Method in the Madness: The design of games as valid and reliable scientific tools

Laura Levy, Amy Lambeth, Rob Solomon, & Maribeth Gandy

Georgia Institute of Technology
85 5th Street NW
Atlanta, Georgia, USA 30308
[laura, amy, maribeth]@imtc.gatech.edu

ABSTRACT

Video games have great value for collecting rich data that can be used to study fundamental research questions in fields such as psychology and human-computer interaction. There is an increased use of games as scientific tools to assess and train the cognitive, as well as non-cognitive traits and abilities of players. However, games designed as research tools have a unique and challenging set of considerations so that they meet both the requirements of a scientific protocol, while also maintaining an acceptable gaming experience. Additionally, the success of a research game requires extensive planning and co-creation support for trans-disciplinary teams. This paper describes five critical considerations for the design and implementation of custom-built games as scientifically valid and reliable experimental test beds. These findings resulted from several years of work creating game artifacts to study cognitive training and assessment of individual differences. This class of games presents unique design challenges that are often at odds with traditional game design conventions. However, there are many promising applications and use cases for thoughtfully instrumented games that are capable of generating data with comparable accuracy to traditional and previously validated measures.

CCS CONCEPTS

• **Software and its engineering**~Interactive games • *Human-centered computing*~User studies

KEYWORDS

Video games, game design, validity, games user research, psychology

1 INTRODUCTION

Video games are an increasingly popular tool for use as an experimental method or environment in scientific investigations. There are many advantages that make games a promising format to collect both rich and fine-detailed data for games user research studies. Both custom-built and commercial games allow for repeatable experiences that research participants can engage with, and many commercial games can be modified to control or manipulate a variable of interest. Traditionally, researchers have utilized games to experimentally test for different interaction types and resulting player behaviors. For example, these kinds of studies have investigated how games can support learning [22], benefit cognitive functioning [4, 9], and assess complex skills [21].

While the issues revolving around unsuccessful research games are often examined through the lens of a lack of human, time, and financial resources; there are a suite of other less-discussed considerations contributing to the overall success of games as research tools. Two of the greatest challenges in using games as experimental test beds are that: 1) the implementation is critically reliant on the goals and protocol of the experimental study, and 2) the inherently trans-disciplinary nature of this work requires a diverse expertise that may not be readily available. The result of not addressing these challenges may lead to either enjoyable games that fail to meet the requirements of scientific inquiry (e.g. not producing relevant or reliable data for analysis) or scientifically instrumented games that no one wants to or can play.

The research design and requirements impact every aspect of the design of the game to be utilized – whether that be an off-the-shelf commercial game or a custom-built one. Many questions must be carefully addressed. For example, what game mechanics precipitate the kinds of data that would answer the research questions? What kinds of games are amenable to and appropriate for the target participant demographic? Does the game maintain an acceptable play experience for the time requirements of the research methodology? Does the game align with the different experimental research conditions of the study and, if not, can it be modified to? Subsequently, the aforementioned considerations of the available resources of time, budget, and expertise have significant impacts for successful implementation of these tools. For example, a team must balance their resources and research goals to make decisions on if a commercial game can be used or if they must build one from scratch, or perhaps if a commercial game can be modified and how.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org
FDG18, August 7–10, 2018, Malmö, Sweden
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6571-0/18/08...\$15.00
<https://doi.org/10.1145/3235765.3235793>

With so many of these factors at play, there is little question for why many research games seem inappropriate for either the player or the research team, and in some cases for both parties. As video games are increasingly considered and applied as scientifically rigorous and valid methodological tools, this domain requires more research on the requisite attention when developing and utilizing games as scientifically rigorous experimental designs.

The contribution of this paper is to introduce five critical issues that are often overlooked in the design and use of games as experimental tools. These factors are lessons learned from our development and use of numerous prototypes, custom-built, and commercial games for the purposes of assessment and training investigations. Ultimately, the design of games as research tools is often at odds with traditional “good” game design principles, and this can be a difficult scenario to reconcile. However, in this early work within this domain, we present these factors as a starting point for trans-disciplinary teams to plan, communicate, and possibly mitigate confounds in using games as a scientific research methodology.

