Designing an Inclusive Playtesting Process Using Cognitive Load Theory

Elaine Fath[1], Alexandra To[2], Geoff Kaufman[2], Jessica Hammer[2]

Schell Games[1], Carnegie Mellon University[2]

Abstract

Designing transformational games requires a keen understanding of player-specific needs and preferences, informed both top-down by learning theories and instructional design practices and bottom-up by extensive playtesting with learners from diverse demographic groups. When not all players are included in these processes, then the final game risks being impactful for only a subset of players. Too often this means that marginalized and other underrepresented groups (based on factors such as socioeconomic status, race, and gender) are excluded. Through an iterative playtesting process at two sites with different demographic characteristics, we identified playtest design issues to consider that may affect players with high levels of cognitive load, which prior work has shown disproportionately affects players from marginalized groups. We explore how cognitive load issues can arise when making decisions about prototype fidelity, game theming, and replayability. Through this case study of our playtesting process and its impact on our iterative design decisions, we propose methods for how these issues can be mitigated in both playtest design and game design.

Keywords: game design, playtesting, cognitive load, inclusion

Designing an Inclusive Playtesting Process Using Cognitive Load Theory

Games can provide students with a safe, informal space in which to adopt and practice new skills, take on new identities and perspectives, and embrace experimentation and failure (e.g., Akilli, 2007; Gee, 2003; Hayes & Games, 2008; Ke, 2009; Papastergiou, 2009; Squire, 2008). However, if designers hope to make these benefits available to all players, it is imperative that they avoid "one-size-fits-all" solutions and, instead, account for and address the unique needs of all players. This is particularly crucial for players from low socio-economic status (SES) households and other marginalized identity groups such as women or racial minorities, whose orientation toward education can be profoundly influenced by their experience of stigmatization and under-representation.

Some educational game designers are already working to take the needs of specific populations into account by including content that is relevant for players of different cultures (e.g., Khaled, 2008; Thompson, 2014) and grounding their designs in a knowledge of the cognitive development of particular learner groups, including low-SES children (e.g., Wilson, 2009). At the same time, such considerations are still far from the norm, as persisting issues of inclusion and diversity in game design continue to result in games that insufficiently address the perspectives and desires of minority groups (Fron, Fullerton, Morie, & Pearce, 2007).

Cognitive load, which can be a barrier to processing information in a multimedia learning context (Van Merrienboer, 2003), is relevant to inclusive educational game design. Mitigating cognitive load in educational contexts can benefit all learners, but can have a particularly significant impact on low-SES and other marginalized groups, for whom cognitive load is often exacerbated by their unique psychosocial experiences (Molzhon 2016).

Prior work has produced models and frameworks for integrating instructional design principles in games (e.g., Enfield 2012; Westera, 2008), as well as case studies demonstrating the detection and management of cognitive load in games (Kalyuga & Plass, 2009). With few exceptions (Wilson, 2009), however, this research has not detailed how playtesting with target learner groups (and low-SES and marginalized groups in particular) informed iterative design decisions.

The present work addresses this gap by detailing the playtesting and iterative design of *Outbreak*, a transformational board game intended to support players, particularly those from low-SES populations, in developing socio-emotional (SEL) skills around curiosity (To et al., 2016; To, Fath, et al., 2017). These skills include formulating questions about an unknown situation, and being willing to admit ignorance in front of a group (To, Holmes, et al., 2017). We iteratively prototyped and playtested the game with students from both low-SES and high-SES communities over 16 sessions. At each session, we observed gameplay, recorded detailed field notes, conducted post-game interviews following play, and administered post-game measures of players' emotional responses to the game to inform our future designs.

During our iterative design process, we discovered that cognitive load theory explained many challenges in playtesting with low-SES players, and that redesigning both the playtest process and the game with cognitive load in mind improved their experience. We do not claim that cognitive load theory is the only possible explanation for what we observed, nor do we aim to exclude other factors. For example, we recognize that differing cultural frameworks also impact the playtesting process (DiSalvo, Guzdial, Bruckman, & McKlin, 2017). Rather than discount these other perspectives, we hope to show that using cognitive load theory principles as a lens explained observed phenomena and improved our playtesting process.

In this paper, we describe and illustrate our discoveries using artifacts from our game design process, field notes and observations, and other qualitative data from our playtests. It is our hope that through this case study, we might help codify practices of embodying instructional design in game design, using playtesting with learners from key demographic groups to observe the impact of design decisions.

Related Work

In the following section we describe the role of playtesting in the game design process, and examine cognitive load through the lens of game design.

Playtesting and Inclusion

Playtesting is a critical part of the iterative game design process, during which players engage with game materials and provide feedback to the designers about the player experience (Choi et al., 2016). Game design educators recommend that playtesting be integrated into the entire design process, including early playtesting, and playtest results are meant to drive game design decisions (Fullerton 2014). Some game design methods, such as the Tandem Transformational Game Design process, explicitly tie playtest results to the development of theory and the identification of new goals for the game (To et al., 2016). In industry, games user research methods and game metrics are used during playtesting to inform the redesign of game levels and other elements (Ambinder 2011; El-Nasr, Drachen, & Canossa, 2016.). In other words, *who* is included in playtesting influences *how* the game evolves.

Marginalized groups, including low-SES learners, racial minority groups, and women, are often unintentionally left out of the discussion of game design (Fron, Fullerton, Morie, & Pearce, 2007). Although playtesting is central to game design, access to one's intended audience may be limited or absent entirely (Fron et al,, 2007). Some methods for inclusive playtesting exist. For example, Vasalou, Khaled, Gooch, & Benton, 2014 explored how to deal with issues of cultural appropriation in game design through co-design activities (Vasalou et al., 2014), while Gerling & Masuch examine inclusion for the "frail elderly" (2011). However, more typically, games that are designed for marginalized populations report on their outcomes rather than their methods.

