Balancing Play and Formal Training in the Design of Serious Games

Rosa Mikeal Martey¹, Jennifer Stromer-Galley², Adrienne Shaw³, Brian McKernan⁴, Tobi Saulnier⁵, Elizabeth Mclaren⁵, Matt Rhodes¹, James Folkestad¹, Sarah M. Taylor⁶, Kate Kenski⁷, Ben Clegg¹, and Tomek Strzalkowski⁸

Abstract
This article discusses the design and development of two serious games intended to train people to reduce their reliance on cognitive biases in their decision-making in less than an hour each. In our development process, we found a tension between rich and flexible experimentation and exploration experiences and robust learning experiences that ensured the lesson content was easily understood and recalled. In line with game-based learning research, initial designs were oriented toward exploration and discovery. Analyses of interviews, playtesting, logs, and surveys revealed that many players were frustrated or confused by the interface and content of the more complex games, even when consistent differences between levels of visual detail or narrative complexity were not present. We conclude that teaching
complex topics such as cognitive biases to the widest range of learners required reducing the games’ playful and exploratory elements and balancing formal training content with simpler visuals and text.

Keywords
educational games, engagement, challenge, multimedia learning

Designing educational games has been the topic of much discussion, research, and debate among game developers, educators, and scholars. The game-based learning research posits that play is the essential component to better and more active learning (Aldrich, 2005; Gee, 2003; Squire, 2011) because the “learning by doing” possible in games can provide motivation, engagement, and reflection (Gee, 2003; McGonigal, 2011). Scholars posit that problem-solving through games can prompt players to recall facts, to analyze, or to implement strategies that enhance learning outcomes relative to more passive or traditional forms of instruction (Whitton, 2014). Traditional forms of instruction offer considerable benefits as well; however, designing effective play spaces can sacrifice critical elements of instructional design (Klopfer, Osterweil, & Salen, 2009). For short educational games, we suggest that design that is not quite as playful as traditional games nor quite as serious as traditional formal training has the best learning outcomes.

A growing body of research tests the characteristics of games for their learning outcomes, but little work examines the process by which developers make design decisions. Accordingly, we describe how playtesting, observation, and formal experiments contributed to the design and development of two serious, educational games intended to train people to reduce their reliance on cognitive biases in decision-making as part of the Sirius program. In the sections that follow, we first describe the games and experiments conducted. We then discuss the main lessons learned in the design and testing of visuals and narrative. The next section addresses the design and testing of lesson content and training techniques, including language style and in-game testing. We conclude with lessons learned overall.

The Games: Context

To identify overarching lessons learned in developing serious educational games, this article draws on quantitative and qualitative data generated from playtesting, pilot testing, and formal experiments run throughout the Sirius program. It is important to note that although these games have playful elements including the characters and dialogs used, they are serious training games designed to help people understand and adjust complex aspects of their thinking processes. As such, these games draw less on the types of playful experiences advocated by Klopfer, Osterweil, and Salen (2009) and more on the structural elements of games encouraged for learning (Gee, 2003; McGonigal, 2011; Whitton, 2014; Wolf, 2003). Thus, the lessons learned and
discussed here are most applicable to training games in particular. Although this limits the context of our conclusions, this study provides helpful insight into how different elements of design and interaction may contribute to creating successful games that train.

In the first phase of the research, we developed nine games which were playtested and piloted with about 125 people, and we conducted four formal experiments with 1,316 people and interviews with 82 participants. In the second phase, we developed six games that were playtested and piloted among about 100 people, and we conducted three formal experiments among 1,224 participants and interviews with 60 participants recruited from three universities. In a computer lab, after consent, participants took a pretest survey, were randomly assigned to play a version of the game or watch a comparison video, and then took a postsession survey. Surveys assessed players’ knowledge of (or “recognition and discrimination” of) and ability to mitigate three biases in each phase, along with game and computer experience, demographics, and personality variables in the pretest and engagement and usability in the posttests.

Different versions of the game were developed to experimentally test a range of game variables: character customization, art richness, narrative richness, repeat play, and game length in Phase 1; in-game rewards, control over content (interactivity), the presence or absence of a structured analytic technique called *structured self-critique*, and repeat play in Phase 2. The results of five of these tests are discussed briefly in this article and are reported in full elsewhere (see Martey et al., 2014; McKernan, Martey, Stromer-Galley, Kenski, & Folkestad, 2015; Shaw et al., 2016). The remaining experimental results are in development and are not reported here.

