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, easily grasped, and conducive to repeated play, among other things. Furthermore, these goals must be accomplished within resource constraints and development schedules, with a multifaceted development effort typically requiring collaboration among a large, interdisciplinary team.

Given the high stakes contingent upon a range of budgeting considerations, 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 presented here 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 using a model specifically designed to generate high-impact knowledge through iterative cycles of research activities, including theory development, prototyping, experimentation, and field study. The underlying premise is that, 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 involves 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 developments in this area and explain how they define and corroborate 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, acknowledging 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. Their 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 consists 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. Finally, 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 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. Effective evaluation takes a variety of forms, and typically includes such activities as checking the final product against stakeholder requirements, verifying 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 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 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].

Relevant 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, in addition to comparing pre- and post-game knowledge states, learning outcomes should be evaluated against an appropriate baseline derived from applicable theory. In other words, we suggest 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 advocated in more general systems-building approaches [5]. Initially, 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, and/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 prudently consider the technical practicality of the system requirements as specified by reference theories and stakeholder needs. Only those features deemed initially feasible are pursued and developed. But even

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcomponent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical Grounding</td>
<td>Evaluation requirements are informed by relevant theories. Theory permeates all other areas of evaluation. Outcomes are gauged against relevant, theory-informed baselines.</td>
</tr>
<tr>
<td>Technical Grounding</td>
<td>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.</td>
</tr>
<tr>
<td>Empirical Grounding</td>
<td>Using multiple methods of evaluation to avoid basing conclusions on a single method.</td>
</tr>
<tr>
<td>External Grounding</td>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
<td></td>
</tr>
</tbody>
</table>
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

Effective system development and evaluation strategy requires the use of appropriate, valid, and focused empirical assessment [5]. Thus, a key component of evaluating the effectiveness of serious games entails careful use of controlled laboratory experiments to properly isolate the featured 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 inappropriate 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 non-representative 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 a measure of ecological validity, and 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, theoretically relevant covariates can be measured and statistically controlled for. Mayer et al. [3] offer 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, applying accepted standards of reliability and validity for any measurement instruments used in assessing 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 practical 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 means 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 external grounding, the serious game system should be evaluated in real-world settings. A system evaluation protocol 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, therefore this aspect of the game should 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 over the long term. To this end, a system evaluation protocol should expand beyond immediate tests of training effectiveness vis-à-vis learning outcomes and examine the lasting effectiveness of training mechanisms in facilitating learning [3]. Measuring such longitudinal learning is an expensive and difficult affair, 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 iteration in developing serious games [18, 19]. Iterative elaboration 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 optimize the overall gaming experience. Such a rapid development cycle allows researchers to constantly obtain feedback on the design, playability, enjoyment, and coherence of game scenarios while simultaneously ensuring 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.

The evaluation process can similarly benefit from an iterative, adaptive approach. Evaluation strategies should coevolve 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, further progressing into performance metrics that can be used as a reference point as the game is more broadly tested. Researchers may understand conceptually how they would like a game to perform at the outset of the game development process (e.g., this game should teach players to make fewer biased decisions), but how those goals are operationally evaluated may not become apparent until after several iterations of a game are created and evaluated.

Furthermore, even multiple versions of the same serious game may require very different forms of evaluation to accomplish the overarching goals of a development strategy. 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 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 intelligence analysts could improve their information gathering and assessment abilities. Within the strategy/simulation genre, the MACBETH project development included three different games advanced and evaluated over the course of two years. Each version of the game was similar in terms of basic game mechanics, with players being immersed in a fictional environment where they gather and analyze intelligence data to prevent simulated terrorist attacks. Players had to quickly synthesize the data to create hypotheses about who was planning the attack, how it would be carried out, and where it would take place. If players fail in any of these
aspects, the attack occurs, they are treated to a dismal concluding animation, and then directed to replay the scenario [18]. If players are successful, the attack is averted, they’re rewarded with a triumphant concluding animation, and then allowed to advance to a higher level of the game.