One challenge around playtesting is that games can be very different from one another, and designers may have very different goals. A single-player augmented reality game meant to improve physical fitness will have different playtest needs and procedures, compared to a multiplayer entertainment-focused strategic board game. Rather than hard-and-fast methods, designers need to be able to identify the *purpose* of a playtest, match that purpose to appropriate methods, and apply the data to game iteration (Choi et al., 2016). For this reason, *theory-driven* approaches to designing inclusive playtests are needed, so that individual design teams can apply them to their games. In this paper, we explore the value of cognitive load theory to explain challenges faced during inclusive playtesting, and identify ways that it can provide solutions when applied to a particular game context.

Cognitive Load Theory and Games

Cognitive load theory describes the limited reserve of cognitive resources available to working memory for the encoding and processing of new, incoming information (Baddeley 1992, Sweller, Van Merrienboer, & Paas, 1998). According to this framework, during experiences that require active and deliberate information processing, cognition may become "overloaded" with irrelevant, ill-timed, or excessive information, detracting from our ability to attend to and satisfy one's current cognitive goals (Kalyuga, 2009). Cognitive load occurs when information enters working memory from too many channels at once, or at too rapid a pace, or when irrelevant incoming information or simultaneous cognitive tasks deplete resources required for effortful cognition, or even when familiar information is introduced in a new context (Moreno & Mayer, 1999; Paas, Renkl, & Sweller, 2004).

In addition to these context-specific variables, a number of learner-specific factors have been shown to influence the experience of cognitive load. For example, research on stereotype threat has revealed that for students from marginalized groups, the activation of stereotypes about one's identity group creates a state of cognitive load that detracts from available for working memory and interferes with optimal problem solving and decision making in learning contexts (Schmader & Johns, 2003). Repeated experiences with stereotype threat can result in perpetually higher levels of cognitive load, ultimately resulting in challenges to performance, persistence, and identification with learning contexts (e.g., Woodcock, Hernandez, Estrada, & Schultz, 2012). Additionally, recent research suggests that students from low-SES backgrounds may be living with a heavier-than-typical level of cognitive load from the daily realities of dealing with poverty, stress, or trauma (Mani, Mullainathan, Shafir, & Zhao, 2013; Sirin 2005).

In a games context, the "information" that might contribute to cognitive load includes everything that encompasses the play experience - from the rules and mechanics to the game materials to the interpersonal and intrapersonal dynamics of play. Games can be understood as complex multimedia experiences, and, as such, position learning the game and learning *from* the game as tasks that both require the deployment of cognitive resources (Mayer & Moreno, 2003). In this context, learners run the risk of attending more to figuring out the game than to learning the content. This is particularly important for groups already experiencing high cognitive load, as noted above.

Games present challenges around cognitive load because many principles of game design run directly counter to recommendations from cognitive load theory, For example, in the pursuit of increasing player engagement, designers typically employ methods such as adding extraneous elements (e.g., striking visual aesthetics and interesting "flavor text" in rule books and game materials). Moreover, it is common practice for designers to purposefully make relevant information difficult to obtain, partially hidden, or initially ambiguous, with the assumption that uncertainty sustains player interest (Costikyan, 2013). In contrast, a cognitive load approach might suggest that rich mechanics and visuals may be deeply engaging but potentially misaligned to the learning goal of the game, especially when treated as two distinct and separate parts of the game experience (Aleven, Myers, Easterday, & Ogan, 2010). Perhaps not surprisingly, game design frameworks that that have used instructional design principles such as cognitive strategies have revealed challenges in creating engaging educational game experiences (Enfield, 2012).

In this paper, we present our attempt to grapple with these challenges using a game design case study approach. In our design process, we utilized cognitive load theory to identify and streamline mechanics that were making the game less engaging or that were distracting from the game's core learning goal. As described below, our playtesting revealed that these design iterations resulted in higher engagement levels, faster playthroughs, and more strategy use directly tied to our learning goals by our players. The change in outcomes was most drastically observed in playtesters where we might expect the higher amounts of cognitive load.

Outbreak Game Design

Outbreak was designed as a part of the SCIPR project, which aims to design and study transformational games to encourage and increase curiosity through play - particularly for adolescents from marginalized or underrepresented groups in STEM. *Outbreak* was designed using Tandem Transformational Game Design which emphasizes iterations of the game alongside theoretical understanding of its transformational goals (To et al., 2016).

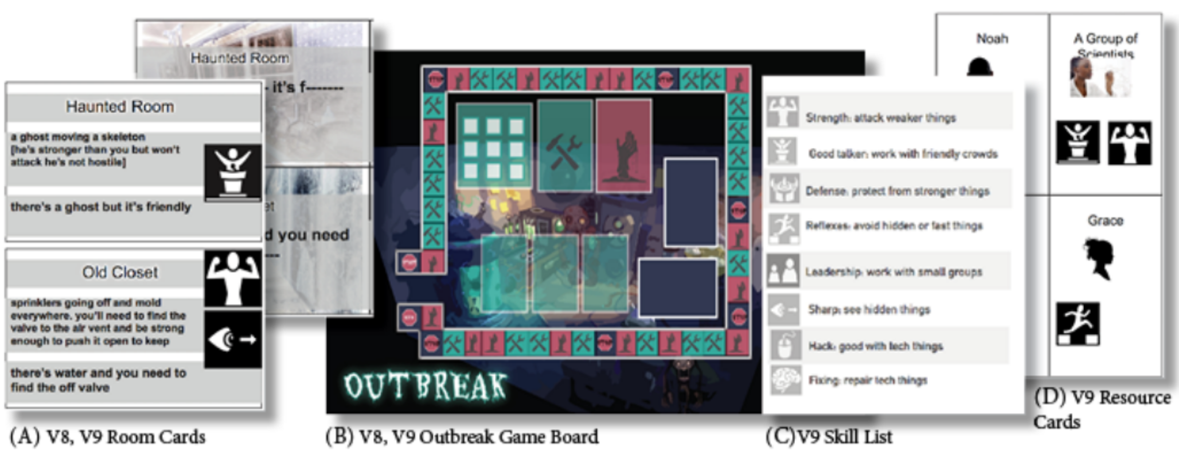

Figure 1. *Outbreak* game with components from V9 including (A) room cards, (B) the game board, (C) the list of skills, and (D) resource cards. Image from (To, Fath, et al., 2017)