The games were 2-D, flash-based, puzzle adventures. The first game was designed to train players to reduce bias blind spot (BBS), fundamental attribution error (FAE), and confirmation bias (CB). FAE is when people assume personality-based factors rather than situational influences account for the behavior of others (see Gilbert & Malone, 1991), CB is the tendency to seek out and remember information that matches or supports one’s view (see Nickerson, 1998), and BBS is the tendency to consider one’s own decisions as being free from bias, even when one can recognize that bias in others (see Pronin, 2007). These biases have been identified as complex concepts to understand and difficult to train (Larrick, 2004).

The game, called *CYCLES Training Center* (CTC), lasted 30 min and asked players to infiltrate a nefarious bias training center and learn how to avoid the biases by completing a series of nine puzzle rooms. Between each room was a “transition room” where players saw lesson summaries in infographics, took multiple choice quizzes, and learned about bias definitions and mitigation strategies via text and voice-over. In the rooms, players explored ways to avoid a specific bias by moving an avatar around the room to flip switches, hit buttons, and talk to other characters such as workers and robots. The best (and final) game of the CTC used minimalistic visual detail and had little narrative content (see Figures 1 and 2).
The second phase of the research developed and tested versions of a 60-min game called CYCLES Carnivale (CC) that trained players to avoid anchoring bias (AN), projection bias (PR), and representativeness bias (RB). AN occurs when an individual is influenced by a number or topic (an anchor) before making an assessment and results in estimates that resemble the anchor too closely (Tversky & Kahneman, 1974; Wilson, Houston, Etling, & Brekke, 1996); PR is when people assume others share one’s own values, preferences, skills, or habits (Loewenstein, O’Donoghue, & Rabin, 2002); RB is a term that refers to a set of errors that ignore or miscalculate the actual likelihood of events, such as drawing conclusions based on a limited (non-representative) or misleading set of examples (Ajzen, 1977; Micallef, Dragicevic, & Fekete, 2012; Tversky & Kahneman, 1974). AN and RB have been identified as particularly difficult to mitigate (Tversky & Kahneman, 1974).

In the game, players were marooned on an asteroid and had to fix their ship by completing a bias training course presented in 17 tents, booths, and shops that involved carnival game scenarios (see Figures 3 and 4). The CC games primarily used first-person views and did not allow for avatar movement. Tents were linear sets of screens players clicked to learn about each bias and answer multiple choice questions. These ended with interactive infographics that summarized the bias lessons. In the booths and shops, players had the opportunity to experiment with different decisions to apply the lessons taught in the tents. Between levels, players took several multiple choice quizzes about the content they learned throughout the game.
The design and development process in the two phases of the CYCLES games used extensive iteration that improved core lesson content and design between each experiment cycle. We collected and analyzed observational, interview, log,
and survey data and made use of extensive playtesting and more formal pilot tests before each experiment. These data informed each subsequent iteration and the corresponding design decisions. Figure 5 shows that the design changes we made between experiment cycles resulted in substantially improved learning outcomes, evident in the percentage of improvements in bias mitigation and knowledge measures for the best game of each experiment by phase. Comparisons of Phase 1 and 2 games are for reference only, as the learning outcome measures assessed different biases; as well, the bias measures were modified during Phase 1. We also suspect it is more difficult to train players in the biases of Phase 2 than those of Phase 1.

Although considerable improvements to learning outcomes were found from the first to the final game of each phase, few of the experiments we conducted showed significant learning or engagement differences between the game variables we experimentally manipulated. Other changes we made between experiments may have been more important in contributing to these improved outcomes. The following sections discuss the insights we gained in development along two dimensions of those changes: playfulness in art style and narrative and instructional content in language and testing. Additional research is needed to examine the specific effects of such changes.