2 RELATED WORK

Learning about players through the use of custom-designed “serious” games, or even commercial games, is increasingly popular and has proven useful in multiple cases. Some of the most compelling research has been to learn about players by analyzing their in-game play behavior with data collected by traditional means, such as measuring personality type through a Big Five Inventory (e.g. BFI-44, [10]). For example, Tekofsky et al. [24] found many statistically significant correlations between play-style in *Battlefield 3* [6] and measures on the five domains of the BFI-44. McMahan et al. used a commercial racing game to compare controller input versus natural interactions, and later they reflected on their work to capture insights into the pros and cons of using commercial games for scientific studies [18]. They point out that using modified and existing commercial games has many advantages in games user research; primarily, it negates the cost, resources, and time necessary to design a game from scratch, while supporting reproducibility and potentially removing researcher bias from the design of the game. However, commercial games often do not allow researchers full control over the experience and play options, which can result in unintended and introduced confounds that may affect the data to be collected and analyzed. Building a custom game lets researchers plan out the intended experience and invent interactions that will generate the types of data desired for analysis.

While qualitative outcomes may also be gleaned in these kinds of studies, we focus our discussion on an aspect of game design purposed to capture valid and reliable quantitative data for scientific inquiry. Informed by McMahan et al.’s analysis of research using commercial games, in this paper we are reflecting on our experiences, over several years, of employing two custom-built games for research into individual differences and cognitive abilities of players. In-game behaviors in *Food for Thought*, a multi-tasking cooking game originally part of a project studying cognitive training game for older adults [17], were linked in a statistically significant way to measurements from traditional assessments measuring

individual differences and cognitive ability [14]. *Bubble Trip*, a personality survey-based game, was found to be statistically both internally and externally valid against the pen-and-paper format traditionally used to assess participants [13]. Despite the inclusion of its game mechanics, elements, and theming, the survey responses made by players in *Bubble Trip* appeared to be unaffected. There are many reasons why researchers should expect the mechanics and elements of a game to affect the player, as well as their responses and behaviors that may alter the data to be analyzed within a games study. We have identified five critical factors that we recommend researchers to consider and design towards when creating or using a game for scientific studies.

Player responses are at the core of data measurement for games user research but little is known for how common game design elements, such as randomness and time urgency, may affect the inputs of a player within the game. If, for example, researchers are interested in measuring personality, then game difficulty, theming, types of mechanics, etc. may affect how the traits of a player are “conveyed” through the collected data in unanticipated or undesirable ways.

3 GAMES AS VALID EXPERIMENTAL TOOLS AND METHODS

Video games present attractive advantages over traditional research design tools and methods. Within a game, researchers can create repeatable and controlled environments so that all participants experience the manipulation in the same way. Alternatively, fine aspects of the game can be altered for between-subjects studies to test or control for specific variables of interest. Potential experimental confounds may be mitigated within games (such as reducing feelings of anxiety that are often present in evaluative environments and may affect performance) by “hiding” the assessment within the game (e.g. stealth assessment [21]). Additionally, in some instances, games can simulate scenarios and environments that would be resource-intensive or cost-prohibitive to recreate in real life.

However, many extant research games suffer from deficits in three main areas of expertise: 1) the scientific domain being measured (e.g. cognitive ability, personality), 2) the proper instrumentation and deployment methodology of the assessment, and 3) engaging and appropriate game design. Unfortunately, the lack of co-design between professionals with these expertise often results in games that are fun to play but not useful as research tools, or games that yield scientifically accurate data but that few prefer to engage with.

The games described in this paper have benefitted enormously from a trans-disciplinary team of game designers, computer scientists, psychologists, and quantitative assessment scientists. However, as we have experienced in our own studies, with this complement of expertise comes the realization that the design of scientifically robust games is often at odds with accepted game design practices. For games to be used in experiments, the design and implementation requires careful consideration for how the game elements and mechanics may impact player attitudes and behavior. Every element within a game’s design may have the potential to act as an independent variable on its own that could affect the data,

resulting in reduced validity and reliability of the game. Scientific *validity* describes a quality that a method or instrument has to measure a factor or variable that it purports to measure. For example, a well-made ruler has high validity in measuring distance. Scientific *reliability* describes how consistent that method or tool is in capturing repeated accurate and stable results for the same measure. For example, a scale that reports the same weight over multiple measures for a 1kg object has high reliability. Games present an exciting opportunity to be both highly valid and reliable scientific methods and tools. However, their design and use in this role is drastically different than their normal operation, as will be described.

3.1 Game case study overview

In this paper, two research case studies are used to illustrate the design decisions, considerations, and outcomes in developing and using video games as scientific instruments. *Bubble Trip* (see [13] for more information) was developed in collaboration with a traditional testing company to explore if and how game elements and mechanics may affect player data responses, and how a game may be used as a scientifically valid survey instrument in comparison to its traditional pen-and-paper analog (see Figures 3, 5 & 8). *Bubble Trip* is the culmination of testing with several prototypes that represented different game mechanics and requiring different levels of skill. Because so little is understood about how game elements and mechanics directly impact player intent and data to be used for assessment, *Bubble Trip* was deployed online in three experimental versions to determine the impact of game features on player responses. This systematic investigation of variables within a gamified assessment is the first step to understanding how elements within a more complex game may affect research goals.

