Serious Games for Reducing Bias in Credibility Decisions: An Evaluation Framework and Case Study

Abstract

Serious games show promise as an effective training method, but such games are complex and few guidelines exist for their effective evaluation. We draw on the design science literature to develop a serious game evaluation framework that emphasizes grounding evaluation in each of four key areas—theoretical, technical, empirical, and external. We further recommend that serious game developers assume an iterative, adaptive approach to grounding an evaluation effort in these four areas, emphasizing some areas more than others at different stages of the development cycle. We illustrate our framework using a case study of a large-scale serious game development project. The case study illustrates a holistic approach to serious game evaluation that is valuable to both researchers and practitioners.

1. Introduction

Serious games hold strong potential as an effective and engaging method of training and education [1]. A serious video game is a game intended to provide training or education through experiential learning, in which players encounter real decision-making scenarios and receive dynamic feedback regarding their actions [2]. Additionally, games have been shown to be more engaging and enjoyable than traditional instruction methods [1].

However, despite the promise of serious games in various training contexts, there exist few guidelines instructing game designers on how to evaluate the effectiveness of a serious game. Unlike traditional training methods in which learning can be rather easily gauged, serious games are highly complex systems with many different aspects that could affect their effectiveness. In addition to building training mechanisms into the players' interaction with the game, designers must also ensure that the game is engaging, reliable, and easily grasped, among other things. Furthermore, these goals must be accomplished within resource constraints and development schedules, and such a development effort typically requires collaboration among a large, interdisciplinary team. With such high stakes, serious game developers can benefit greatly from clear guidance regarding how to effectively evaluate a game's performance.

The framework developed in this paper provides a holistic approach to evaluating serious games. The

guidelines expand beyond the few examples of evaluation guidelines currently found in the literature [e.g., 3] to address the evaluation process using a systems building approach. We draw on the design science literature—in particular the Nunamaker et al. [4-6] model—to generate recommendations guiding the effective evaluation of serious games. The resulting framework highlights the value of iterative, adaptive evaluation, and consistent grounding in both theoretical and practical/external constraints, while ensuring that experimental testing is performed with maximal validity and generalizability. The guidelines produced are thus valuable for both future researchers and practitioners interested in leveraging serious games as an effective training tool.

2. Background

Before presenting our evaluation framework, we first summarize the Nunamaker et al. [4-6] design science approach, which heavily informs our evaluation approach and recommendations. We also summarize prior literature related to serious game evaluation.

2.1. The Design Science Approach

The Nunamaker [4-6] design science research (DSR) methodology informs our serious games evaluation framework. The model seeks to generate high-impact knowledge through iterative cycles of research activities, including theory development, prototyping, experimentation, and field study. The underlying premise is that as compared to short-term approaches, greater knowledge can be achieved by iteratively refining high-impact ideas in a sustained, agile program, shepherding those ideas from initial concept to standalone solution. The success of these programs of research is enabled and sustained by multidisciplinary academic and practitioner partnerships, the use of which both improves the breadth of available expertise and helps refine and add crucial detail to research program goals.

The emphasis on a sustained, adaptive program is a key reason for adopting the Nunamaker methodology for the current effort. Serious game development typically involves high uncertainty and requires agile responsiveness to many unforeseen factors. Finding and making the case for success is therefore expected to require more than a single investigation. A serious gaming framework requires

addressing the problem from multiple perspectives and iteratively learning from successes and failures.

2.2. Prior Work in Serious Game Evaluation

Recognizing the complexity of designing and evaluating effective serious games, past research has begun formulating initial frameworks on serious game evaluation. Much of this research provides observations on how to evaluate the learning outcomes of serious games, often in a primary or secondary education context. We draw on these observations to help inform our proposed framework, which focuses on the components of evaluating serious games from a holistic systems building approach. Below we summarize the most relevant research in this area and explain how it informs and corroborates our research.

Mayer and colleagues [3] provide a framework for evaluating serious games in an advanced learning context. Their framework consists of eight steps and provides the most comprehensive evaluation recommendations for serious games. Their framework is limited, however, in that the serious game and its learning outcomes receive the primary focus, with comparatively less focus on systems-related considerations. In particular, their framework does little to address the iterative nature of systems development. As we will argue, admitting that systems development is a nonlinear, cyclical activity implies the need for a more iterative, adaptive evaluation approach.

Other researchers have offered less comprehensive guidelines, addressing specific portions of the evaluation process. Nacke et al., [7] provide a summary of methodologies for assessing the game play experience, but less focused on the design and development of the game. The paper also suggests assessing how the player context (ethnography, multiplayer scenarios, etc.) influences the experience. De Freitas and Oliver [8] introduce a four-dimensional framework for evaluating the effectiveness of serious games within a curriculum. Amory [9] explains that the design and evaluation of serious games should consist of several dimensions or spaces, including the game space, the problem space, and social space. Finally, Astor et al. [10], suggests three requirements that serious games should entail. First, serious games should provide an engaging learning environment within a learning context. Second, the game should reward users for desired outcomes. The system should provide feedback in an unobtrusive and meaningful way.