**Designing Playfulness: Visuals and Narrative**

**Visuals**

Visuals are considered an important element of what makes games and their content comprehensible (Mayer, 2011), engaging (Bailenson, Yee, Merget, & Schroeder,
2006), and memorable (Houts, Doak, Doak, & Loscalzo, 2006) because they provide additional information learners can draw on in processing lesson content. Research has found, however, that when players must spend considerable effort interpreting and integrating complex images and video in multimedia learning, they can become distracted and confused due to information overload (Mayer & Moreno, 2003). Minimalistic visuals may therefore be more effective than richly detailed visuals in educational games (Wolf, 2003). We tested the impact of a learning game’s visual richness by comparing a full color, richly textured version of the CTC game to a black-and-white version with minimal shading (see Figure 6). We found no significant differences in learning or engagement between the two games (see Martey et al., 2014).

These findings do not suggest that visuals are unimportant. In the development process, playtesting and interviews demonstrated that players were responsive to art style, especially in its contribution to the game’s narrative. One interviewee who played the minimal art style CTC game noted that the simple visuals “helped me focus on the stuff that I needed to click on or answer.” Interview analyses suggested that although players preferred the richer, more colorful art style, the complex visuals may have increased their expectations of the game’s playfulness, making them more critical of its narrative and production value than those who played the less visually complex version.

We improved our design of later games by balancing simple visuals that allowed players to easily understand the context and tasks with visuals that were rich enough to be engaging and memorable. For example, in both the CTC and CC games, we used infographics to summarize the learning content. Throughout development, we reduced visual complexity and used more consistent images and

![Figure 5. Pre- to posttest improvements in bias mitigation and knowledge by phase and game.](image)
icons. Simpler visuals were important in reducing cognitive load and helping players focus on the learning content, especially because learning about cognitive biases is a difficult process.

We also iteratively improved interface design to more clearly indicate expectations for players. In some places, for example, players were confused about what was clickable on the screen or how to select objects, resulting in players clicking many objects and taking a long time to complete challenges. To address these struggles, we added visual indicators such as cursors that became feet to indicate walking, hover effects with hints about why players might click an object, and flashing borders when players made mistakes. For example, one RB challenge in the CC game,
“Bottle Blast,” asked players to select a type of object to knock over stacked bottles. To determine which was best, players had to reveal the chances that each object was heavy enough to knock over the bottles by clicking on a label (see Figure 7). Lab supervisors noted that some players could never determine that they needed to click those boxes. We made the labels more visually prominent, changed their text, and added arrows and highlighting if players repeatedly failed to click them. We found that making strategic visual changes reduced the number of extraneous clicks and the time it took to complete each challenge in both the CTC and CC games.

Visuals can also provide valuable feedback to learners about their performance. Game points, rewards, visual and audio fanfare, and game progression serve as an instantaneous form of self-monitoring, allowing players to gauge their current skill level and thus encouraging players to improve upon their deficiencies and pursue tasks more diligently (Kelle, Klemke, & Specht, 2013; McGonigal, 2011). More generally, real-time performance feedback is considered a vital component of effective learning systems of all types (Butler, Godbole, & Marsh, 2013).

Our research shows that simply including reward features in a game as a form of feedback is not sufficient to improve learning outcomes. In our experiment assessing quantity of in-game reward features such as visual, audio, and textual fanfare, prizes, and collectible objects, we found that the presence of numerous reward features had no impact on learning outcomes and, moreover, had little impact on players’ reported sense of feeling rewarded. That is, there were no differences in how rewarded players felt across the two games (McKernan et al., 2015). However, those who felt more rewarded, regardless of the quantity of reward features in the game, did find the game more usable, engaging, and appealing, which may enhance
players’ motivation to play (Przybylski, Rigby, & Ryan, 2010). Although the presence of numerous types of in-game rewards may not have influenced learning, some forms of visual feedback are necessary to provide players with a sense of the consequences of their actions. Therefore, in the final CC game, we integrated additional visual and textual feedback to enhance the game’s appeal, such as a pie that hit the screen for incorrect responses or confetti that dropped after correct responses (see Figures 8 and 9).

Overall, we sought to avoid overloading players with too many cues to understand easily and quickly by reducing some of the playful and detailed visual information and feedback. We concluded that sacrificing some playfulness in the visuals in favor of the information-processing benefits of a more formal approach contributed to substantial improvements in training across game iterations without reducing player engagement.