In each scenario, the player has to reconcile carefully crafted bits of data so as not to fall prey to cognitive bias. If players succumb to their cognitive biases within the game (e.g., seek to actively confirm hypotheses rather than disconfirm them, or attribute greater weight to dispositional as opposed to situational factors), their actions are more likely to result in incorrect conclusions and failing the scenario. Players also receive 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 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 different versions 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 our findings, the best performing condition of the core game mechanics (i.e., in-game training, feedback, and game type) was retained in future versions of the game. Thus, at the conclusion of the experiment series, the final game contained core mechanics that had undergone rigorous, 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, while 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 play testing feedback [18, 19]. MACBETH evolved significantly from a limited conceptual version (referred to as 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 into what aspects of the game were most important for evaluation, which were then emphasized in later aspects of the game. 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, thus this design criterion 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. Also 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 gauged according to more stringent performance standards, the tests of learning outcomes were more targeted, and we focused on assessing the games with larger, more diverse groups of players.

The result of this adaptive evaluation approach was a theory-based serious game 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 using longitudinal measurements to estimate their real-world viability. In the sections that follow, we will summarize the ways in which each of the four key areas of evaluation used within 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; thus, the research team consulted the extant bias mitigation literature in designing the games, as well as the techniques used to evaluate them. To this end, the research team drew from the Heuristic-Systematic Model [21] and other dual-process 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 consulted to develop game mechanisms addressing various types of bias [1, 23]. For example, prior research suggested potential interventions useful in promoting more systematic processing on the part 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 the extant bias mitigation literature, and combined with other measures specifically created 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 theory-based comparison methods. A professionally produced, well-conceived training video was used to provide a more traditional state of the art training approach to bias mitigation, and function as a general benchmark for comparing the games’ interactive approach. The training video employed many of the same interventions used in the MACBETH games, albeit within the confines of the more passive 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 another more conventional theory-driven training approach.

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 prompt feedback regarding the most effective design and gameplay mechanisms, this prototyping strategy also enabled the research team to iteratively adjust the evaluation strategy and develop effective metrics establishing performance standards useful for assessing later stages in the development process.

As with many system development efforts, the MACBETH research team was forced to make trade-offs so as 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 its effectiveness experimentally, it became clear the functionality did not improve the game in terms of bias mitigation and training efficiency. Rather than continuing game development along this line of functionality, and further expending resources to create a multiplayer version that might not provide the desired outcome, these features were removed in later development iterations, and key mitigation evaluation requirements were adjusted accordingly. Thus evaluation requirements became more fluid, with initial conceptions of game designs adjusted as early iterations of the games were assessed.

4.6. Empirical Grounding

Learning outcomes were also evaluated using a variety of methods, each designed to capture unique aspects of the games’ effectiveness. The research team used pre- and post-test comparisons of learner performance on tests designed to measure the intended learning outcomes. Additionally, learner performance was re-tested with a follow-up post-test approximately eight weeks after the initial experimental sessions. Several scales were used for assessing bias knowledge and bias mitigation to ensure adequate treatment 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, which were pilot tested before use to assure their reliability as well as their correlation 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, another pilot test was performed to compare them to and correlate them with physiological responses (e.g., heart rate, skin conductance, and eye-tracking). During this process, it became apparent how differences in cognitive absorption were indicated in the self-reported perceptual measures relative to the physiological responses, and it was deemed useful to further employ eye tracking along with perceptual self-report measures to better assess player engagement in subsequent experiments [25].

Players’ behavior while engaged was also recorded and analyzed, and the logs of players’ activities were captured and subsequently examined across the three experiments for evidence of improved game play. These supplementary objective measures provided additional, rich information contributing to a more comprehensive evaluation
than would otherwise be possible with simple pre-test/post-test comparisons of self-report 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 necessitated by 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 further 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, thereby 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 interesting in their own right, the covariates’ primary function in the statistical analyses was to control for their effects so as to better isolate the 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, and manipulation checks, as appropriate. The research team took great care to ensure measurement strategies and experimental procedures were sound, even conducting a special pilot test for the purpose of verifying the measurement procedures, as well as cross-validating several perceptual measures of engagement and flow with physiological measures (e.g., eye-tracking). All of these procedures served to ensure the validity of conclusions drawn during later stages of game evaluation.