Food for Thought was created as a cognitive training game for older adults to ameliorate age-related cognitive decline [1, 17]. *Food for Thought* has since been used as an experimental method to test for participant locus of control, feelings of anxiety, and video game music psychology [14, 15] (see Figures 4, 6, & 7). Both *Bubble Trip* and *Food for Thought* represent years of an iterative participatory design process involving almost 1000 participants (aging from 14 to 82) engaging with the games in both playtesting and experimentation.

4 DESIGN CONSIDERATIONS

4.1 Make the game just “good enough”

Commercial games are generally developed for mass appeal and to particularly delight a subset of users within their demographic. These games are loaded with features, customizations, storylines, Easter eggs, and sophisticated graphics. In contrast, a game intended as a scientific tool likely needs to have much of this stripped away to reduce confounds and afford a repeatable and controlled experience. For example, researchers using commercial games for certain kinds of controlled studies should consider that the communication of pathos and emotionality may impact player affect in unintended ways. Emotions and feelings, as well as cognitive load, may be so altered by the theming of a game as to influence the inputs players make that are later analyzed. Here, we use the psychological term of

affect as a concept rooted in physiological arousal states that have impacts on cognition [2], as opposed to discussions of affect that surround mental boundaries of play experienced by the player (e.g. Huizinga’s “magic circle” [12]). For example, if the game has temporary effects on player affect (through raising basal cortisol arousal levels) then that could result in data capture bias for a personality game (e.g. recording players as being more optimistic about the future or being open to new experiences). Because so little is currently known regarding these kinds of impacts on player assessment, we emphasize the need in returning to the fundamentals of what makes a game “good enough” to scientifically test the influences of these elements, and how they may be mitigated or leveraged into a game as a scientific tool. Though modifications to commercial games have a long history in games user research, there are many benefits to creating custom games for research if the right human and financial resources are available.

Creating a new game, or modifying an existing one, for use in a study has very different design considerations that often go against traditional good game design conventions. There exists a balance between having “enough” of a game that engages the player and produces the desired data, but where the design does not interfere with the validity of using the game as a scientific instrument. Luck, skill, chance, and randomness are common elements in popular games and are often the features that add the most fun and excitement to the play. However, these elements may introduce non-deterministic “noise” in the game data. These elements may introduce a disconnect between the original intention of the player and the actual outcome. If the data collected within the game is to be analyzed to understand some characteristics of the player (e.g. personality type), then any elements that may interfere with the responses of a player (e.g. skill, chance) must be avoided.

In our development of *Bubble Trip*, we were tasked with designing a game experience that would allow us to test the viability of using game mechanics and presentation to answer an existing and validated personality survey (the HEXACO [3]). The need to employ a brief, straightforward, and familiar interaction made the use of a gamified assessment very attractive. Surveys are simple and ubiquitous across social media, and allow for a highly controlled experience while also collecting relevant data for the variable(s) of interest. To understand how presentation modes, game mechanics and elements could impact player responses to a gamified survey, a simple experience that allowed for precise control of game elements and mechanics was the first step to understanding the appropriate design of more complex games for assessment and research.

We created four prototypes representing popular game mechanics (in otherwise simple games) that players used to answer a personality survey. Game clones of Space Invaders, Pachinko, and Angry Birds explored varying levels of skill and chance. To answer the questions, players could make their Likert scale responses by shooting a missile at an invader corresponding to their choice, drop a ball in the slot of their answer, or launch a bird to land on the desired response (see Figure 1).

Figure 1: Three Likert-style survey prototypes cloning well-known games that represent varying levels of skill and chance, (a) Space Invaders, (b) Pachinko, and (c) Angry Birds.

However, these mechanics and presentation styles were found to be too jarring with our playtesting participants. Even with affordances that allowed for greater control over the game response (e.g. “bumping” the Pachinko screen to encourage the ball to fall in the desired bucket), the disparity between the intended and actual responses was too great, which frustrated participants and significantly affected the collected data. The fourth prototype, originally called *Stay in the Lane*, had players maintain their position within a lane corresponding to a Likert scale response. Players read the personality survey question, then moved their icon (pictured as a green dot) into the scale response of their choosing while avoiding obstacles that moved from the bottom to the top of the screen (see Figure 2). Colliding with an obstacle would bump the player towards the top of the screen, moving them further away from their objective of answering the question.

Figure 2: Original prototype of a personality survey-based game, called *Stay in the Lane*.