**Narrative**

Narrative complexity, too, highlights a tension between generating player engagement and distracting from the learning process. Narrative in learning contexts has been found to increase learning and transportation (Green, Brock, & Kaufman, 2004), promote reflexivity (Conle, 2003), lead to exploratory learning (Mott, Callaway, Zettlemoyer, Lee, & Lester, 1999), and provide motivating learning scenarios (Rowe, McQuiggan, Mott, & Lester, 2007). This prior research suggests that
an engaging, memorable story with robust characters can contribute to effective learning in games.

For our short games, however, using richer narrative meant less space for lesson content. In an experiment with the CTC games, we compared the learning outcomes of a game with a richer, more complex narrative to a version with a simplified narrative and did not find significant differences in learning or in engagement (Martey et al., 2014). The next iteration of the CTC game, therefore, shifted some of the game’s cognitive load into richer content learning by reducing narrative richness (e.g., by removing much of the characters’ backstory). Interviews and analyses of player engagement reinforced this notion: We found that players recalled many of the narrative details in the rich narrative game, suggesting that they were devoting cognitive energy to understand the plot and characters. That narrative was reduced in subsequent games, but engagement actually increased slightly. For example, comparing the rich narrative CTC Game 2 to the final CTC Game 4, average immersion went from 3.8 to 4.2, transportation increased from 2.7 to 3.7, and general engagement went up slightly from 3.2 to 3.5. Likability decreased from 4.5 to 4.2. This reduction in narrative did not diminish engagement.

We also iteratively reduced narrative complexity across the CC games’ development process. For example, the first CC game story involved the player being shot down from space onto an alien carnival. Playtesting and interviews revealed that the alien characters and objects, such as rocket ships, may have made it difficult for players to transfer the bias training to real-world decision-making contexts.

Figure 9. Phase 2, Game 3 success feedback.
Accordingly, the final version of the game replaced the aliens with humans and integrated more real-world examples in the training tents and challenge booths (e.g., having the player purchase tires rather than a rocket ship engine). The number of errors players made within challenges and the number of times they had to repeat failed challenges dropped 15% each from CC Game 1 to Game 3, likely in part due to these changes.

Overall, our development process in both phases demonstrated that although richer narrative provided a more playful and game-like experience, in order for players to quickly process information relevant to their decision-making and improve their ability to understand key concepts, we needed less emphasis on solving narrative dilemmas and more real-world examples. We concluded that these adjustments contributed to improvements in the CTC game’s training outcomes from 44% mitigation in Game 2 to 56% mitigation in Game 4 and from 27% in the first CC Game to 47% in the last one (Game 3).

**Designing Learning: Language and Testing**

**Language**

The CTC and CC games aimed to introduce complex concepts about cognitive biases to a population accustomed to learning (college students) who had little if any previous knowledge of biases. Learning theory suggests that overly complex language impedes cognition (Mikk, 2008; van Weert et al., 2011) and that effective learning requires content presentation that suits the particular subject matter and learner population (Stodolsky, 1988).

In developing the text for the CTC and CC games, we started with formal definitions of the biases and used playtesting and pilot tests to adjust language complexity to the level of our learners and the learning context. We sought to reduce text and use consistent phrasing of bias definitions in everyday (rather than academic) language that was easier for players to process and recall, especially because the games were less than an hour each and few players had prior knowledge of biases. For example, in the first version of the CC game, we defined AN as “AN causes people to overemphasize recent information seen: numbers, facts, or events. The anchor is the piece of information that pulls people in its direction when they make an estimate or decision.” We changed this by the final game to, “Unconsciously relying on the first piece of information you encounter when making a decision.” In both games, we reduced some of the humorous, playful, and narrative text content in favor of formal lesson content and examples. We used automatic text-parsing tools to assess the degree of challenge our game text posed to players (Kincaid, Fishburne, Rogers, & Chissom, 1975) and found that we had successfully simplified the vocabulary from the first to second phase, with an average reading comprehension level of 8.4 in the final CTC game and 7.2 in the final CC game.
Throughout this process, however, we actually increased the total number of unique words in the games, contrary to our perceptions of the process during development. The discrepancy between our perception of changes to language and these findings underlines the importance of formal tools for assessing design goals during development. Although we did not reduce word count across each game’s iterations, these goals did help streamline the design process because we could reuse effective language throughout the game rather than reinventing ways to explain the core concepts.

