the design processes involved in many serious games involves only three stages: design, development, and research implementation with live players. Especially with younger learners, it is unethical to implement an untested game as the design may lead to negative, unintended consequences such as learners building incorrect mental models of complex systems that are difficult to change. Re-teaching these concepts may take more time than if it simply had been taught correctly in the classroom the first time and it robs students and teachers of valuable time needed to learn other knowledge and skills, which is extremely important given the breadth of content states demand students learn in today’s high stakes, standardized test driven culture.

The questions then are myriad. How do we ensure our game is usable before we put it in classrooms for beta or pilot testing? How can we ensure that our content is correct and our models sound to avoid the aforementioned challenges? How do we ensure that the game is playable on the systems that average schools or universities have available before we start collecting data? This is where usability testing comes into the picture of serious games assessment.

© 2010 Sense Publishers. All rights reserved.
Usability testing is the process of examining those components and processes in a software product that impact the ability of the user to complete successfully those tasks that the designer intends them to be able to complete. As Notess, Kouper, and Swan (2005) note, “(t)he development of effective tasks for usability studies can be strengthened if it is an iterative process. The first draft of a task list is rarely the final product (p. 301).” Real world learning activities are complex. In games, they may be even more so, given the amount of cognitive stimulus from the game tasks and multimedia being pushed at learners. Kuniavsky (2003) notes that usability tasks therefore must “be representative of typical user activities and sufficiently isolated to focus attention on a single feature (p. 270).”

However, Notess et al (2005) also state a caveat to Kuniavsky’s maxim. “Tasks that are too simple misrepresent the complexity of real-world work, but tasks that are too complex may frustrate the test participant, who is already under pressure from the perceived evaluative flavor of participating in a ‘test’ (p. 304).” This is especially problematic in testing for the usability of a game as a researcher may defeat the play and authentic “flow” (Chen, 2007; Csikszentmihalyi, 1991), which would make the outcomes of the usability test inaccurate and lead to unintended design consequences that could cause a game to fail.

Notess et al (2005) go on to note that: “Moreover, crafting tasks to focus on isolated features undermines authenticity because users rarely set out to ‘use features’ but rather to accomplish meaningful work…at the same time, development teams think in terms of features and want to know whether they have made effective design decisions (p. 304–5).” This is a major tension in developing usability tests for serious games that must be addressed. How can we evaluate authentically the user experience with our products without tainting the process which will likely result in skewed or unusable results?

For the development of the Chalk House literacy game (Jones & Warren, 2008; Warren & Jones, 2008), we created a five stage usability process that leveraged both inauthentic and authentic data collection environments and test implementations including both usability lab-based experiences as well as naturalistic use scenarios in order to tackle basic user problems in the sterile environment prior to using the game with children in a classroom environment.

Example of Usability Set Up

There are multiple ways to set up a usability testing lab or naturalistic setting for conducting this type of research. Some are more intrusive while others are less. It is important to minimize the impact of researchers and their data collection tools on participants as was found during the Quest Atlantis project’s Anytown research (Warren, Dondlinger, Stein, & Barab, 2009). During the research phase of the study, there were 16 audio recorders, four video recorders, and, depending on the day, anywhere from 5–9 researchers in the small computer lab observing 26 students and a teacher. This likely had an untoward effect on student use of the game by distracting them or putting undue stress on the players. Further, it was unnecessary
as the volume of data wound up being excessive and therefore a large portion wound up unused.

Instead of such a robust setup, there are simpler ways to set up a usability test to uncover the relevant data one needs to produce an effective serious game. One can either set up a usability testing lab or conduct usability in the naturalistic setting of the classroom. Both have benefits and challenges, but both likely are necessary before putting a serious game out into the world for a large number of users.

When deciding the order in which to test usability, we have found that testing in the lab prior to moving into the field is very helpful as it helps eliminate many usability issues. Some are difficult and others are just a nuisance, such as: server speed; interference with data collection due to unpredictable issues such as students playing with audio recorders; user failure to follow directions when left to their own devices; or the time limits of class periods. However, all can grind your usability test to a halt or make your data useless, requiring additional time and participants that may be difficult to find.