Integrating this research with the Nunamaker [4-6] design science methodology, we create a more

holistic approach for evaluating serious games from a system building perspective.

3. A Framework for Evaluating Serious Games

The framework presented in this section draws conceptually from the Nunamaker et al. [4-6] DSR approach to provide a holistic approach to evaluating serious games. The guidelines we provide highlight the importance of grounding an evaluation strategy in each of four key conceptual areas, and are summarized in Table 1. Before presenting the framework, we first provide a definition of evaluation as it is used in this work.

3.1. Evaluation Defined

Software and systems are typically designed to satisfy certain requirements and achieve a specific purpose. Accordingly, part of any systems development effort—including that of serious game development—should include some form of evaluation in order to verify that the system meets specified requirements. Such evaluation takes a variety of forms, and typically includes such activities as checking the final product against stakeholder requirements, verifying that the system performs adequately under conceivable usage patterns, ensuring the interface is useable and logical for intended end users, and so on [11]. In referent literatures—most prominently those of the computer science and systems engineering disciplines—these procedures are frequently referred to as system verification and/or validation. To avoid confusion with the term “validation,” which is used most often in the IS literature in the context of measurement validity [e.g., 12], we employ the term “system evaluation” as follows:

System evaluation is the confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application of a system have been fulfilled. The purpose of evaluating a system is to acquire confidence in the system's ability to achieve its intended mission.

3.2. Theoretical Grounding

As prior DSR scholars have suggested, “kernel theories” can be usefully employed to advise design solutions [4, 13]. Serious games’ system requirements should thus be grounded in theory appropriate for the context. The reference theories used would typically be related to the learning outcomes, for which relevant theory will suggest

Table 1. Summary of Evaluation Framework

Category	Subcomponent
Theoretical Grounding	Evaluation requirements are informed by relevant theories. Theory permeates all other areas of evaluation. Outcomes are gauged against relevant, theory-informed baselines.
Technical Grounding	Evaluation is an iterative process, with requirements evolving and adapting as the system development matures. Evaluation requirements must be balanced with practical cost of implementation and technical feasibility.
Empirical Grounding	Using multiple methods of evaluation to avoid basing conclusions on a single method. Produce as much generalizability to the target population as is feasible. Use randomization, and measure and control for relevant covariates, to isolate the focal training effects. Measurement instruments used in evaluating the game's effectiveness should be rigorously evaluated in terms of measurement validity and reliability.
External Grounding	Evaluation is partially guided by stakeholder needs and expectations. Serious games should be evaluated in real-world settings. Longitudinally evaluate permanence of learning outcomes. Maintain focus on practical significance of results, avoiding overemphasis on statistical significance.

effective training methods. When strong theory is used as a starting point to generate system requirements, the process of building, evaluating, and using the system can ultimately provide the research team with insights that augment existing theories with new insights [13].

Strong theory should also be used in formulating tests of learning outcomes for serious games. The comparison of before-and-after states of knowledge using a pre-/post-testing paradigm is common in serious game evaluation [3, 14], due in part to this method's simplicity and ease of application in learning environments. However, comparing a post-test to its pre-test counterpart may be limited to the scope of the testing instrument. Therefore, we suggest that, in addition to comparing pre- and post-game knowledge states, learning outcomes should be evaluated against an appropriate baseline derived from relevant theory. In other words, we suggest that a serious game should be evaluated against the traditional form of training it is designed to replace.

3.3. Technical Grounding

Serious game development and evaluation should follow an iterative, adaptive process, as is suggested in more general systems-building approaches [5]. As a starting point, initial requirements and evaluation criteria are derived from kernel theories and stakeholder requirements. As development and evaluation of a serious game system proceed, designers will almost certainly identify opportunities and/or weaknesses in the original set of system requirements. An effective evaluation strategy will thus be a flexible one in which evaluation requirements are not held constant but are adjusted according to findings from early iterations in game development or changes in stakeholder requirements.

Serious game evaluation will also be affected by technical and operational feasibility constraints, some of which may not surface until later stages of the development process. Of course, care is taken at early stages of game design and development to carefully consider the technical requirements of the system requirements as specified by reference theories and stakeholder needs. Only those features that are deemed initially feasible are pursued and developed. But even carefully considered features may prove infeasible or require more resources than can be reasonably justified, and thus proper technical grounding entails continuous reevaluation of system requirements against their technical feasibility.