Outbreak is a cooperative question-asking game for two to five players, in which the group must collect sufficient antidotes to cure a disease before the game ends. To do this, players search a different room inside the scientist's haunted house each round of gameplay. Most players assume the role of scientific investigators, while one player takes the role of their robot assistant. The robot can safely enter any room, which equates in the game to reading a secret description of the room, including what is necessary to neutralize any threats (e.g, Figure 1A). However, the robot cannot describe what it sees. It can only respond to questions put forward by the investigator players during the timed question-asking phase. Players then use the answers to these questions to decide which of their limited resources (Figure 1D) with related skills (Figure 1C) they will use to overcome the challenge in the room to win the antidotes they need. For example, an angry mob might be calmed by a card that has someone who is friendly, while a card with good defense might neutralize a room with dangerous electrical wiring.

Methods

This paper presents a case study of the playtest and iterative design process for the cooperative tabletop game *Outbreak*, in which we draw themes from the specifics of a given situation (Lazar, Feng, & Hochheiser, 2017). Over the course of seven weeks, we conducted 16 playtests of *Outbreak* with 9-14 year old players at two field sites in Pittsburgh, PA, USA. During this time, the game was iterated to help achieve its transformational outcomes (for further details, see (To, Fath, et al., 2017)). Over the seven weeks, seven versions of the game were prototyped - four of which were deployed in the official playtests detailed in this paper. The playtest methods and game materials were also iterated in order to include the maximum number of players across two very different sites.

Field site one (referred to as the community center) was a summer day camp located in a community center designed to provide local neighborhood youth a safe and enriching environment for campers to enjoy healthy, developmentally-appropriate learning experiences and activities. The camp enrollment is from mid-June to mid-August and students must pay and register to attend. The camp focuses on building self-esteem, social skills, and includes a maker space for STEM learning. Attendees of the camp come primarily from the Homewood neighborhood in Pittsburgh, PA which is a predominantly African American neighborhood (98.3% African American as of 2000 census) and historically has a lower income per capita than both the Pittsburgh city and Pennsylvania state average and a higher unemployment rate.

Field site two (referred to as the science center) was a post-care program for a wide variety of summer science camps located at a local science center. Summer camps at the center tend to only be a week long but some campers enroll in multiple camps. The post-care program is an additional paid program for parents who want to pick up their children from 4:30-6pm when the camps have already completed their activities for the day. The post-care program offers supervised, unstructured play. While the site does not collect demographic information about attendees, a single week of camp (with post-care) costs twice as much as the community center's program cost for the entire summer. We therefore conclude that these families are relatively affluent. Our observations suggest that the racial demographics of the campers are more similar to the overall demographics of Pittsburgh, PA than at the community center (Bureau, 2010).

At both sites players had varying relationships and familiarity with each other. Players were randomly assigned to groups and playtests were scheduled as part of the regular activities of the program (To, Fath, et al., 2017).

In all playtests, we tested the same version of the game at both sites on the same day, which gave us access to qualitative field notes that we could use to cross-compare the same gameplay experience across two groups. Participants played *Outbreak* in groups of three to five, with a researcher taking the role of the robot player. This researcher also obtained consent and taught the rules of the game. An additional researcher was present in the space to take notes. Participants played until they won, they lost, or 40 minutes had passed.

To capture data about participant inclusion in the playtest process, we relied primarily on observation and post-game focus group interviews. Due to the limitations of our field sites, we were unable to record audio or video, as children who had not consented to participate in the study were regularly present. We therefore developed a field notes template that allowed researchers to capture visible emotional responses to the game, unusual player behavior, and the gist of side conversations between players. When possible, researchers noted the game outcome, whether players succeeded in a particular room, and other observations related to playability and balance (To et al., 2016). To study the game's transformational outcomes, we also captured questions asked by players and used a customized valence-arousal measure to connect player emotional reactions to specific moments of the game (To, Holmes, et al., 2017). In the focus group interview, participants were told that their feedback would be helpful in aiding the game designers working on the game to change the game and make it better. They were asked what they liked most about the game, what they would wish to change about the game, and for any other additional feedback they'd like to share. While other elements of the playtests changed in response to our observations about player inclusion (see below), these elements were held constant across the entire study period.

The data captured represents a diverse range of playtests. Some participants played the game only once, while some played multiple times over several weeks; playtests occurred in a range of physical locations from a formal lab setting to a cafeteria in a science center; and some players played multiple versions of the game. Although we had quantitative data, for example in the valence-arousal maps, we felt it was inappropriate to perform formal statistical analyses across groups. Instead, we used qualitative methods to understand participant behavior.

Following each playtest, the entire team (i.e., playtesters, game designers, researchers) met to discuss common themes and look together at the data in an exploratory and informal manner. Cognitive load was identified early in the design process as a promising and relevant theory, but only as it broadly might pertain to the design of educational materials. After our second playtest, we identified cognitive load during a post-play debrief as being relevant to the specific design of game materials such as length and complexity of text on game cards. From that point forward we added analyzing cognitive load of the game to our observations and post-play debrief sessions, including examining play speed, rules comprehension, changes in the self-report measures, and the use of "better" (i.e., more strategically helpful) questions during the question-asking phase of game play.

Using those themes and informal data discussion, the team then also participated in an open brainstorm session for changes that might improve the game. Finally, the game designers would generate a master changelist from this brainstorm and start prioritizing potential changes based on their ability to increase playability for our target audience, the relative ease of implementing the change, and which specific overarching design goal the change would address.

Results and Discussion

Based on our qualitative field notes, as well as learning and engagement measures from playtesting, we believe that cognitive load theory can meaningfully inform the playtest process. We observed that using cognitive load theory when designing playtests and interpreting playtest data can be more inclusive of children from a range of backgrounds, particularly those who may struggle with increased cognitive load. In the following section we discuss three areas of consideration: prototype fidelity, theming, and replayability. Table 1 summarizes what sort of playtesting feedback may be tied to each of these three areas, along with the related Cognitive Load principles discussed for each.