Stay in the Lane reduces many non-deterministic and non-reproducible elements from gameplay. Random rewards, enemies, and behaviors of physics engines and AIs can all affect how researchers ultimately understand the player and their data. *Stay in the Lane*'s interaction is straightforward and understandable, while the avoidance objective of the game was enjoyable enough to engage players and not significantly affect their responses (as compared to a traditional measure). Through

playtesting, however, we learned that players had a natural desire to collect the obstacles rather than avoid them. In the game's subsequent iteration, *Bubble Trip*, players had the option to collect bubbles as they answered questions. *Bubble Trip* has an underwater theme, where players navigate a fish on the screen to avoid enemy jellyfish, collect bubbles, and – most importantly – read questions from a personality test and make a selection using a five-point Likert scale represented by seashells (see Figure 3).

Figure 3: *Bubble Trip* presents players with survey questions that they answer by touching the colored shell corresponding with their Likert scale response.

Our game design process used a combination of rapid prototyping and qualitative feedback given by student participants, in conjunction with domain knowledge provided by our test design collaborators. In standard game design, emphasis rests on designing an experience that is as fun and engaging as possible. However, the design process of *Bubble Trip* was focused on identifying the simplest theming and mechanics that would satisfy the player and meet our scientific requirements. In this way, *Bubble Trip* was designed to be as un-cluttered as possible from the start of development, while still being enjoyable to play with the expectation that future elements and mechanics could be layered onto it without disrupting core play or introducing experimental confounds.

When designing the game for this study, we also considered two important factors that were derived from our research needs. The first need was scientific validity, as the game needed to reflect and build upon established research in our target domain of assessment design. The second need was in supporting the experiment itself. Here, the game was required to produce meaningful quantitative data from player sessions, in addition to being able to set up distinct and controlled experimental conditions.

The underwater theme was appealing and enjoyable enough for our diverse set of college play testers, while being subtle and inoffensive enough to not affect the validity of the game as a data collection measure (as supported by our statistical analyses into assessment factorial validity and reliability against the traditional measure, see [13]). Theming and aesthetics should be carefully considered [17]. If these elements are jarring or not preferred, then the attitude or mood of the player may be altered, affect their play behavior, and ultimately corrupt the collected data. In a study of

older adults playing *Boom Blox* [5], participant perceptions of the visual art style as “foreign” and “un-American” hindered their enjoyment of and attitudes towards the game. The seemingly innocuous animal characters in the form of blocks were off-putting enough in this demographic to affect their play experience. Similarly, in the cooking-themed, *Food for Thought*, we encountered specific scenarios of unexpected resistance from our population of participants. For example, one sample of predominantly older Jewish adults expressed discomfort for certain recipe levels containing pork products (e.g. cooking bacon for a breakfast, see Figure 4).

Figure 4: *Food for Thought* is a multi-tasking cooking game designed as a cognitive training game for older adults.

Fatefully, this study was also conducted over Passover and participants expressed displeasure in making game recipes that included bread, corn, and rice. Players became insistent that they were making these recipes for their “Christian neighbors”, even though the game narrative involved cooking for a cartoon monkey taste-tester. While it is impossible to accommodate every player with the design, content, and theming of a game, these design decisions can have significant impacts on the experience and play behavior of an individual, even when these design choices (e.g. making a game cooking themed) may seem mundane to the researchers. For games instrumented for scientific use, these impacts on attitude and behavior could tamper with the data used to learn about the player. While unintended consequences of game theming can never be completely predicted and avoided, it is important to recognize their potential impact and take measures to mitigate influences on data, if that is a concern.

4.2 Leverage familiar experiences

Games for research studies are often short experiences so that the experimental manipulation can be engaged with quickly and data collection can begin. This requires that the player have the ability to jump into gameplay with a certain level of proficiency and understanding that reduces their need to learn about the game world and acquire new skills. These kinds of games do not benefit from extensive world building, over-arching narratives, explanations of resources, and character backstories. In fact, researchers should be wary of the influence of these elements to confuse and overwhelm players to the point of affecting their data. Additionally, these elements reduce the accessibility of the game, particularly for older adults [1, 17]. Whereas, a fantasy world containing mages and elves

may be familiar to a seasoned game player of that genre, it could be impenetrable for a casual or non-game player. Using relatable and intuitive environments, interactions, and metaphors allows for many types of players to begin play with little instruction and required learning time. McMahan et al. also point out in their reflections that there can be benefit to the research validity of the study for game mechanics and theming to leverage “familiar real world tasks” (e.g. driving a vehicle [18]).

Therefore, *Bubble Trip* leverages its underwater metaphor to enable study participants to immediately begin playing with only one instruction screen (see Figure 5). We found through playtesting and qualitative feedback in our experiments that floating bubbles, antagonistic jellyfish, and a fish that can swim in any direction are intuitive, understandable, and approachable enough for young and older players (ages 14 to 58) alike to participate.