3.4. Empirical Grounding

An effective system development and evaluation strategy requires the use of appropriate, valid, and focused empirical evaluation [5]. Thus, a key component of evaluating the effectiveness of serious games entails careful use of controlled laboratory experiments to properly isolate the intended game mechanics and observe the nature and extent of the expected training outcomes. Empirically grounding a serious game evaluation strategy will provide the research team with objective data with which theoretical and technical aspects of the game design can be evaluated and iteratively adjusted through the development process.

The results of any one test of a phenomenon are subject to multiple forms of error or bias [15]. Researchers must take care to ensure that, to the extent possible, the results obtained in one test can be reasonably expected in the population of interest. First, using a single method to evaluate a serious game can lead to mono-method bias, which has long been a focus of IS methodologists, particularly in the

realm of survey research [16]. Using multiple methods to measure the same phenomenon counteracts this type of bias, reducing the likelihood of incorrect conclusions. In the context of serious games, it is important to evaluate learning objectives and other success metrics using multiple methods. Nacke et al. [7] provide a comprehensive discussion of various methods researchers can use to observe and evaluate users' interactions with a serious game system.

Second, an effective serious game evaluation strategy will avoid the use of nonrepresentative evaluation subjects [15]. Steps should be taken to ensure that evaluations subjects are not significantly different from the population of interest—a serious game intended to train adolescents regarding online safety, for example, would be ineffectively evaluated with adult test participants. An alternative to this would be to use multiple variations of a given testing protocol, run with subject pools that collectively differed from each other [17]. Such precautions can provide researchers and stakeholders with confidence that outcomes observed while evaluating a serious game in a laboratory setting can be expected as the game system is used by its intended user population.

A third component of effective empirical grounding entails isolating the effects of a game's training mechanisms from the effects of a plausible set of covariates and alternative explanations. The first key strategy used to accomplish this goal is to randomly assign experimental subjects to treatment groups. To further isolate the effects of training mechanisms within serious games, however, theoretically relevant covariates can be measured and statistically controlled for. Mayer et al. [3] produce a sizable list of potential covariates that may apply in a given serious gaming context. Game designers should carefully consider all variables that could influence learning outcomes in their chosen serious game context so that their influence can be accounted for in subsequent statistical tests.

Lastly, strong empirical grounding entails ensuring that other aspects of serious game testing are held to established methodological standards. Published serious game studies have traditionally lacked methodological rigor [14]. Serious game designers should place appropriate attention on the instruments and methods used in the course of evaluation, using accepted standards of reliability and validity for any measurement instruments used in measuring learning outcomes. Using and reporting rigorous research methods will generate confidence, among stakeholders as well as academic reviewers, in the validity and strength of conclusions drawn regarding the serious game being evaluated.

3.5. External Grounding

We have argued that theoretical, technical, and empirical considerations should guide an effective serious game evaluation strategy. However, these considerations should be tempered with real-world requirements of stakeholders and real-world value, given the target population and intended success outcomes of the serious game [5]. If, for example, stakeholders emphasize engagement and playability as supreme requirements for a particular game, this will likely have an effect on the extent to which theoretical mechanisms can be explicitly incorporated into training mechanisms. Developers of serious game systems will thus need to consider the theoretical and practical origins of system requirements, account for appropriate prioritization of those requirements, and cater their evaluation strategies accordingly.

To provide effective external grounding, the serious game system should be evaluated in real-world settings. A system evaluation protocol that is limited to a contrived laboratory setting will be less effective at predicting real-world success or failure of the system [4]. Thus, serious games should be tested in situations that are as close to their intended use cases as possible. The serious game experience is one in which users learn while competing or having fun, and this aspect of the game should therefore be carefully evaluated.

Further evaluating the real-world impact of a serious game, designers should note that learning objectives obtained in the short term are not as valuable as those obtained more permanently. To this end, a system evaluation protocol should expand beyond immediate tests of training effectiveness vis-à-vis learning outcomes and examine the long-term effectiveness of training mechanisms in facilitating learning [3]. Measuring such longitudinal learning is an expensive and difficult affair, however, so research teams will need to balance this need with available resources and expected attrition rates.

While on the topic of practical significance, we present a related, cautionary guideline. Running many iterations of a game-testing protocol will likely necessitate a large number of statistical tests. Game designers should be aware that many repeated statistical tests could inflate Type I error, leading to false conclusions. In addition, tests of serious games may be susceptible to an unwarranted focus on statistical significance (particularly when using very large sample sizes) rather than on meaningful effect sizes. Game designers should be sure to apply appropriate weight to statistically significant findings, and focus effort on mechanisms that

produce meaningful effects in the context of the serious game being developed.