Table 1
Examples of playtest feedback that teams may receive, and which Cognitive Load Theory Principles may help address and resolve that feedback.

Playtest Feedback	Potential Area of Focus	Cognitive Load Principle(s)
Sporadic levels of interest that differ widely between players. Voiced frustrations about game material quality or organization. Random rather than strategic choices made during gameplay.	Prototype Fidelity	Essential & Incidental Processing (Mayer & Moreno, 2003) Coherence and Signaling Principles (Mayer & Moreno, 2003) Strategic Choice and Mental Processing Resources in Games (Kalyuga, 2007)

Players discussing game theme/story in unexpected ways or not at all. Players discussing rules in incorrect ways that relate to the story. Players focusing on irrelevant information as though it mattered.	Theming	Extraneous Information (Mayer & Moreno, 2003) Information Channels (Moreno & Mayer, 1999; Paas, Renkl, & Sweller, 2004)
Players who have played once before can play much more quickly. Players who play the game multiple times rate it more highly than first-time players.	Replayability	The Pre-Training Principle (Mayer & Moreno, 2003; Kalyuga, 2009) Segmenting Principle (Mayer & Moreno, 2003) Working vs Long-Term Memory (Baddeley 1992)

Prototype Fidelity

A key principle of iterative game design is to get prototypes into testing as early as possible, and to iterate many times during the design process (Fullerton 2014). These early, rapid prototypes usually focus on mechanics and gameplay rather than visual design or production quality. High-quality materials can detract from players' ability to rate gameplay on its own merits, and so are often left for later stages (Martin & Hanington, 2012, Schell, 2014). However, we found that low-fidelity prototypes come with their own set of challenges around engagement, learning, and even the ability to play the game - but only for some players.

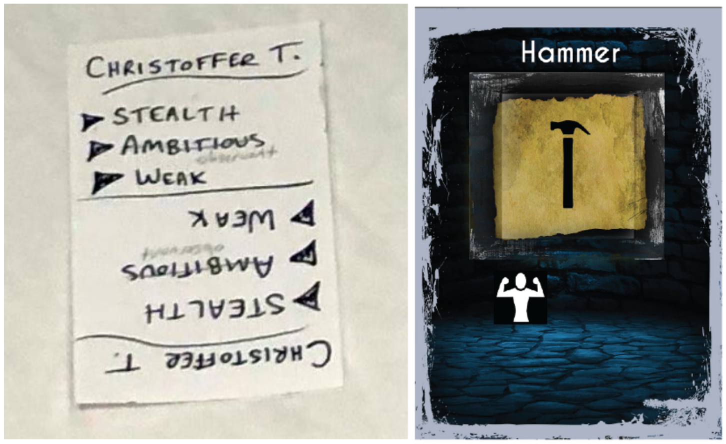

Figure 2. Initial resource card prototypes written on white card stock with black permanent marker [left] and final resource cards printed on playing cards in color [right] from *Outbreak*.

We deployed the same early prototypes at both research sites: black-and-white printouts, accompanied by hand-drawn cards (Figure 2). At the science center, players rated early playtests more positively in the post-game measures, asked more strategic questions during gameplay, and completed the game more quickly (on average, 20 minutes vs. 1 hour) than their peers at the community center. For example, most players at the science center ranked 9-10 of the 11 key game moments as having been positive, joyful, or “gripping” (i.e., afraid but interested), while responses from players at the community center were more distributed and sporadic. Most community center player responses were ranked in the “bored" area while others put a disproportionate number as “gripping,” with few positive responses. In other words, science center players were more consistently interested and interested in the same things as their fellow players, compared to their counterparts at the community center who were generally more disinterested or reported more negative emotions.

Because the problems appeared in player interactions with mechanics, our first hypothesis for explaining this difference was that the game mechanics were cognitively taxing for our community center players. This would imply that the *essential processing*, or the basic work required to make sense of game activities, was too high. For example, in *Outbreak*, essential processing includes taking turns, drawing cards, asking questions, and eliminating choices. However, our observations did not align with this hypothesis. Rather than critiquing the game *activities* during play or in post-game interviews, players criticized the game *materials*. Our field notes report, for example, that community center players were "frustrated by the fidelity of the game" and "bothered about cards and board low-fi."

If the problem was the fidelity of the materials, we could investigate this possibility by improving the quality of our prototype. We removed our paper-printout board (Figure 3 [left]), backed all materials with cardstock or cardboard, (Figure 3 [right]), and printed everything in color. We made no changes to game mechanics or core interactions. After deploying the new version at both sites, we observed that players at the science center did not change their play behavior. However, players at the community center played in ways that were similar to science center players, rather than similar to our previous community center playtests. They completed game loops more quickly, and asked more directed, productive, and strategic questions.

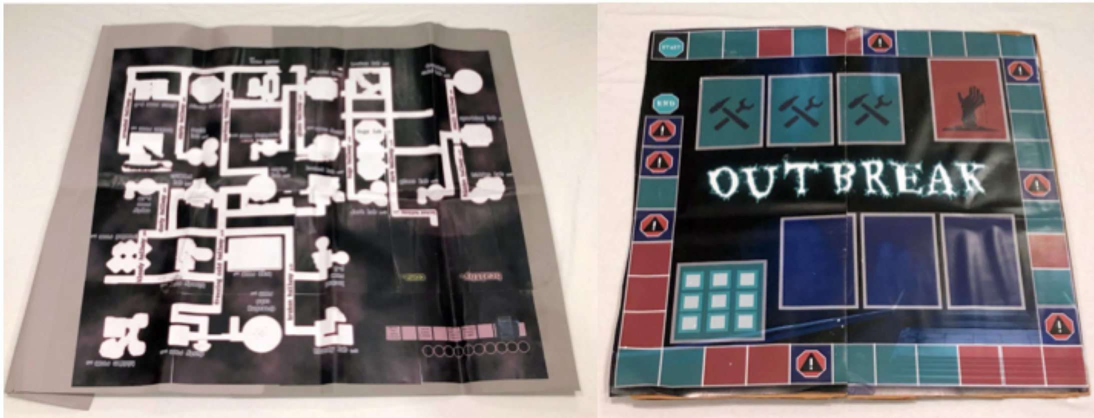

Figure 3. A low-fi game board primarily printed in black and white on construction paper with a fixed map of rooms [left] and a later more high-fi colorful game board with a track backed with cardboard [right].