The increase in words used was in part due to adding feedback with specific explanations for why the player’s actions were correct or incorrect. Most of the feedbacks in early game iterations were designed to be brief corrections to incorrect player actions, but with each game design cycle, we added feedback that also explained to players why certain actions were successful. This ensured that players saw core learning content regardless of their performance on challenges. For example, a room about CB asked players to assess robots of different colors and shapes to determine which were flammable. Incorrect selections resulted in an exploded robot, serving as visual feedback that their answer was wrong, but no text was used. We added text to the explosion event in order to explain why the wrong choice was biased, such as “Testing the hypothesis that Fire Bots are red by testing red ones is CB.” We also added text to correct selections to explain why they worked, such as

You overcame CB and discovered Fire Bots are round but not always red, disproving the hypothesis. You also did not exhibit CB in the last test. If you had, you would have picked the round, red bot. Testing the round, blue bot gave you new information and showed you that not all Fire Bots are red.

The CC game went through a similar process of language and feedback refinement from the first game to our final and ultimately best-performing game. Taken together, these findings suggest that with these changes, we may have traded complex language in lesson segments for additional types of learning structures and more useful feedback that explained why player actions were correct or incorrect (Butler et al., 2013). That is, by reducing some of the cognitive load of understanding bias definitions, we were able to devote player attention to feedback and testing more effectively.

Testing

Research on learning suggests that testing can significantly improve learning by encouraging learners to apply concepts to novel scenarios or demonstrate their understanding of those concepts (Roediger & Butler, 2011). Testing also encourages players to retrieve previously learned information in more effective ways than restudy or summaries, a phenomenon known as the testing effect (see Rowland, 2014, for a review). Further, tests most potently encourage learning and transfer of lessons to novel situations when they are accompanied by feedback that includes a comprehensive explanation of the answer (Butler et al., 2013; Kornell & Rhodes, 2013).
Both the CTC and the CC games included several types of testing experiences. Both used multiple choice quizzes administered between levels on plain screens distinct from the games’ visual setting (see Figures 10 and 11). The CTC games of Phase 1 used between 8 and 16 quizzes focused on bias recognition and definitions asked in groups of one or two in the transition rooms as part of the content summary and review. Observation and interviews revealed that participants found these quizzes too easy, however, so we increased their difficulty. As a result, the average correct responses went from 81% in the first CTC game to 68% in the final CTC game.

We also added testing moments to the final CTC game in the form of puzzles to open the door to leave each challenge room. These were “complete the sentence” format and were generally more difficult than the quizzes (see Figures 12 and 13). This nearly doubled the quantity of testing moments in the CTC game.

Initial versions of the CC game used more quizzes in each level (3–4) relative to the CTC game, included considerable detail in response feedback, and in most versions of the CC games, players had to answer additional quizzes if their initial response was incorrect. We worked carefully to calibrate quiz difficulty in the context of the learning process. For example, the first CC game started with quizzes that were more difficult than those we had used in the CTC games. We further increased their difficulty by selecting quizzes that were challenging to play testers and pilot participants. We reduced quiz difficulty for the final CC game; however, as we found in interviews and playtesting that they were too frustrating for players. To do so, we removed the trickier “best definition” and “most effective” questions and returned to using 10–20 of the simpler identify-and-mitigate types of questions from

Figure 10. Quiz screens in CYCLES Training Center games.
the first CC game iteration. In the first CC game, participants got 57% of quizzes correct on average, and by the final game, they got 76% correct.

Another design approach that we felt improved the effectiveness of the quizzes was their level of detailed feedback. For example, quiz feedback in the CTC games and early versions of the CC games generally identified the correct answer when players were wrong without explanatory text. In the final CC game, however, quiz feedback included comprehensive explanations, such as “This is a form of RB, assuming a small sample (like 1 bulls eye) can allow one to predict future performance. When making your choice, consider the size of the sample.” In the final CC game, players made 16% fewer biased selections, repeated challenges 21% less often, and made 18% fewer clicks over the minimum required to complete each challenge. We believe this was indicative of a reduction in struggles with the interface and improved understanding of the learning content (see Martey et al., 2014).