Lab

The usability lab, housed at the University of North Texas in the ThinkTankTwo Project, used a combination of technology to capture fully the interactions during testing (figure 1). Selection of software was dependent upon the computer platform being tested on and product availability. The usability lab produced reports for software and curriculum developers and researchers and outcomes from each stage of evaluation and use will be saved on our secure documentation server. The negative to conducting research in the lab is that it lacks the authentic problems

---

![Figure 1. UNT ThinkTankTwo usability lab.](image-url)
that will be found when implementing the game in the real world classroom. The benefit is that we were able to have several participants run through the software without time restrictions and spend a lot of it asking questions and learning how best to refine the game to improve use.

**Naturalistic Classroom Setting**

To test the usability of the product in the field with all its restrictions and challenges, we also conducted several tests with users in their local classrooms to identify problems with networks, graphics use, school day strictures and other technical issues that face teachers and students in schools. This usability testing in a naturalistic setting allowed us to overcome one of many possible criticisms of usability testing in a lab, which is that it lacks authenticity and therefore inadequately predicts challenges to product use by end users. Figure 2 shows the classroom lab setting in which we conducted our first use test with students and teacher.

Through this process, we discovered large numbers of technical challenges, some of which were unique to individual schools and some districts, and others that were challenges for all teachers in terms of using the software due to widespread policy regarding computer use in today’s public schools. Some of these included the blocking of the port through which the game is delivered, “freezing” of machines that caused the software to have to be reinstalled each time they rebooted, and graphics that showed up on some graphics cards and not on others. Newer challenges have included the release of new operating systems as well as 64-bit architectures that had to be tested repeatedly, in some cases necessitating massive overhauls of the game software or mechanics.

![Figure 2. Classroom lab setting.](image)
USABILITY TESTING IN ACTION

To illustrate what usability testing is through a live experience, we offer an example from our own work with the *Chalk House* game. This was developed for middle school and elementary school learners respectively in order to leverage the motivating power of the game to encourage learners to practice reading and writing skills. The process of developing the game followed the ADDIE model (Bichelmeyer, 2005), which involved six months of initial analysis of the users (teachers, school systems, technology specialists and students), followed by a detailed design documented thoroughly to allow revision based on feedback, a two and ½ year development process to create the content, dialogue, and digital space, followed by usability and alpha testing prior to implementation of the game in schools for evaluation and assessment of learning in response to the product. Before one can understand the impact of the usability process, one must understand the game for which we developed and used it. The following section describes the design and structure of the *Chalk House* literacy game.

*The Chalk House Serious Literacy Game*

The game was developed in the Created Realities Framework, which has similar elements to multi-user virtual environments like Active Worlds and Second Life. It focuses on encouraging students to read and write in middle school by putting them in the role of a reporter investigating the mysterious happenings occurring in the reportedly haunted Chalk House mansion. As students explore, the *Farewell Times* editor requires that student players report the information they have gathered in the form of newspaper stories that give accounts of the mysteries they uncover through unfolding game tasks that require reading comprehension, decoding, and vocabulary expansion. Further, the characters themselves ask questions that challenge student vocabulary and comprehension throughout the game environment, pictured in Figure 3.

*Figure 3. A view of chalk house within the created realities client.*
Within the game environment, players encounter 3-D graphics, 2-D images, and text that either provide direct information about the game tasks or must be deciphered with the help of in-game objects or collaboration with peers. In the spirit of games like Cyan’s Myst™ and The Adventure Company’s Syberia™ series, game information and challenges come in the form of clue objects. These are items such as handwritten notes, links to in-game computers, art, and numeric ciphers that students must read and comprehend to locate keys, open doors, and learn the fates of the missing inhabitants.

**Game/Learning Task**

To complete the learning tasks, students must then retell the stories they experienced through text and images in their own words, writing them in a word processor and submitting them through the system. The system itself drives student play, learning activities, and delivers rapid feedback regarding their progress, helping to replace some of the more onerous teacher functions such as direction-giving and providing a context for learning (Warren, Barab, & Dondlinger, 2008; Warren, Dondlinger et al., 2009; Warren, Stein, Dondlinger, & Barab, 2009). This allows the teacher to spend more time giving students additional or remedial support within the computer lab. Players receive feedback on their work from experts in the form of pedagogical agent (Baylor & Kim, 2005) characters such as the Copy Editor, who comments on their use of vocabulary within their written piece as well as their depth of comprehension of text passages that informed the development of their written piece. Some of the characters are pictured in Figure 4.