3.6. Applying the Framework Iteratively

Prior work in serious game development has emphasized the value of iterative development in developing effective serious games [18, 19]. Iterative development allows the research team to develop portions of a game, test the quality and effectiveness of those portions, make small adjustments as new insights are gained, test the effectiveness of those changes, and so on. Such a rapid development cycle allows the researchers to constantly obtain feedback on the design, playability, and coherence of game scenarios while simultaneously ensuring that key theoretical aspects of the training mechanisms are maintained throughout the iterative process. In this manner, an effective serious game evolves and matures as the design team learns progressively more in the development process.

In presenting our evaluation framework, we submit that the evaluation process can similarly benefit from an iterative, adaptive approach. Evaluation strategies will necessarily evolve as researchers observe the effectiveness of various game mechanics. Those features that prove useful in early evaluation can be emphasized in later stages of evaluation, even evolving into performance metrics that can be used as a reference point as a game is more broadly tested. Researchers may understand conceptually how they would like a game to perform at the outset of a game development process (e.g., this game should teach players to make fewer biased decisions), but how those goals are operationally evaluated may become apparent only after several versions of a game are created and evaluated.

Furthermore, even multiple versions of the same serious game may require very different forms of evaluation in order to accomplish the overarching goals of a development effort. Early versions of a game will likely contain exploratory mechanisms that researchers hope will evoke learning outcomes. Such features may be best evaluated using exploratory, qualitative methods such as focus groups or one-on-one playtests in which players' behaviors and thought processes are carefully recorded and discussed. As a game matures and its functionality becomes more sophisticated, evaluation will require a corresponding increase in sophistication. In these later stages of the development process, evaluation can be both more targeted (since the designers have incorporated the training mechanisms proven most effective in early stages) and more generalizable (as more stable versions of the game can be tested reliably with broader testing samples). Thus, we advocate a highly

iterative approach to both development and evaluation, and we recommend that the four key methods of grounding found in our framework be applied throughout the iterative process.

4. The MACBETH Project

4.1. Overview

The MACBETH (Mitigating Analyst Cognitive Bias by Eliminating Task Heuristics) project was a multi-year project that was managed by the Air Force Research Lab (AFRL) and funded by the Intelligence Advanced Research Projects Agency (IARPA). The project was designed to create novel training activities so that intelligence analysts would improve their information gathering and assessment abilities. The MACBETH project included three different games that were developed and evaluated over the course of approximately 2 years. The MACBETH games were in the strategy/simulation genre and each version of the game was similar in terms of basic game mechanics. Players were immersed in a fictional environment where they gathered and analyzed intelligence data to prevent simulated terrorist attacks, and the players had to quickly synthesize the data to create hypotheses about who was planning the attack, how the attack would be carried out, and where the attack would take place. If the player failed, the attack occurred and the player had to replay the scenario [18].

In each scenario, the player had to reconcile carefully crafted bits of data so as not to fall prey to cognitive bias. If the players followed their cognitive biases in the game (e.g., actively confirmed hypotheses rather than disconfirmed hypotheses, attributed greater weight to dispositional rather than situational factors), their actions usually resulted in incorrect conclusions and failing the scenario. Players also received corrective feedback and instruction during the game about how to approach decision making and data gathering in an unbiased way. This feedback and instruction was designed to equip players with transferable tools that will not only lead to success in the game, but also hopefully assist in mitigating cognitive bias outside of the game.

4.2. Development and Evaluation

During the development of the MACBETH games, numerous pilot tests and play tests were performed to ensure playability and proper function of the games. In addition, three large scale experiments were conducted testing the effectiveness of the game [20]. Although some modification occurred in the measures between experiments, the

performance of the experimental conditions was judged by analyzing knowledge of cognitive biases, mitigation of cognitive biases, and engagement during learning. Based on the experiment findings, the best performing condition of the core game mechanic in each experiment (i.e., in-game training, feedback, and game type) was retained in future versions of the game. Therefore, at the conclusion of the experiment series, the final game contained core mechanics that had undergone rigorous and repeated evaluation.

4.3. Iterative Evaluation

Following the Nunamaker et al. [4-6] approach to systems design, the MACBETH games were developed iteratively, with an early focus on strong theoretical grounding and addressing the needs of external stakeholders (i.e., AFRL, IARPA, intelligence analysts). The development team used a process known as rapid iterative prototyping (RIP), in which various portions of the games were created, played, evaluated, and then adjusted according to feedback received [18, 19]. MACBETH evolved significantly from a limited conceptual version (which we called the “paper prototype”) to a full-featured game employing sophisticated training mechanisms [18].

Given the iterative nature of the game development process, evaluating the games was also an iterative, evolutionary process. Creating the early versions of the games and testing them with small groups of testers provided insights regarding what aspects of the game were most important for evaluation, and these could be emphasized in later iterations. For example, an early version of the game employed a skeuomorphic paradigm, staging the gameplay as if the player (who was acting as an “analyst”) was sitting at a desk—complete with a wall-mounted bulletin board and a paper in-basket—and gathering evidence. These early games were purposefully designed to satisfy requirements that the research team thought important. As the game was evaluated, however, it became clear that the skeuomorphism was a distraction that provided little value in terms of learning outcomes, and these design criteria became irrelevant in evaluating later versions of the game.