Another element of in-game testing that likely contributed to better training was that many of the questions presented scenarios different than those used in the games to explain the biases in the lesson segments. For example, in the CTC game, players attempted to fix various broken robots and machines, but quizzes included real-world scenarios such as a mother with a crying child in a store. This was intended to facilitate transfer from the game to other contexts and help players learn not just how to solve biases in the game but also to consider the concepts about biases they learned from different perspectives.
Analyses of players’ performance on the in-game tests showed that they were closely associated with posttest performance, especially in the CC games of Phase 2. For example, posttest performance on bias knowledge was on average 13% higher among those who got at least 80% of the quizzes correct in the CTC games and 17% higher among those who got 80% of the CC game quizzes correct. We concluded that performance on in-game testing could be a useful gauge of how much players are learning when games are used in training without immediate posttests. Although testing, especially with formal multiple-choice formats such as used in the CTC and CC games, departs from traditional game design approaches, we concluded that our more formal in-game testing was a vital part of producing effective training in the CTC and CC games and the improvements in learning outcomes we found across game iterations.

**Observations and Conclusions**

The design and development process for the educational games we created as part of the Sirius program afforded us the opportunity to combine playtesting, pilot testing,
survey, and log data to identify key themes in what did and did not work in producing potent learning outcomes. Much of the research on educational games focuses on ways that games are different than other types of learning contexts, especially games’ playful characteristics (see Gee, 2003; Klopfer et al., 2009; McGonigal, 2011). Less attention is paid to the techniques that address how learners need to process knowledge and what it takes to use such knowledge in practice (Ball & Bass, 2000). This article addresses that gap by identifying key lessons in two elements central to engaging play (visuals and narrative) and two central to good instruction (language and testing). We found a series of trade-offs between playfulness and formal instruction, and our adjustments helped reduce players’ cognitive load and maximize the learning outcomes.

Importantly, some of the decisions we made were not based on statistically significant differences between variables used in the formal experiments, such as art style, narrative complexity, and reward features. Instead, we combined game developer techniques of playtesting and team review with social science tools of experimental results to calibrate the complexity of these features to suit the learner population and short learning game context. Our strong improvements in learning outcomes suggest that we were successful in reducing the time players spent on learning how to interact with the game and understand the story and enhancing the cognitive resources they spent on learning how to recognize, define, and mitigate cognitive biases.

These choices were driven, too, by the type of game appropriate to the Sirius program. First, our games needed to provide robust training for players across the
spectrum of experience and interest in games. For example, games that require precision timing or complex combinations of clicks to advance are less accessible for novice players than they are for experienced players. Those who need more time to learn the controls or interface must focus on the mechanics of learning—or of the game—before they can focus on the content. Similarly, certain types of challenges can impede learning, such as struggle with the structure, expectations, and feedback in educational materials (Csikszentmihalyi & LeFevre, 1989; Orvis, Horn, & Belanich, 2008), especially in multimedia learning contexts including educational games (Mayer & Moreno, 2003). As a result, educational games may be more effective among those who are more comfortable with games than those who rarely play them (Clark, Fleck, & Mitroff, 2011). Designing for both experienced and inexperienced game players, then, requires a careful balance of playful, game-based experiences and more traditional lesson content.

Second, these games were designed to take an hour or less to complete. Therefore, our choices were intended to maximize learning with relatively brief exposure. For example, although discovery-based learning has been found to be extremely effective (Squire, 2011), short learning experiences may not benefit as much from this approach. Similarly, it is possible that creative problem-solving and higher level decision-making would be better with more abstract, self-directed learning experiences. But for understanding and recall of basic principles even around a complex topic such as biases and their mitigation, our concrete and simple design was effective. As a result, our games were based less on the narrative and discovery principles of learning design and focused more on the most efficient recall and integration techniques such as topic spacing and effective testing. Indeed, when participants took follow-up surveys assessing their bias mitigation and knowledge several months later, the participants demonstrated significant retention of this training (see Clegg et al., 2014, 2015).

Third, because these games were designed for a classroom or other guided learning context rather than the marketplace, we did not prioritize interest in buying or playing them again. Although richer narrative, complex visuals, and elaborate feedback may be more appealing, playful, and fun for players, we found minimal differences in engagement levels across different versions of the games. However, we did find that the most effective learning resulted from playing our games a second time a week or two later (see Clegg et al., 2015). This suggests that the appeal of the games may remain an important factor in motivating players to return to them later and thus improve their learning.