Players' difficulties completing game loops can be explained by the concept of incidental processing (Mayer & Moreno, 2003). Incidental processing is what happens when non-essential media material requires cognitive processing, such as when music plays during the narration of a video. Many games make use of this idea to increase the challenge level. Nintendo's *Mario Party* minigames, for example, often work on the concept of distraction. They draw the eye with distracting "red herring" movement of non-essential game components. However, in this case, the additional processing required by our less polished game materials impaired players' ability to play the game at all. Particularly for players who may already experience high cognitive load for out-of-game reasons, low-fidelity prototypes may cause an incidental processing "tipping point." At this moment, players are no longer able to interact with the core of the game and the game materials feel overwhelming.

This theory also explains why we observed lower-quality questions at the community center with our lower-fidelity prototypes. Cognitive load theory predicts that cognitive overload leads to random search behaviors, as strategic choice is too complex for the mental processing resources available (Kalyuga, 2007). We witnessed random search behavior at the community center with our low-fidelity materials. At the time, it seemed unrelated to board fidelity: "[players] want to ask open-ended, storytelling questions, not to optimize skill quality questions ('do we need to be fearless?', 'Would I be scared?')." However, the issue was resolved by improving the fidelity of the material, which we hypothesize freed up cognitive resources for asking strategic questions.

Creating higher-fidelity prototypes has its own risks. They are more expensive and time-consuming to produce. The added time slows the iterative design process. Game materials that are too high-fidelity will also throw up false positive feedback. Players want to please the interviewer about a game that is perceived as closer to done (Schell, 2014). There are also known psychological effects related to materials fidelity. Users will rate good-looking screens as better, even if the content is identical (Martin & Hanington, 2012). But audiences with heavier-than-average cognitive load burdens may need high-fidelity materials to provide useful playtest data. Comparing data from a low- and high-fidelity prototypes of the same version of a game may help identify whether the issues are from core mechanics or incidental processing of game materials.

Theming

Game theming is a central narrative concept that ties the entire game together, which can be conveyed through imagery, text, or other materials (Schell, 2014). Theming invokes *schemas*, or sets of related ideas that provide meaning and structure to game activities. For example, in a game about pirates, players might expect loot, exploration, or double-crossing. While each person may not come to the game with the exact same expectations, theming can help players understand mechanics (Rosewater, 2012).

At the same time, cognitive load theory tells designers to reduce "extraneous" information. Extraneous information is information that is interesting but irrelevant (Mayer & Moreno, 2003). For example, in a video about lightning, a designer might remove interesting images of objects damaged by lightning that have no pedagogical value. While it may be relatively easy to tell what is extraneous information in an educational video, it is less easy to understand what is extraneous in a game. For example, flashing lights and a shower of stars might tell players they have obtained a reward. Is this core to the game experience, creating a "juicy" user interaction (Swink 2009), or is it extraneous to the core interactions of the game?

We discovered the challenges of theming early in our design process. We identified key concepts we wanted to signal with our choice of theme, such as cooperative play, expectations of failure, and exploring the unknown. We originally selected a zombie theme because it involves small groups of survivors exploring an unknown environment, at great risk to themselves, under pressure from the zombie hordes. However, when we brought this version of the game to our players, we found that players at the community center were disengaged from the theme. Players at this site did not talk about the story during the game or interviews, except using words that were directly in the instructions. Confusion about zombie horror tropes was also revealed in questions such as "Would zombies be good?" If community center players struggled with the theme, we suspected that it might function as extraneous cognitive load for them, rather than as a cognitive scaffold. By contrast, science center players seemed familiar enough with zombie tropes that they were correctly strategizing to theme (e.g., we have to save people quickly before they succumb to infection or attack).

Figure 4. A collection of the codesign materials used to identify which aesthetics seemed mature, creepy, and fun. Description words were placed on about 20 pictures with different color filters.

To understand why we were seeing this effect, we turned to culturally relevant pedagogy (Irvine, 2010), which encourages designers to think about how a learner's pre-existing cognitive schema might differ from their own. Game designers are not exempt from this challenge, as the design principles of games have emerged from a limited market and are hardly immutable (Fron, Fullerton, Morie, & Pearce, 2007). In our case, we turned to the literature and discovered that zombie survival stories are primarily white middle-class fantasies that allow them to imagine what "roughing it" would be like (Preston 2010), and the players at the community center would likely have much less exposure to this genre. We cross-checked this by interviewing our own players and found corroborating results. While players at both sites knew about zombies, we found differences between the sites when asked to describe a zombie apocalypse. The stories of participants at the science center were richer, more detailed, and more descriptive than those collected at the community center.

Figure 5. Image of a codesign activity to identify scary stories our target audience was familiar with. "A girl like Bloody Mary who is trapped here until she can find friends to help her out of the room"

To develop a theme that would connect better with players, we turned to co-design (Sanders & Stappers). We retained our core concepts of cooperation, failure, exploration, and the unknown, but worked with participants to identify a theme that would better evoke those concepts. We asked players to place adjectives onto different pictures describing their opinion of them (e.g., Figure 4) We also asked players to create an artifact about a spooky story, after a short discussion of spooky stories and urban legends they'd heard (e.g., Figure 5).

We decided to re-theme the game to “haunted house" after hearing that concept come up the most frequently when discussing failure, exploration, and the unknown. We then ran another session where we collected a list of the threats players expected to see inside of a haunted house:

body horror (disembodied hands), haunted dolls, monsters, angry animals, and ghosts. These elements were then incorporated into the game.