In addition, our evaluation strategies evolved as the games matured. For example, we employed far more qualitative evaluation techniques (e.g., focus groups, interviews, one-on-one playtests) in early versions of the games. At this stage of development, our understanding of the effectiveness of the game was limited, and we benefitted from the more open-ended nature of these evaluation strategies. At this early stage, the game setting, interaction paradigm,

and mechanics were still in flux and could be adapted as we learned which features were proving effective in helping participants think about biased decision-making. As effective game mechanics were discovered and the games became more sophisticated, our evaluation strategy adjusted accordingly. The games were evaluated according to more stringent performance standards, the tests of learning outcomes were more targeted, and we focused on evaluating the games with larger, more diverse groups of players.

The result of this adaptive evaluation approach was the development of theory-based serious games that catered to the needs of external stakeholders. They were developed within reasonable technical constraints that effectively leveraged limited resources, were tested extensively using multiple, valid methods with large numbers of subjects from the intended user population, and were further evaluated outside the lab and using longitudinal measurements to evaluate their real-world viability. In the sections that follow, we will summarize the ways in which each of the four key areas of evaluation found in our framework were applied in the development and evaluation of the MACBETH games.

4.4. Theoretical Grounding

The MACBETH project was deeply influenced by established bias mitigation theory, and the research team consulted established bias mitigation literature in designing the games as well as the evaluation techniques. The research team drew from the Heuristic-Systematic Model [21] and other dual-system models of information processing [22] to understand how biases are the result of heuristics used by decision-makers. Additional literature relating to specific biases was used to develop game mechanisms that addressed various types of bias [1, 23]. For example, prior research suggested potential interventions that might improve the systematic processing of game players by encouraging them to seek disconfirming information and question their assumptions [24]. Further, pre- and post-test measures targeting each of these biases were derived from prior literature, and were combined with additional measures created specifically for the MACBETH project [1, 23].

Given the existence of established theory regarding bias mitigation, the research team was able to compare the learning outcomes of gameplay with the learning outcomes of a theory-based comparison method. We used a professional, well-conceived training video that provided “state of the art” bias mitigation training as a general benchmark to

compare the games' training effectiveness with that of traditional training approaches. The training video employed many of the same interventions that were used in the MACBETH games, albeit within the confines of video medium. This strategy allowed us to not only establish that the games were improving learning and behavior, but also to quantify these effects in comparison to other theory-based approaches.

4.5. Technical Grounding

As mentioned above, the evaluation requirements for the MACBETH project evolved along with the design of the games. A rapid prototyping strategy was employed to develop and refine the games [18], and they were repeatedly evaluated along a variety of dimensions. While providing rapid feedback regarding design and gameplay mechanisms that were effective, this prototyping strategy also enabled the research team to iteratively adjust the evaluation strategy and develop success metrics that became performance standards later in the development process.

As with many system development efforts, the MACBETH research team was forced to make trade-offs in order to balance the utility of certain features of the MACBETH games with the resource cost of developing and improving those features. For example, one version of the game was designed to support multiplayer functionality. Upon testing this functionality with experimental evaluations, it became clear that the functionality did not improve the game in terms of training effectiveness. Rather than continuing to develop this functionality and further expend resources to create a multiplayer version that provided the expected outcome, these features were removed in later development iterations, and our evaluation requirements adjusted accordingly. Thus our evaluation requirements were somewhat fluid, with early conceptions of the game requirements adjusted as early iterations of the games were evaluated.

4.6. Empirical Grounding

Learning outcomes were evaluated using a variety of methods, each of which captured unique aspects of the games' effectiveness. The research team used pre-test and post-test comparisons of learner performance on tests designed specifically to measure the intended learning outcomes. Additionally, learner performance was re-tested with a follow-up approximately eight weeks after the experiment session containing the pre- and post-tests. Several scales were used for assessing bias

knowledge and bias mitigation to ensure adequate coverage of each bias. Where scales did not exist or only included a limited number of items (insufficient for pre-, post-test, and follow-up), we created additional items and scales. Prior to using these items and scales, we piloted them to examine their reliability and correlations with existing measures.

In addition, players provided answers to perceptual measures of the games' engagement, including being rated on scales of cognitive absorption. To validate these perceptual measures, we performed a pilot test that compared them to physiological responses (e.g., heart rate, skin conductance, and eye-tracking). We noticed differences in how cognitive absorption was manifest in the self-reported perceptual measures and the physiological responses, and we therefore used eye tracking, accompanied by perceptual, self-report measures to capture engagement in subsequent experiments [25].