Our design and development process was ultimately a process of using playtesting, pilot, and experimental data to inform how we balanced the level of playfulness and formal teaching in the games’ visuals, narrative, language, and testing. Additional research on these factors is needed to rigorously test how this balance influences outcomes. The increasingly effective learning outcomes we found across iterations within each phase demonstrated that this balance can result in a powerful
learning tool for those already drawn to educational games as well as for those who might hesitate to engage with them.

Authors’ Note
The views and conclusions contained in the article are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory contract number FA8650-11-C-7176.

Notes
1. Scales for bias measures were developed and tested independently by the authors for Phase 1 and by a joint team of all performers for Phase 2.
2. Percentage of improvements is calculated by first converting raw scores to percentage of maximum performance, then pretest % – posttest %/(1 pretest %).

References


Author Biographies

**Rosa Mikeal Martey** is an associate professor in the Department of Journalism & Communication at Colorado State University. Her research focuses on games, identity, social interaction, and digital media as well as educational games and multimedia learning. Her work can be found in New Media & Society, Information, Communication & Society, and Games Studies, among others.

**Jennifer Stromer-Galley** (PhD, University of Pennsylvania, 2002) is an associate professor in information studies. Her expertise is on communication processes and effects through digital communication channels. She has over 30 publications that focus on dimensions of digital media around influence, leadership, political communication, and training.

**Adrienne Shaw** has been studying video games since 2005. She is the author of *Gaming at the Edge: Sexuality and Gender at the Margins of Gamer Culture* (University of Minnesota Press, 2014). Her research has been published in Ada: Journal of Gender, New Media, and Technology, New Media and Society, Critical Studies in Media and Communication, and Games and Culture, among others.

**Brian McKernan** is a visiting research professor at Syracuse University. McKernan’s research focuses on examining the educational and civic significance of digital games and using these insights to help develop game-based learning software and other innovative applications.

**Tobi Saulnier**, PhD in electrical engineering and previously worked at GE Research, is the founder and CEO of 1st Playable Productions, where she led initiatives in embedded systems, new product development, and software quality. She is a frequent speaker at industry conferences in the areas of usability, learning game design, and production processes.

**Elizabeth McLaren** is a production coordinator and lead artist at 1st Playable Productions. She oversees day-to-day production of software projects and focuses on both children’s and adult games.

**Matt Rhodes** is an associate professor in the Cognitive Psychology department of Colorado State University. He specializes in human memory, memory and age, and metacognition. His work addresses judgments of learning and metacognitive illusions. Recent work can be found in Psychological Bulletin, Memory, and the Journal of Experimental Psychology: General.
James Folkestad is an associate professor in the School of Education at Colorado State University. His work focuses on educational technology, activity systems for learning, innovative inference game-based learning, and social network analysis.

Sarah M. Taylor has supported efforts in the Intelligence Community and U.S. DoD to improve analysis methods and apply computer-based methods to analysis and information processing for over 30 years. She was the principal research engineer for Lockheed Martin natural language processing efforts. She has participated in numerous DARPA and IARPA programs, most recently SIRIUS, Metaphor, and SCIL, fulfilling roles as SME for analysis and making significant contributions to the emerging field of computational sociolinguistics.

Kate Kenski (PhD, University of Pennsylvania) is an associate professor in the Department of Communication and School of Government & Public Policy at the University of Arizona. She has published over 30 articles, books, and book chapters on political communication, public opinion, and research methods.

Ben Clegg is a professor of cognitive psychology at Colorado State University. His research is on aspects of human performance, particularly how to apply principles from cognitive psychology to real-world tasks and skills. His work examines issues such as training, automation, workload, and situation awareness; sequence learning; and implicit learning.

Tomek Strzalkowski is the director of the Institute for Informatics, Logics and Security Studies and professor of computer science at SUNY Albany. He has done research in computational linguistics, sociolinguistics, information retrieval, question answering, human–computer dialogue, and formal semantics of natural language. He has directed research projects sponsored by IARPA, DARPA, EC, and NSF and published over 100 papers and books.