We observed that the new theme was successful at both sites. Players spent time talking about the haunted-house themed flavor text, talking about the characters and what they were like, and talking about the game scenario as it related to the skills on the cards. However, this came at a price. We noted players would often read an entire card out loud, including flavor text, and then become confused as to what was relevant versus irrelevant information. We also observed players extrapolating solutions from the flavor text or imagery when in fact the text was there for humor only. This effect was more pronounced at the community center, to the point where it significantly affected gameplay. For example, we observed the strongest readers often adopted the role of the board manager, managing the essential processing that the other players were unable to handle, such as keeping track of turn-taking or of hidden information. We hypothesized that we might be seeing a channel mismatch. The core activities of our game included reading and speaking (e.g., asking questions); putting flavor information into text would conflict with these core activities, and the effects of the additional cognitive load would become visible for the players where reading itself provided a substantial challenge.

To address this issue, we experimented with removing much of the non-essential textual descriptions, and simplifying the images down to icons (Figure 6). We were concerned that these choices would lead to player disengagement, because we would no longer be conveying our theme effectively. However, we found that the simpler materials produced higher engagement from the table as a whole and faster turns. Players were also more likely to succeed at the game (i.e., completing rooms and receiving antidotes). Player engagement with the narrative remained, driven largely by channels such as the detailed box art (which situated the game inside of a haunted house) and the introductory story, which was now more aligned to a story most participants were familiar with.

Figure 6. Resource cards with flavor text and complex imagery [left] and simplified resource cards with fewer words [right].

While theming and relying on schemas can be a useful tool to reduce cognitive load in players, the risks in this approach are twofold: over-generalization of audiences and variance in players' pre-existing knowledge. In just a handful of interviews it can become easy to quickly overgeneralize for a given audience. To mitigate this risk, when designing for audiences whose prior knowledge and experiences differs from the designers', story should likely not carry the torch. If it does, test the story early and often. To mitigate the risk of variable audience prior knowledge, it can be helpful to create an “assumption map" of what the game design team's expectations are (regarding theme and story) before a playtest in order to explicitly test those assumptions (for more on assumption maps, see Martin & Hanington, 2012.

Replayability

Game designers approach replayability in games in a variety of ways, such as procedural generation, randomness, or emergence. The board game *Settlers of Catan*, for example, relies on a randomly-created tile map for different gameplay each time. Replayability can also come from players making different choices in different playthroughs, as in *Dragon Age*'s romances, or from players repurposing and revisioning the game, as in speedrunning.

Replayability in games creates challenges for playtesting. On the one hand, it is important to diversify the number of players who are testing a game, to avoid over-fitting the game's design to the preferences or needs of a specific group. On the other hand, cognitive load theory argues that playing a game multiple times frees up cognitive resources because some of the information has been committed to working memory, also called the pre-training effect (Kalyuga, 2009).

In our playtest process, we addressed this issue by having some players return from previous playtest sessions, playing alongside new players. This simulates a typical board game play situation where at least one player has read the rules beforehand, or has previously played. This also allowed us to have some players reflect on the differences between game versions, while still learning about the effects of new versions on new players.

Early versions of *Outbreak* used a fixed map with a limited number of rooms (Figure 3). We found that this created issues with replayability. For example, players would learn how to game the system. Once it became clear to the players that some questions were particularly useful for successfully clearing a room, they began to ask the same questions over and over again without variance. Once a single player discovered this dominant strategy, all players would adopt it. We were concerned that this dominant strategy would make further playthroughs less engaging for players. Instead, however, we saw a positive effect on engagement. This was particularly evident at the community center, where players were often participated less in their first play of the game. Once the dominant strategy came into play, players at the table would remind one another of turn order and other rules, and success rates for turns went up and players more frequently won the game.

We interpret this evidence using the pre-training effect. As players were getting better at playing the game, they had more cognitive resources for other kinds of activities, like helping one another or checking their actions against the rules (which, being held in working memory, were easier to access). In our redesign, therefore, we aimed to redirect more of those resources at the learning goals of the game, namely asking questions. But, we wanted to keep the useful outcomes of the pre-training effect, particularly for our community center players.

To accomplish this, we reframed what we had been calling a "dominant strategy" as an interim plateau. We expected that by their second or third playthrough, players would discover a particularly effective set of questions, as we indeed observed at both our sites. However, we could introduce a new level of expert play by offering high-risk, high-reward moves for players who felt comfortable with the basic mechanics. Players were given the ability to select rooms with varying difficulty (i.e., rooms have a displayed signal of level 1, 2, or 3 difficulty where level 3 offered the most risk - three threats - for the most reward - three antidotes). Additionally, we built in some cards that were near-replicas of others, with identical room titles, in order to make memorization of the actual solution less likely. After these changes, we found that players did engage in discussion over choosing risky vs. safer rooms. We did not see observable differences between community center and science center players regarding these conversations.

Understanding the pre-training effect helped us playtest in a way that was aligned with our eventual goals for deployment. While first-time players needed to have a good experience, we wanted the game to be played multiple times. We also knew that the game might be played in different contexts, (e.g., at home, at school), and with different sets of players, (e.g., a single expert, a group of experts, or all first-timers). That meant we needed to understand how new players played, to be able to identify when the game rules and procedures had been internalized through repeated play, and to observe gameplay afterwards as well.

We note that in order to align with our planned deployment strategy, we needed to know what that deployment strategy would be. If the game has a "legacy" model, such as *Risk Legacy*, which adapts as it is being played over time but can only be played once, that demands a different playtest strategy from a game that uses remixing to present a randomized, new puzzle every time and where players will encounter a small subset of all possible puzzles no matter how long they play. Answering these questions is particularly important for classroom deployment. One in three game-using teachers feel that not knowing *how* to use games in the classroom is a barrier to using them more often (Takeuchi & Vaala, 2014), and supporting materials need to take into account the fact that multiple playthroughs will drastically impact the learning outcomes. This, too, is also an inherent risk with the strategy of multiple play-throughs: deployment to classrooms is already a difficult problem (Klopfer et al., 2009). A game that requires repeated playthroughs may present additional barriers to deployment and evaluation, which may be mitigated by such steps as sending the game home to be played or allowing the game to be an independent station choice after its initial introduction.