Players' behavior while playing the games was also recorded and analyzed. The logs of players' gameplay activities were captured and subsequently examined across the three experiments for evidence of improved game play. These more objective measures of game effectiveness provided additional, rich information that contributed to a more comprehensive evaluation than would have been possible with simple pre-test/post-test comparisons or other perception measures.

The MACBETH games were tested extensively using student subjects, in line with the target population for the games and the guidelines established by the project sponsors. Given that student subject pools are typically homogenous in many respects [17], we used two key strategies to increase the generalizability of our findings within the constraints regarding the target population (i.e., college-aged students). First, we chose to test the games in two different locations using substantively different student samples. Our results occasionally differed across these two pools of experiment participants, which provided additional insight regarding the games' effectiveness across different demographic subgroups. Second, we used a series of iterative experiments with minor variations. This strategy allowed the research team to examine results across a wider range of test subjects and build consensus regarding the utility of various training mechanisms. These strategies served to increase the generalizability of our findings.

To isolate the effects of training mechanisms, a first step was to incorporate randomization in the study design, thus distributing participants' individual differences across the cells. The research

team also measured and controlled for a variety of relevant covariates. While the interpretations of the significant covariates were somewhat interesting in their own right, the covariates' primary function in the statistical analyses was to isolate the effects of training mechanisms and allow for a more accurate overall assessment of the games.

To ensure the validity of our testing results, all measurement approaches used in the pre-test and post-test evaluations were carefully derived from theory. These measures were developed and pilot-tested using rigorous, established validation techniques following standard methodological paradigms. Furthermore, all experimental procedures used in the evaluation of the MACBETH games were designed in accordance with established experimental practices, with careful controls, randomization, manipulation checks, and so on. The research team took great care to ensure that measurement strategies and experimental procedures were sound, even conducting a special pilot test whose sole purpose was to verify the measurement procedures, as well as cross-validate several perceptual measures of engagement and flow with physiological measures (e.g., using eye-trackers). All of these procedures gave the team confidence that the conclusions drawn during later stages of game evaluation were valid.

4.7. External Grounding

The initial game design, as well as our initial evaluation focus, were guided by the requirements established by key stakeholders, who suggested bias mitigation thresholds that the researchers were required to target. These thresholds remained in focus throughout the development and evaluation, which served to keep the research team and the game being developed focused on the needs and wishes of our sponsoring partners. Additionally, we also partnered with another non-academic organization that provided expertise and oversight of our development and testing of the MACBETH games. Specifically, external guidance was sought to ensure game content was believable and would be consistent with the challenges intelligence analysts actually faced.

Beyond these methods used to maintain alignment with stakeholder needs, we further grounded our evaluation externally by ensuring that we could be reasonably confident of the games' effectiveness in real-world usage over time. We used single vs. repeated game-playing experiences to compare the games' effectiveness over time within subjects. Additionally, the research team employed an 8-week follow-up evaluation to gauge the effectiveness of the serious games in terms of long-term bias mitigation.

Finally, MACBETH was tested using an optional "take-home" game. Some participants were invited to play the game at home, and they were free to play it as much as they desired. This unique strategy allowed for a separate, real-world evaluation of the game's enjoyability (i.e., did people voluntarily play the game?) and effectiveness (i.e., did additional playing from home increase bias mitigation?).

5. Discussion

Serious games show potential for training where traditional instruction methods have historically been ineffective [22]. Far from a simple solution, however, such games are highly complex systems that require substantial resources for successful implementation and testing [3, 18]. We have argued that evaluating serious games' effectiveness and viability can benefit from a systems-building perspective, accounting for the broader context of the serious game development effort. We draw on the Nunamaker et al. [4-6] design science approach to develop a serious game evaluation framework that emphasizes iteratively grounding evaluation in each of four key areas—*theoretical, technical, empirical, and external*.

The resulting evaluation guidelines highlight the value of iterative prototyping, adaptive evaluation, and consistent grounding in both theoretical and practical/external constraints, while ensuring that experimental testing is performed with maximal validity and generalizability. We then illustrated the application of these guidelines using a case study of a large-scale serious game development effort.

The most important contribution of our framework is its unique perspective on serious game evaluation, namely that serious games researchers should evaluate their games within the broader context of the system as a whole. With few exceptions, prior work has examined the process of serious game evaluation more narrowly, focusing extensively on evaluating games in terms of learning outcomes. Although learning outcomes are arguably the most important feature of serious game evaluation, they should not be overemphasized at the expense of other considerations. A serious game that proves itself effective in laboratory settings, for example, may still fail in actual use unless it is evaluated under real-world circumstances, which can reveal previously unknown issues and insights [6]. These precautions are pertinent as games become increasingly complex and sophisticated, requiring substantial resources and careful management.