Figure 5: The only *Bubble Trip* instruction screen.

The cooking theme for *Food for Thought* was chosen for older adults because it is an activity that almost everyone has performed at some point in their lives. People understand that multi-tasking can make cooking more efficient, though performing too many simultaneous tasks can result in recipe errors, like under- or over-done food. This made the feedback on game performance easily interpreted, even by a first-time player (see Figure 6). And in our interviews and focus groups with older adults, many were wistful for a deceased partner that cooked for them or at a loss of their own ability to cook regularly.

Figure 6: The player receives positive feedback from a monkey taste-tester for her hamburger, broccoli, and corn casserole in *Food for Thought*.

A secondary benefit of this minimalistic design approach is that it eases the expertise, time, and resources that researchers would need to create a more complicated game. Storylines, cut scenes, and world building take time to plan, create, and implement while also cluttering the experience for research participants. Streamlining a game in this way not only makes the game more accessible for individuals in a study, it also alleviates strain on developers, artists, and programmers to implement.

4.3 Make them play the “right” way

Complex and rich game worlds provide many ways to play, explore, and accomplish goals. These sophisticated games can be used to learn about player types, strategies, and characteristics (e.g. assessing trait creativity in *Elder Scrolls* [23]), but present complications when more controlled methods are required by specific research goals. Every game element, interaction, and mechanic should be carefully scrutinized for the way it allows players to engage and if this interaction is the one desired by researchers to produce the data of interest. For example, the design of *Bubble Trip* could have been made more fast-paced and exciting but, as we learned in playtesting, this affected the ability of the player to read, understand, and answer the survey questions.

Additionally, many games are vulnerable to cheats, exploits, and glitches in code and gameplay. Some clever, or at least very lucky, players in *Food for Thought* discovered a game glitch that allowed them to interact with both the mini-game in the foreground of the screen while also attending to the cooking ingredients in the background of the game (see Figure 7). The intention was that players would switch back and forth between the mini-game and main game, challenging players’ multi-tasking abilities. Exploiting this game bug allowed players to achieve much higher scores than normal and play in an unintended and unexpected way. Normally in games, where the goal is simply for the player to have fun, unexpected emergent strategies may not be a problem, but when the game is being used to carefully study a particular phenomenon it may taint the experimental data. Initially in the *Food for Thought* study, these players appeared in the data analysis as multi-tasking aficionado outliers and were only exposed as “cheaters” when the data was probed further. In this case, the most important dependent variable for this study – cognitive abilities relating to multi-tasking – was confounded by a subset of participants exploiting a bug and playing in an unintended, though creative, way. A game with a different research goal of identifying player creativity, as measured by unique and rare emergent playstyles, may benefit from keeping such a bug (see [21] for use of game mechanics to measure creativity).

Figure 7: A user plays a spatial reasoning mini-game (based on blending a smoothie) in *Food for Thought* with the main game screen visible in the background.

A research game should be controlled so that all players have similar experiences by encountering the same obstacles, rewards, and stages at the same time. This puts further emphasis on the importance of playtesting with a wide selection of participants of varying experience with games. This has been crucial to our design process in furthering the scientific reliability and validity in *Bubble Trip* across these demographic variables. As previously mentioned in Section 4.1, to achieve this may mean to reducing many elements that would normally be included in a more typical game for entertainment. This may also mean making a game mechanic less efficient to force players into the kind of play that you intend. In *Bubble Trip*, to prevent players from making thoughtless or accidental survey responses, the game requires the player to bump the Likert response shell of their choice twice as confirmation of their response. While this slows down the gameplay and injects an interface element that may be viewed as frustrating by some players, this action helps to ensure that the player intended their response, while also encumbering players by forcing them to read and reflect on the question instead of only focusing on the fun elements of bubble collection and jellyfish avoidance.

4.4 Considering the research design and data collection goals from the beginning of development

For a game intended as an experimental tool, the design of it is highly dependent on the research methodology, study hypotheses, and data collection goals. For a cognitive training game, like *Food for Thought*, multiple sessions over the course of many months were anticipated to be necessary to impart cognitive change in older adults. This requirement necessitated a game that not only represented the types of cognitive tasks that were hypothesized to result in cognitive benefit, but also could be played for an hour at a time every day over the course of three weeks. The gameplay and theming had to be enjoyable, or at least tolerable, enough for this duration of play for a specific demographic. Alternatively, *Bubble Trip*, with its gamified 60-item survey, need only be played once by an individual in a brief web browser encounter. The frequency and duration of game interventions significantly affects the design and experience of a research game. Future research is necessary into the impacts on data validity for games that divide and deliver the assessment or measurement over time in multiple play sessions.