![Figure 4. Three characters from the chalk house newsroom.](image)

**Vocabulary and Reading Comprehension Skills**

Reading passages are encountered in the form of character dialogue and interaction with 3-D objects that provide information that can help learners complete game tasks, add detail to their writing, or provide further clarification of advanced vocabulary terms that students must understand to solve puzzles or to write their stories. By having students engage with text that encourages them to engage continuously in a process of textual comprehension and interpretation followed by retelling, our
goal is to help develop automaticity in reading more rapidly than it traditionally emerges (Samuels, 2002; Stewart, 2004).

Because of the high rates of systemic feedback, the student-directed-by-system nature of the game, and the presence of “correct” answers related to reading comprehension, the structure of this game fits most closely with objectivist, cognitive processing approaches to learning. It is because of this that the game diverges from other attempts to develop games for learning such as those completed in science subject areas. The focus on including computer-directed opportunities for engaging in reading practice is substantially different from other work, which used social constructivist, inquiry-based learning (Barab, Sadler, Heiselt, Hickey, & Zuiker, 2006) or problem-based learning instructional design approaches (Warren, Stein et al., 2009).

Achieving Learning Goals

In terms of teaching skills related to the reading comprehension goals, the dialogue texts of various characters, objects, and archival materials in the space requires that students read for multiple purposes such as: 1.) identifying relevant information, 2.) developing a context for their reading, 3.) entertainment, 4.) rereading when comprehension is unclear, 5.) investigating the unknown, and 6.) making predictions and inferences using textual features and imagery such as pictures, paintings, and three-dimensional images. As such, students are required to switch between reading purposes as they work through the learning tasks.

Because the necessary information required to complete each task is embedded within the texts, students must read to decode vocabulary terms, understand multiple meanings of words, and recognize information unique to multiple genres such as news stories and folklore. Without these strategies, they cannot succeed in the game. Student success at learning tasks is also expected to act as an indicator of individual sustained silent reading, which has been difficult to ensure in traditional classroom silent reading periods (Marshall, 2002).

Game Design Mixed with Learning Objectives

With instructional objectives established, the design and development of the learning activities was guided by many game design principles commonly identified in the game design literature. This instructional design incorporates Salen and Zimmerman’s (2003) definition that states that three major design elements are required for a game. These include (a) a rule-based interactive system, (b) a quantifiable outcome characteristic (also known as a win scenario), and (c) artificial conflict and play characteristics. In the case of Chalk House, we also included a fourth element from simulations called the (d) modeling reality characteristic, as the skills students would be practicing in the game would have to mirror real reading and writing skills to help enhance transfer between what they learned in the game world and what they need in the real world. Thus, learners were put in an authentic writing and reading role, which is that of the newspaper reporter and the narrative, while including a few fantastic elements, functions in a manner very close to our own real world.
WARREN ET AL.

CONDUCTING THE USABILITY TESTING

So, now that we have described Chalk House in some of its complexity, a question remains: how can I tell if it will work for my end users? There are a number of components that one must either set up or consider prior to conducting usability testing, especially with serious games. This section will provide an example of these questions as well as provide an example of the phases of usability testing as were conducted with the Chalk House literacy game prior to play testing and widespread use in schools.

Chalk House Usability Tools

In support of usability testing before putting the game content into 3-D development, a web-based 2-D game rapid prototype tool was developed to allow the developers to create, edit, debug, and test learning modules before any 3-D development or graphics were required. A screenshot of the 2-D simulator can be seen in Figure 5.

![Figure 5. CRG 2-D usability simulator.](image)

This usability simulator allowed for testing the game interactions with teachers before designing the 3-D environment, which is important given the high cost of commercial quality graphics. More important, it allowed the users to focus on the
content and instruction. The 3-D graphics we provided early on before changing our underlying game engine usually distracted them with graphics, which they focused on rather than use questions and also may have swayed them affectively in a way that skewed the results, which we discovered through analysis of the qualitative data.

Further, the server is designed to collect detailed information on user interactions for later quantitative analysis, which has been useful for conducting unobtrusive usability testing when noting bottlenecks and stopping points in game play as well as conducting unobtrusive observational qualitative research (Pershing, Warren, & Rowe, 2006). Teachers are able to login into a web-based teacher interface to examine student interactions, activity completions, grade assignments, review quizzes, and examine other progress indicators.