Table 2
Team Iteration Strategies Based On Cognitive Load Principle

Area of Focus	**Cognitive Load Principle**	**Ways to Iterate**
Prototype Fidelity	Essential & Incidental Processing (Mayer & Moreno, 2003)	Improve the quality of game materials without altering mechanics. Run an additional playtest to evaluate

	Coherence and Signaling Principles (Mayer & Moreno, 2003)	differences in how strategic players are being with choices. Alternate tests occasionally between low- and higher-fidelity materials. Move relevant images and text on game materials closer together.
Theming	Extraneous Information (Mayer & Moreno, 2003) Information Channels, Coherence Principles (Moreno & Mayer, 1999; Paas, Renkl, & Sweller, 2004)	Use principles from Culturally Relevant Pedagogy and Codesign (Sanders & Stappers) to identify what themes will work with players' pre-existing cognitive schema and which are unfamiliar. Explore ways to convey theming that are used at different times or on different channels from the channels used to play the game.
Replayability	The Pre-Training Principle (Mayer & Moreno, 2003) (Kalyuga, 2009) Segmenting Principle (Mayer & Moreno, 2003)	Consider deployment context early. How often and where will players play in the wild? If players may only play once, how can you reduce the information needed to understand the game one time? Test at least some groups of playtesters with at least one person who has played previously. Compare differences. If the game has a dominant play strategy, does that strategy fit with what you want players to master? If not, consider removing it. If so, what mechanics can you introduce later on to compel players who have mastered the first dominant strategy and need a higher challenge?

Limitations

In this paper, we present a case study for a single game, *Outbreak*. This choice allows us to explore multiple design iterations and solution spaces to problems with cognitive load, described in Table 2. However, because it is a case study, we cannot fully rule out alternative explanations for the phenomena we understand as evidence of cognitive load - nor do we intend to. As noted earlier in this paper, we believe that cognitive load theory provides *one* useful lens for improving the playtest process, not the *only* useful lens for doing so. While cognitive theory principles need no further verification, as they have been extensively researched, we have not chosen to compare their impact on the playtest process with the impact of potential alternate explanations. While we find the evidence presented here persuasive, additional studies on this topic are needed.

The case study design also limits the generalizability of this work. For example, some of our findings are specific to non-digital games (e.g., backing early prototypes with cardboard), and we expect that there are additional tensions that emerge when designing in other genres (e.g., mixed-reality games, competitive games) with cognitive load in mind. The value of this work is to expose some of the ways in which game design practice can come into tension with instructional design theory, and to suggest ways that designers can move forward in their own projects.

Another limitation of this project is that we do not directly link design improvements to learning (transformational) outcomes. The data analyzed in this paper is from design iterations, observations, and qualitative field notes, not pre-post tests or other measures of learning gains. It is possible that our metrics of design improvement (e.g. playability) would have unexpected effects on the learning goals of the game. While we consider this unlikely, we cannot rule it out.

Finally, while our subjects were drawn from multiple sites and demographic categories, they all resided in the same mid-sized American city. While they come from a range of cultural backgrounds, they also have significant cultural similarities. Similarly, while the designers in the project were trained in a range of places, all received academic game design training, and most had professional work experience in North American small-to-medium-size game companies. It is possible that their knowledge of the state of the field is incomplete.

Conclusion

In this paper, we develop a theory-driven approach to inclusive playtesting and describe how it was implemented in the iterative design process of a cooperative board game. Players drawn from marginalized groups may experience increased cognitive load during playtesting for a variety of reasons, such as stereotype threat. Cognitive load theory has implications for a range of game design and playtest design processes, including prototype fidelity, the degree and type of game theming, and how many times playtesters encounter the game.

Our work contributes to improving the game design process both by making playtest design more inclusive and by providing designers with a lens to use for making decisions about iteration post-playtest. Researchers can also use this work to understand the game-specific challenges of working with cognitive load theory. On the surface, cognitive load theory is in tension with with game design. To completely reduce cognitive load in a game, researchers might aim to create a game with no extraneous detail, theming, or uncertainty - and also no fun. By looking at the specific tensions between game design and cognitive load, designers and researchers can address these tensions with meaningful intention.

In our future work, we hope to explore the lens of cognitive load theory as it relates to other types of games. The value of working deeply with one game as an exemplar helps us explore cognitive load theory within an iterative design process. By exploring cognitive load in the playtest process for additional games, we can identify additional factors about the games themselves, such as how they are intended to be deployed, that would affect the way cognitive load theory manifests.

We could also think about generalizing beyond cognitive load theory, and developing other lenses for analyzing playtest processes. Just as Schell's lenses help designers look at their game in many different ways, to break habits and increase creativity, a book of lenses for playtesting could help designers - not just trained researchers - conduct more inclusive and rigorous iterative design processes.

Acknowledgements

Thanks to the SCIPR (Sensing Curiosity in Play and Responding) project and co-PIs Justine Cassell, Jessica Hammer, and L.P. Morency, generously funded by the Heinz Family Foundation. We also thank Anny Fan, Catherine Kildunne, and Eda Zhang for their work iteratively designing and playtesting *Outbreak*.

Reference

Akilli, G. K. (2007). Games and simulations: A new approach in education.

Aleven, V., Myers, E., Easterday, M., & Ogan, A. (2010). Toward a framework for the analysis and design of educational games. In *Digital Game and Intelligent Toy Enhanced Learning (DIGITEL), 2010 Third IEEE International Conference on* (pp. 69–76).

Ambinder, M. (2011). Biofeedback in gameplay: How valve measures physiology to enhance gaming experience. In *game developers conference* (Vol. 2011).

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Bureau, U. C. (2010). Pittsburgh PA census map.