Game design decisions based in academic theories, such as cognitive psychology, also benefit the validity of its use in experiments. General usability and human factors psychology principles, like presentation and placement of important text, benefit the design overall. More impactful decisions may be gleaned from other disciplines, such as assessment design. *Bubble Trip* originally presented the survey question and Likert responses at the bottom of the screen, but play testers complained of difficulties reading the text and then scanning upwards for bubbles and jellyfish. This result makes sense, as the visual system of humans prioritizes scanning from top to bottom and the previous presentation antagonized this convention.

The design, placement, and orientation of the Likert scale was also informed by cognitive psychology and assessment design theories. The shells fit with the underwater theme and the double semantic metaphor of color (red is “negative” and green “positive”) and check or “x” mark (for agree or disagree, respectively) iconography visually assisted intuitive and quick decision-making. The orientation of the Likert scale, either horizontal or vertical, was also discussed. The horizontal layout was used not only because of its similarity to the traditional survey analog, but also to remove a potential bias for players weighting higher placed Likert responses (e.g. strongly agree) as more meaningful than lower placed responses (e.g. strongly disagree) [8]. For *Bubble Trip*, all of these issues were explored early on via easily manipulated low-fidelity prototypes that were approachable to the game designers as well as collaborators from other domains and early play testers (see Figure 8). This seemingly obvious lesson culminated through iterative discussions with stakeholders of many disparate fields, but is commonly violated in many research games – likely due to lack of access to its underlying principles (e.g. assessment design, sensation and perception).

Figure 8: Paper prototypes of *Bubble Trip* allowed the trans-disciplinary team to easily explore alternative game mechanics and screen layouts (e.g. placement of HEXACO question and Likert Scale).

Care was taken in the design and aesthetics of *Bubble Trip* to also remove any elements that could impact the affect or mood of the player that may ultimately influence the way they responded to the questions. Overtly threatening or even cheerful themes, gameplay, and even music could alter the current mood state of the player, as we found through qualitative methods in both playtesting and experimentation. State and temporary changes in feelings, emotions, and moods can all alter the way people feel about themselves and their cognitive abilities [2, 20]. For a survey measuring personality type, ensuring the experimental environment or method does not affect participant mood state is paramount. A game that significantly influences mood, either positively or negatively, may interfere with the validity of the collected data and the reliability of its use over time.

Music represents a complicated set of constituent variables and is a particularly salient element that can affect both mood and cognitive performance, as demonstrated in disciplines outside of games research [7, 11]. While it is accepted that audio and music are important thematic features of engaging games, less is understood for the interaction of the exact musical elements (e.g. tempo, mode) on player physiological responses that result in changes on cognitive functioning within games. This makes music a particularly tricky variable that should be carefully considered in games intended for research. For example, game music with fast tempo was found to result in players making more mouse movements within *Food for Thought*, though these movements were typified as being extraneous to the gameplay itself [15]. Observational accounts of players revealed that many were dancing to the game music and using the mouse to trace over game art to the beat of songs while they waited for ingredients to finish cooking. Players also interpreted the same cooking levels as being more enjoyable with fast music, and interpreted their own performance as higher with faster tempos though that was not the case. Audio has the potential to make a tedious game more enjoyable, but consideration to the effects it may have on player affect and data are essential.

Finally, describing a player data collection plan is beneficial to the design and implementation of a research game overall. This may require identifying the type, frequency, duration, and aggregation of data that will serve to answer the research questions put forth by the study. Substantial effort was made by our team to develop a logging framework for *Bubble Trip*, which allowed researchers to pull a variety of discrete data including numbers of collected bubbles and jellyfish collisions, as well as more complex data such as high frequency avatar position. Capturing as much data as possible, while seemingly enticing, can quickly become burdensome both to researchers who need to analyze it and the logging systems that capture it. Structuring a clear data collection and analysis plan between all stakeholders of developers, designers, statisticians, and other researchers can benefit considerably the kinds of analyses and conclusions drawn. At times, difficult decisions must be made about the kinds and frequency of data collected, and this may be further exacerbated by many unknowns for how players will ultimately engage with the game. For example, it may not be important to capture the large amount of data necessary to deduce cursor dwell times and mousing activity. However, if a critical variable can be reinforced by this data (e.g. ensuring that players were reading

information and not rapidly clicking through) then the effort to build such an infrastructure will greatly bolster the results.