As the Chalk House game included a usability testing phase before being placed into use in a classroom, modifications and corrections could be made repeatedly with a small number of non-middle school students to ensure the product was functional, rather than making corrections during the implementation process or after the summative evaluation. The usability phase has also been useful for allowing the identification of classroom implementation and other use issues with teachers and professionals, which allows these bugs to be identified, documented, and corrected before kids see it. This allows for a much more successful launch of the game with kids than other processes that overlook usability testing.

Usability Phases Conducted for Chalk House

This section describes the five phase usability test series that we conducted with Chalk House over a period of 18 months prior to completing development of the product for widespread use. Through these phases, and by leveraging the use experience of students and teachers, we were able to refine the product in many ways that would not have occurred to us as designers. Ignoring this experience likely would have resulted in a product that no one would have felt comfortable using in their schools as bugs stopped play at various points, text directions were unclear at certain points, and certain components that make users feel comfortable with game play such as loading screens and certain common key stroke commands were lacking.

Usability Phase 1 in usability lab or online with non-teachers, design and development staff. In this phase, usability data were gathered according to evaluation procedures informed by Hix and Hartson (1993). A trained usability tester talked though the software and curriculum materials with each non-teacher, or design and development staff person borrowed from another project who had no direct knowledge of the Chalk House game. Following an introduction to the goals and purpose of the testing, the teacher was given a paper packet that included specific questions about their ability to complete tasks in the Chalk House game as they used the software. Each tester was asked to “talk aloud” into an audio recorder or video camera about the software as they used it, noting frustrations with user interface, directions, keys and layout, visual items, and providing other feedback about their
general ability to complete the learning tasks in the game as mediated by their ability to use the technical components of the game. Use of the product through screen capture using Techsmith’s Camtasia™ or Adobe Captivate™ was collected so that we could match the audio recordings with on-screen video use to corroborate the usability issues with specific benchmarks in the game. This would allow us to make changes or ask follow-up interview questions with users to understand better any confusion or other issues not immediately clear to the developers. After quantitative and qualitative analysis of survey and interview data was complete, the findings were used to revise the game before using it with classroom teachers. For example, we learned through this phase that our lack of using standard graphical elements to denote active and inactive quests led us to mirror the structure of quests from several massively multiplayer online games that learners more instinctively understood.

**Usability Phase 2 in computer labs with end-user teachers.** Reading teachers were then brought into the lab to play the Chalk House game. Usability data was gathered according to evaluation procedures described for Phase One. Different from Phase One, the teachers’ pedagogical and assessment expertise was tapped as we asked them to examine and comment on the evaluation instruments in additional think-aloud protocols. Based on this feedback and the results of the usability surveys, an in-depth, semi-structured interview was also conducted with the teachers. By focusing in this manner, we were able to gain heightened insights into teacher use and feelings about the product and training materials, school restrictions and impediments that would make the product unusable in schools as well as to identify any other major instructional design issues that may have been missed prior to this usability test. The results of this phase led us to many changes; however, the most important was that the interface design as we had it 15 months ago led many teachers to accidentally exit the program. Worse, several teachers could not tell if the program was loading and shut down the program as they though it had crashed due to the lack of a load screen which would show the percentage completed as the game began.

**Usability Phase 3 in computer labs with students.** During this phase, five sixth grade students, in the presence of their parents, were brought into the lab to participate in the testing. A usability tester walked the student through the software and the students were asked to provide feedback throughout by the method of “talk aloud” in the same manner as the first two phases. As in the first two phases, a digital recording of the screen and student question responses was made to show all their actions while using the software. After the use, the students were interviewed and asked to take one or more of the evaluation instruments. The purpose of this evaluation was to gain insight into a student’s use and feelings about the product and the findings were used to refine the game further prior to use with a real class full of sixth grade students. With the students, we learned quickly that many of the quests we had developed lacked adequate direction in the early stages to take new users from one non-player character (NPC) to another and that the previous character quest giver needed to have text that repeated directions to send the player to the
next stage of the quest. Without it, the players, especially those without strong gaming backgrounds, were frustrated and did not want to continue.