4.7. External Grounding

The preliminary game design and initial evaluation were guided by the requirements established by key stakeholders, who provided bias mitigation thresholds the game development was required to target. These thresholds remained in focus throughout development and evaluation, keeping the research team focused on the needs of our sponsoring partners. Additionally, we partnered with a consulting firm to provide professional oversight for developing and testing the authenticity of various aspects of the MACBETH game. Specifically, external guidance was sought to ensure game content was believable and would be consistent with the challenges actually faced by intelligence analysts.

Beyond these methods used to maintain alignment with stakeholder needs, we further grounded our evaluation externally by ensuring the games’ effectiveness under real-world conditions. To this end, single vs. repeated game-playing experiences were compared to assess the games’ playability within subjects over an extended time period. Additionally, the research team employed an 8-week follow-up evaluation to gauge the effectiveness of the 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, to test whether players voluntarily play the game on their own, and if so, does that additional play further increase bias mitigation.

5. Discussion

Serious games show potential for training where traditional instruction methods have been ineffective [22]. Far from a simple solution, however, such games are complex systems requiring substantial resources for successful design, 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 emphasizing 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 considerations, while ensuring that experimental testing is performed with maximal
validity and generalizability. Our development of the MACBETH game demonstrates the application of these guidelines using a case study of a large-scale serious game development effort.

The most important contribution provided by this approach is its unique perspective on serious game evaluation, namely that serious games researchers should assess the effectiveness of their games within the broader context of the game system as a whole. With few exceptions, prior work has examined the process of serious game evaluation more narrowly, focusing primarily 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 advocate 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 serious game literature, which conceptualizes the development and evaluation of serious games as though they occur 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 objectives of the games, and the features that will be used to accomplish them. Had we jumped straight to development of MACBETH without iterative conceptual designs and team discussions, it is likely the games would have lacked an effective approach to bias mitigation.

As goals and conceptual design choices crystallize, the next stage of serious game development is to demonstrate how the game can work effectively within certain perimeters. 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 may best be applied. 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.

MACBETH was expected to be imperfect in the initial testing. Follow-up changes and experimentation produced evidence that that game could be engaging and reduce bias under certain circumstances. The next stage is to demonstrate how the game can have real-world value. 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 this framework in the holistic evaluation of serious games. Using the MACBETH project as an example, we have illustrated how the framework may be applied, demonstrating 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 provided an example of grounding the game in theory, using the heuristic-systematic model of information processing. We demonstrated the use of a rapid prototyping process for developing the MACBETH games, and the trade-off decisions between resource cost and functionality encountered in technically grounding the software. We discussed the empirical grounding of MACBETH by providing examples of pre-, post-, and delayed follow-up-surveys, psychophysiological testing, game-log analysis, and population selection. Finally, we advocated externally grounding the development process by partnering with professional consultants to provide ecological validity while fulfilling sponsor requirements. These examples constitute specific operational guidelines for practitioners wishing to implement our iterative development framework.

6. Conclusion

Serious gaming has demonstrated great potential as a uniquely effective training methodology, but such a technique involves an extremely complex developmental process, with few guidelines available for effective evaluation. This research has drawn 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 serious game developers assume
an iterative, adaptive approach to grounding their evaluation efforts in these four areas, emphasizing a shift in focus as appropriate for different stages of the development cycle. Using a case study of a large-scale serious game development project, we have illustrated our framework within a holistic approach to serious game evaluation that can provide guidance 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


[20] Dunbar, N.E., "Credibility Assessment and Intelligence Analysis Training in a Serious Game (Sirius04) - Phase 1 Final Report!", Intelligence Advanced Research Projects Activity, 2013.