By this methodology, a carefully constructed data collection plan can help to identify previously unexpected sources and types of desired data that can be contrived through carefully designed game mechanics and interactions. For example, the floating bubbles in *Bubble Trip* not only provide for a different way of play but also allow researchers to know when players are still actively engaged with the game but not responding to the question yet (as opposed to leaving the browser game window open and unattended), while also generating data that can be used to make correlations between their survey responses and fish movement/bubble collection tendencies. Similarly, *Food for Thought* has screens that gate players from the next step in the game. This serves to discourage players from mindlessly clicking rather than playing strategically and allowed us to collect dwell time information that informed researchers on different player strategies (e.g. do certain types of players spend a lot of time planning their approach by dwelling on the level's upcoming "recipe", or looking at an in-depth feedback screen regarding their score post hoc (see Figure, 6) and how does this affect their subsequent performance?).

Ideally from the start of development of a research study game, the research goals, data collection and analysis plans should help to inform the game's design, functionality, and operation ultimately supporting the final analysis of the data that will be used to learn about the player.

4.5 Support trans-disciplinary co-creation of games

A scientifically instrumented and designed game requires expertise from many disciplines including game design, research methods, and statistics. The danger of not involving the expertise from relevant stakeholders may result in a product that is playable but not valid, or a game that is scientifically valid but unappealing to play. For an assessment game designed to learn about the player, like *Bubble Trip*, an existing and validated scientific instrument was gamified to begin nascent and selective research into this domain. Many surveys and test instruments are carefully ordered so that questions are presented in groupings that load on different factors. For example, the 60-item HEXACO used in *Bubble Trip* has 10 interspersed questions that measure six factors of personality. Re-ordering these questions or changing the presentation style breaks the intended design of the survey and there would be the expectation that overall results would also be affected. Removing questions to shorten a test or restructuring the response scale (e.g. change a 7-point Likert scale to 5-points) all affect the validity and reliability of the instrument, as well as reducing the ability to compare its results against population data known for the original method. Our involvement of expertise from assessment designers, statisticians, and cognitive psychologists helped to ensure that the design of the assessment would not be unwittingly broken when turned into a game.

With so many professionals from varied fields collaborating asynchronously and remotely, there was a great need to leverage processes and develop tools that facilitated team member contributions. For *Bubble Trip*, we created a web-logging framework called *Gloggr* (see [13] for more information) with a simple graphical

user interface that allowed all researchers, regardless of programming experience, to query the logging database for their specific needs at any time convenient to them. Though the development of *Gloggr* involved significant time and resources, the investment in its framework can be used for future projects and unburdens developers from the time previously needed to fulfill data requests from non-technologist collaborators.

Food for Thought, due to the participation time required by the experimental design, required extensive level building so that participants would play only new levels over the 15 hours of the experiment. A large team of undergraduate psychology majors were available to design levels, but had insufficient programming experience to reliably code within Unity3D. The solution was to create a Unity3D level designer so that students with no programming experience could create unique, challenging, and appropriate game levels at a rapid pace to contribute to the overall game. Students imagined or looked up cooking recipes they wanted to create within the game, located the corresponding art assets, and used the level designer to set cooking station orders, time pressures, and mini-games. The user-friendly level creator resulted in the swift production of hundreds of student-generated cooking levels that minimized development time and allowed for quick implementation of the study.

When teams are made of members with varying proficiencies in programming, coding, and development, tools that support all levels of contribution and access are beneficial to the research goals of the project. Team management software, like Confluence, is also useful for sharing literature, design documents, and justifications of scientific requirements to inform on-going game design so that all involved can be well versed in a complex subject matter. Inter- and trans-disciplinary teams offer extensive expertise and experience to a project, and the most benefit is received when all stakeholders are involved from the beginning of development. When resources allow, development of accessible tools and platforms to support dissemination of data, results, and information can result in well-informed game designs that produce highly valid and reliable scientific data.

5 DISCUSSION

One of the most significant challenges in designing games for scientific studies is the tension between including enough game-like elements that produce an engaging game, but also selecting the right elements that will not interfere with the validity and reliability of the game as a scientific method or tool. Despite the simplicity and stripped-down nature of *Bubble Trip*, exit questionnaires revealed that our sample of over 400 participants found the game to be enjoyable and rewarding (mean = 5.9 of 7 points, 7 = strongly agree, "Did you enjoy playing this game?" and similar questions). Likely, this is because of the short play duration and neutral theming of *Bubble Trip*. Even *Food for Thought*, a game with highly repetitive levels and a significant amount of playtime, was generally liked, or at least tolerated, by participants playing it for 15 hours within a study. By sacrificing standard game design conventions, an "enjoyable enough" game can be produced that still engages participants while also collecting valid and reliable scientific data. Compromising

standard game design conventions does not necessarily have to result in a boring, yet scientifically useful, game. Balancing the research goals, intended duration of playtime, and demographics can lead to informed decisions producing an entertaining and appropriate game that yields accurate scientific results.