We also provide a unique focus on iterative refinement of evaluation criteria, following the Nunamaker et al. [4-6] design science approach. This

focus on iterative evaluation distinguishes our framework from prior literature. Prior work conceptualizes the development and evaluation of serious games as though it occurs in a single iteration—the game is designed, built, and evaluated. An agile approach to the complex goals underlying serious game development is likely to achieve greater impact. In the earliest stages, overarching concepts are developed and regularly evaluated qualitatively by teams with disparate expertise. Little or no major experimental evaluations are necessary in the earliest stages. The major goal of the initial stages is to identify the overarching goals of the games and the features that will be used to accomplish those goals. Had we jumped straight to development of MACBETH without iterative conceptual designs and team discussions, it is possible the games would have had a poorer approach to bias mitigation.

As goals and conceptual design choices start to solidify, the next key stage of iterative serious game development is to show that the game can work, at least in limited circumstances. Several rapid iterations using mock-ups or storyboarding combined with survey or verbal feedback can help identify unexpected roadblocks, refine the problem space, and show where theory is lacking. In the case of the MACBETH games, for which engagement was a key metric, feedback from practitioners on initial mock-ups led to revisions that made the games more consistently interesting. Experimental results of the first prototype showed that the games had promise, but needed improvement before they could demonstrate real value.

From the beginning, MACBETH was expected to be imperfect in the initial testing. Follow-up changes and experimentation ultimately produced evidence that that game could be engaging and reduce bias, at least in some circumstances. The next stage is to show that the game has real-world value. To accomplish this goal, MACBETH needed to increase its impact in terms of engagement and bias reduction. The games were tested with broader, more diverse samples. More importantly, a version of the game was created as a take-home game that participants were free to play as much or as little as they liked. These procedures allowed the research team to more effectively judge whether the game was engaging and potentially interesting in non-laboratory settings.

Practitioners can also benefit from the use of our framework for evaluating serious games. The principles in our framework provide practitioners guidance for holistic evaluation of serious games. Using the MACBETH project as an example, we illustrate how to apply the framework to a real-life scenario. For example, the case study demonstrates

the process of interactive development, beginning with playtesting using paper prototypes, to using small groups of play testers, to testing the game over time with large populations. We provide an example of grounding the game in theory, using bias-mitigation theory. We also discussed the rapid prototyping process for developing the MACBETH game, and the trade-off decisions between resource cost and functionality that were encountered in technically grounding the software. We discussed the empirical grounding of MACBETH, providing examples of pre-, post-, and follow-up-surveys, psychophysiological testing, game-log analysis, and population selection. Finally, we discuss the process of externally grounding the MACBETH game, partnering with external parties and fulfilling sponsor requirements. These examples constitute specific operational guidance for practitioners who desire to implement our framework.

6. Conclusion

Serious games show promise as an effective training method, but such games are complex and few guidelines exist for their effective evaluation. We draw on the design science literature to develop a serious game evaluation framework that emphasizes grounding evaluation in each of four key areas—theoretical, technical, empirical, and external. We further recommend that serious game developers assume an iterative, adaptive approach to grounding an evaluation effort in these four areas, emphasizing some areas more than others at different stages of the development cycle. We illustrate our framework using a case study of a large-scale serious game development project. The case study illustrates a holistic approach to serious game evaluation that is valuable to both researchers and practitioners.

7. Acknowledgment

This research was supported by the Intelligence Advanced Research Projects Activity (IARPA FA8650-11-C-7178) via the Air Force Research Laboratory (AFRL). The views and conclusions contained herein 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.

8. References

- [1] Dunbar, N.E., Miller, C.H., Adame, B.J., Elizondo, J., Wilson, S.N., Lane, B.L., Kauffman, A.A., Bessarabova, E., Jensen, M.L., and Straub, S.K., "Implicit and Explicit Training in the Mitigation of Cognitive Bias through the