Choi, J. O., Forlizzi, J., Christel, M., Moeller, R., Bates, M., & Hammer, J. (2016). Playtesting with a Purpose. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play* (pp. 254–265).

Costikyan, G. (2013). *Uncertainty in games*. Mit Press.

DiSalvo, B., Guzdial, M., Bruckman, A., & McKlin, T. (2014). Saving Face While Geeking Out: Video Game Testing as a Justification for Learning Computer Science. Journal of the Learning Sciences, 23(3), 272-315. doi: 10.1080/10508406.2014.893434

El-Nasr, M. S., Drachen, A., & Canossa, A. (2016). *Game analytics*. Springer.

Enfield, J. (2012). *Designing an educational game with ten steps to complex learning*. Indiana University.

Fron, J., Fullerton, T., Morie, J. F., & Pearce, C. (2007). The Hegemony of Play. In *DiGRA Conference*.

Fullerton, T. (2014). *Game design workshop: a playcentric approach to creating innovative games*. CRC press.

Gee, J. P. (2003). What video games have to teach us about learning and literacy. *Computers in Entertainment (CIE)*, *1*(1), 20.

Gerling, K., & Masuch, M. (2011). When gaming is not suitable for everyone: Playtesting wii games with frail elderly. In *1st Workshop on Game Accessibility*.

Hanington, B., & Martin, B. (2012). *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport Publishers.

Hayes, E. R., & Games, I. A. (2008). Making computer games and design thinking: A review of current software and strategies. *Games and Culture*, *3*(3–4), 309–332.

Irvine, J. J. (2010). Culturally Relevant Pedagogy. *Education Digest: Essential Readings Condensed for Quick Review*, *75*(8), 57–61.

Kalyuga, S. (2009). Instructional designs for the development of transferable knowledge and skills: A cognitive load perspective. *Computers in Human Behavior*, *25*(2), 332–338.

Kalyuga, S. (2007). Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective. *Educational Psychology Review*, *19*(3), 387–399.

Kalyuga, S., & Plass, J. L. (2009). Evaluating and managing cognitive load in games. In *Handbook of research on effective electronic gaming in education* (pp. 719–737). IGI Global.

Ke, F. (2009). A qualitative meta-analysis of computer games as learning tools. *Handbook of Research on Effective Electronic Gaming in Education*, *1*, 1–32.

Khaled, R. (2008). Culturally-relevant persuasive technology.

Klopfer, E., Osterweil, S., Salen, K., & others. (2009). Moving learning games forward. *Cambridge, MA: The Education Arcade*.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Research methods in human-computer interaction. Morgan Kaufmann.

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science*, *341*(6149), 976–980.

Martin, Bella.Hanington, Bruce M. (2012) *Universal methods of design :100 ways to research complex problems, develop innovative ideas, and design effective solutions* Beverly, MA : Rockport Publishers, 2012. Print.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52.

Molzhon, A. (2016). *Exploring the influence of socioeconomic status on the executive function and theory of mind skills of preschoolers*. Virginia Commonwealth University.

Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology*, *91*(2), 358.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science*, *32*(1–2), 1–8.

Papastergiou, M. (2009). Digital game-based learning in high school computer science education: Impact on educational effectiveness and student motivation. *Computers & Education*, *52*(1), 1–12.

Preston, J. (2010). Prosthetic white hyper-masculinities and `disaster education'. *Ethnicities*, *10*(3), 331–343.

Rosewater, M. (2012). Piggybacking.

Sanders, L., & Stappers, P. J. (2012). *Convivial design toolbox: Generative research for the front end of design*. BIS.

Schell, J. (2014). *The Art of Game Design: A book of lenses*. CRC Press.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology*, *85*(3), 440.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453.

Squire, K. D. (2008). Video games and education: Designing learning systems for an interactive age. *Educational Technology*, *48*(2), 17.

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296.

Swink, S. (2009). Game feel. *A Game Designer's Guide to Virtual Sensation. Burlington, MA*, *1*.

Takeuchi, L. M., & Vaala, S. (2014). Level up Learning: A National Survey on Teaching with Digital Games. *Joan Ganz Cooney Center*.

Thompson, J. (2014). Engaging girls' sociohistorical identities in science. *Journal of the Learning Sciences*, *23*(3), 392–446.

To, A., Holmes, J., Fath, E., Zhang, E., Kaufman, G., & Hammer, J. (2017). Modeling and Designing for Key Elements of Curiosity: Risking Failure, Valuing Questions. *Proceedings of DiGRA 2017*.

To, A., Fath, E., Zhang, E., Ali, S., Kildunne, C., Fan, A., Hammer, J., & Kaufman, G. (2016). Tandem Transformational Game Design: A Game Design Process Case Study. In *Proceedings of the International Academic Conference on Meaningful Play*.

To, A., Fan, A., Kildunne, C., Zhang, E., Kaufman, G., & Hammer, J. (2016, October). Treehouse Dreams: A Game-Based Method for Eliciting Interview Data from Children. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*(pp. 307-314). ACM.

Van Merriënboer, J. J. G., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist*, *38*(1), 5–13.

Vasalou, A., Khaled, R., Gooch, D., & Benton, L. (2014). Problematizing cultural appropriation. In *Proceedings of the first ACM SIGCHI annual symposium on Computer-human interaction in play* (pp. 267–276).

Westera, W., Nadolski, R. J., Hummel, H. G. K., & Wopereis, I. G. J. H. (2008). Serious games for higher education: a framework for reducing design complexity. *Journal of Computer Assisted Learning*, *24*(5), 420–432.

Wilson, A. J., Dehaene, S., Dubois, O., & Fayol, M. (2009). Effects of an adaptive game intervention on accessing number sense in low-socioeconomic-status kindergarten children. *Mind, Brain, and Education*, *3*(4), 224–234.

Woodcock, A., Hernandez, P. R., Estrada, M., & Schultz, P. (2012). The consequences of chronic stereotype threat: Domain disidentification and abandonment. *Journal of Personality and Social Psychology*, *103*(4), 635.