**Usability Phase 4 in the classroom and lab.** Usability testing then moved into a real sixth grade reading classroom for this phase. In the lab, the game supplemented several hours of reading and writing practice within the game context, which was a shortened version of the final intervention. This phase specifically examined using the teacher training materials to guide students as they played the game, challenges in implementation of the intervention during the school day, and examined the challenges of including the product as part of their curriculum as students learned to use the product as a learning tool as well as game.

Using the quantitative and qualitative data collected through think aloud protocols with a few students, video observation, in-class researcher observation, and informal interviews, the intervention process was then analyzed to determine gaps and problems facing use within the existing implementation protocol as designated by the teacher guide. After completing analysis, the process was improved so that a **replicable startup and use procedure** would be available in the next phase to test the scaled-up interventions, which would take place in several classrooms over the next several weeks of use as part of the supplementary intervention use of the Chalk House game. Additional usability data, in the form of semi-structured interview questions with both successful and unsuccessful students, also were gathered as available or necessary using checklists to ensure task completion. In addition, we conducted formal interviews with the teacher, technology coordinator, principal, and five children at the conclusion of each use period formatively or at the end of the intervention, summatively, to examine their willingness to use the product in the future. Based on this feedback, the product was revised again before moving to the final phase. Revisions included eliminating several quests to keep the game within the number of hours that the school day and labs permitted and a revision of the graphics quality to drop the polygon count to something low that the school server could handle without overwhelming it. At one stage, teachers could not open their server-based attendance program due to the demand of the online game. Without revision, very few schools would have been able to run the game at all.

**Usability Phase 5 in classroom and lab.** This phase was defined as a full pilot test, using at least three intact classrooms, to be conducted prior to the interventions. The goal was to examine the research questions. Specifically, this phase examined the training materials and product, as it will be used during the full classroom intervention. The summer teacher training for the product also was tested. Teachers, specifically the reading teacher and technology coordinator, were brought into the labs for a teacher-training workshop. Usability testing then moved into a new sixth grade classroom. Just as in Phase 4, data were gathered using the quantitative usability data collection tools previously described as well as several hours of qualitative data in the form of video, student and teacher artifacts from the intervention (i.e. written articles, reflection pieces, reading test scores), and structured and semi-
structured interviews with the teacher, technology coordinator, principal, and ten children. All of these data were analyzed and used to make final refinements to the game prior to scaling the intervention further to allow use in two school districts. Some of these refinements included: changes to the font size and color to make them more visible in many spaces, corrections to feedback systems that made directions clearer to students as they used the system, and further work to drop polygon counts and other non-essential stressors on local server systems to ensure usability at the local school level.

CONCLUSION

In conclusion, we believe that no design and development of a serious game is complete without the inclusion of at least one phase of usability testing, if not multiple phases. In the end, this stringent usability process provided us with detailed feedback on the structural components of the game that often interfere with learning or understanding concepts, and therefore result in the failure of the game. The challenge of taking the time to engage in such a process often leads developers to skip one or more phases before implementing it large scale with learners. However, this causes a number of problems and may cause one’s game to fail with users who are used to the seamless experience they find in their leisure entertainment games like Blizzard’s World of Warcraft or EA’s Madden series, wasting countless hours of design and development because of a poor experience that could have been repaired before the game was released widely.

Further, due to the increasing complexity of game designs, the ease of developing quality interfaces that users like, and the wider availability of games in classrooms, it is ethically important that our product be tested with users to ensure that the product is the highest quality we can produce before it ever winds up in front of target users. This can help ensure that users are not harmed through loss of instructional time, poorly functioning systems that frustrate and fail to deliver instruction, or that inadequately support learner cognition due to underlying design or development flaws.

Finally, by involving the users in the game development, it helps designers and developers see the game through users’ eyes, which is easy to neglect, especially when the designer is highly educated, or worse, arrogant. When bringing a serious game to the end user in education, there are many parties involved, not just the student learner. Without buy-in to your process from the teacher, your game will not find its way into the classroom and your technology specialist will not be purchasing it, making recouping development costs impossible. Without buy-in from parents, principals, or department chairs (depending on your situation and level), you will not be researching your game. Instead, your brilliant serious game may find itself coated in dust at the bottom of the bargain bin in your local game shop. Through usability testing, we can avoid this and produce games people want to play.

REFERENCES


127