- Use of a Serious Game", *Computers in Human Behavior*, 37(2014), pp. 307-318.
- [2] Jarmon, L., Traphagan, T., Mayrath, M., and Trivedi, A., "Virtual World Teaching, Experiential Learning, and Assessment: An Interdisciplinary Communication Course in Second Life", *Computers & Education*, 53(1), 2009, pp. 169-182.
- [3] Mayer, I., Bekebrede, G., Hartevelde, C., Warmelink, H., Zhou, Q., Ruijven, T., Lo, J., Kortmann, R., and Wenzler, I., "The Research and Evaluation of Serious Games: Toward a Comprehensive Methodology", *British Journal of Educational Technology*, 45(3), 2014, pp. 502-527.
- [4] Nunamaker Jr, J.F., Chen, M., and Purdin, T.D.M., "Systems Development in Information Systems Research", *Journal of Management Information Systems*, 7(3), 1991, pp. 89-106.
- [5] Nunamaker Jr, J.F., and Briggs, R.O., "Toward a Broader Vision for Information Systems", *ACM Transactions on Management Information Systems (TMIS)*, 2(4), 2011, pp. 20.
- [6] Nunamaker Jr, J.F., Twyman, N.W., and Giboney, J.S., "Breaking out of the Design Science Box: High-Value Impact through Multidisciplinary Design Science Programs of Research", *AMCIS 2013 Proceedings*, 2013
- [7] Nacke, L., Drachen, A., and Göbel, S., "Methods for Evaluating Gameplay Experience in a Serious Gaming Context", *International Journal of Computer Science in Sport*, 9(2), 2010, pp. 1-12.
- [8] De Freitas, S., and Oliver, M., "How Can Exploratory Learning with Games and Simulations within the Curriculum Be Most Effectively Evaluated?", *Computers & Education*, 46(3), 2006, pp. 249-264.
- [9] Amory, A., "Game Object Model Version Ii: A Theoretical Framework for Educational Game Development", *Educational Technology Research and Development*, 55(1), 2007, pp. 51-77.
- [10] Astor, P.J., Adam, M.T., Jerčić, P., Schaaff, K., and Weinhardt, C., "Integrating Biosignals into Information Systems: A Neurois Tool for Improving Emotion Regulation", *Journal of Management Information Systems*, 30(3), 2013, pp. 247-278.
- [11] Rakitin, S.R., *Software Verification and Validation for Practitioners and Managers*, Artech House, Inc., 2001.
- [12] Straub, D., Boudreau, M.-C., and Gefen, D., "Validation Guidelines for Is Positivist Research", *The Communications of the Association for Information Systems*, 13(1), 2004, pp. 63.
- [13] Kuechler, B., and Vaishnavi, V., "On Theory Development in Design Science Research: Anatomy of a Research Project", *European Journal of Information Systems*, 17(5), 2008, pp. 489-504.
- [14] Hainey, T., and Connolly, T., "Evaluating Games-Based Learning", *International Journal of Virtual and Personal Learning Environments*, 1(1), 2010, pp. 57-71.
- [15] Shadish, W.R., Cook, T.D., and Campbell, D.T., *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Wadsworth Cengage learning, 2002.
- [16] Burton-Jones, A., "Minimizing Method Bias through Programmatic Research", *MIS Quarterly*, 2009, pp. 445-471.
- [17] Dennis, A.R., and Valacich, J.S., "Conducting Experimental Research in Information Systems", *Communications of the association for information systems*, 7(1), 2001, pp. 5.
- [18] Dunbar, N.E., Wilson, S.N., Adame, B.J., Elizondo, J., Jensen, M.L., Miller, C.H., Kauffman, A.A., Seltsam, T., Bessarabova, E., and Vincent, C., "Macbeth: Development of a Training Game for the Mitigation of Cognitive Bias", *International Journal of Game-Based Learning (IJGBL)*, 3(4), 2013, pp. 7-26.
- [19] Eladhari, M.P., and Ollila, E.M., "Design for Research Results Experimental Prototyping and Play Testing", *Simulation & Gaming*, 43(3), 2012, pp. 391-412.
- [20] Dunbar, N.E., "Credibility Assessment and Intelligence Analysis Training in a Serious Game (Sirius04) - Phase 1 Final Report ", in (Editor, 'eds.): *Book Credibility Assessment and Intelligence Analysis Training in a Serious Game (Sirius04) - Phase 1 Final Report Intelligence Advanced Research Projects Activity*, 2013
- [21] Chaiken, S., "Heuristic Versus Systematic Information Processing and the Use of Source Versus Message Cues in Persuasion", *Journal of personality and social psychology*, 39(5), 1980, pp. 752.
- [22] Stanovich, K.E., and West, R.F., "Advancing the Rationality Debate", *Behavioral and brain sciences*, 23(05), 2000, pp. 701-717.
- [23] Bessarabova, E., Piercy, C., King, S., Vincent, C., Dunbar, N.E., Burgoon, J.K., Elkins, A., Wilson, D.W., Jensen, M., Miller, C., Wilson, S.N., and Lee, Y.-H., "Mitigating Bias Blind Spot Via a Serious Videogame", in (Editor, 'eds.): *Book Mitigating Bias Blind Spot Via a Serious Videogame*, San Juan, PR, 2015
- [24] Einhorn, H.J., and Hogarth, R.M., "Confidence in Judgment: Persistence of the Illusion of Validity", *Psychological review*, 85(5), 1978, pp. 395.
- [25] Dunbar, N.E., Jensen, M.L., Elizondo, J., Piercy, C., Bessarabova, E., Twyman, N.W., Burgoon, J.K., Valacich, J., Adame, B., Miller, C., and Wilson, S., "Measuring Engagement in a Serious Game", in (Editor, 'eds.): *Book Measuring Engagement in a Serious Game*, London, 2013