In addition to close considerations for how the game mechanics and elements may impact player behaviors and responses, we suggest a design that allows for gameplay that is not inherently tied to the interactions that produce data for the variable(s) of interest. For example, *Bubble Trip* allows for a player to effectively “fail” at playing the game (e.g. by not collecting bubbles, colliding with jellyfish) but this does not impact the reliability of the personality survey responses or reveal itself in frustrations in gameplay through the exit questionnaire. This degree of separation between the gameplay and assessment was important in this case to generate valid survey results and allow for multiple styles of gameplay, while largely removing the effects of gameplay on the survey responses.

The trans-disciplinary nature of games for science necessitates buy-in and expertise from a variety of professionals. Based on theories of executive control and cognitive change, the design of *Food for Thought* was informed by classical theories of cognition and successful interventions from fields as varied as improvisational comedy [19] and taxi driving [16]. All too frequently in serious game design, only one kind of expertise is represented. Participatory design processes help to eliminate issues where, for example, a great game is built that has no scientific merit or a scientifically rigorous game is built that no one wants to play. However, even more importantly is the nuance that can exist in the design of scientifically valid games that can be easily overlooked. For example, innocently invalidating a carefully constructed scientific instrument (e.g. survey assessment) by omitting or reordering certain portions.

5 CONCLUSION

Games are an exciting and promising environment for data generation and collection because of the very components that make them game-like. Games should be fun, engaging, and rewarding to play. In some research situations, these elements may serve to reduce anxiety in the participant and free them to play in a natural way that generates data that most reflects a truth about who the player is. The controllable and repeatable nature of custom-built, or even some modified commercial games, creates rich environments to collect finely detailed data to learn about player characteristics and strategies. In this paper we presented five considerations for designers of games for research:

- *Make the game just “good enough.”* The goal of this type of game design is not to create a wildly popular “AAA” title, but rather to create a game intervention that is just engaging enough that participants will play it, while providing a player experience that meets the stringent requirements of a controlled scientific experiment. This is not only due to the practical need to limit scope and required resources for game implementation, but to eliminate potential confounds in the experiment.
- *Leverage familiar experiences.* To reduce the learning curve, the experimental game may use elements or interfaces that the target demographic is already familiar with, either from real life (e.g. cooking and driving) or from well-known game tropes (e.g. Pac Man and Space Invaders). A caveat is that the designers should ensure that their theming choices, no matter how banal they may initially seem, are acceptable and appropriate for the target demographic, as even familiar themes may have negative connotations for some.
- *Make them play the “right” way.* In games intended for entertainment, it may actually improve a player experience when unanticipated strategies or play styles emerge over time (e.g. “speed runs” of games where errors in the game physics engine allow players to take shortcuts through levels). However, if the game interface and mechanics were carefully designed with additional goals in mind (e.g. physical rehabilitation, cognitive training, personality assessment) then such “cheats” may nullify these aspects of the experience. Careful design and subsequent testing is required to ensure that players will only be able to play in the intended manner.
- *Consider the research goals at the beginning.* Game mechanics can be designed in such a way that they allow for collection of additional data that will help to shed light on later research results (e.g. specifically what aspects of a game experience actually caused positive cognitive training outcomes). And knowledge from the relevant “serious” domain (e.g. cognitive psychology, educational theory, medicine) may have implications for all aspects of the game design (e.g. design of multi-modal stimuli, motor skill requirements for the interface, educational pedagogy). However, the research questions, experimental methods, desired data, and potential analysis approaches must be identified and carefully integrated into the game design process from the beginning.
- *Support trans-disciplinary co-creation.* Given the importance that input from other fields may have in this type of games research, it is important to identify processes and tools that support true trans-disciplinary co-creation. It can be challenging to help collaborators from an entirely different field, who may have little experience with game design (or even playing games) transfer their domain knowledge in an actionable manner. However, if their expertise is not leveraged substantively in the game design process it is likely that the resulting game will have elements that impede the intended research.

While specific challenges exist for the development and implementation of games as scientific tools and methods, this does not negate the creation of entertaining and engaging games that still produce valid scientific data. With a deliberate game design process considering the unique requirements of scientifically valid games, a new suite of powerful tools can be added to the arsenal of researchers.

6 FUTURE WORK

More research is required to understand how game elements and mechanics affect player data, and how sensitive this data is to erosion by a game. For example, to what extent do theming, aesthetics, and color choices affect participant mood, cognitive performance, and skill development? How do both diageic and non-diageic game sounds and musical scores impact the ability of a player to make assessment responses that would be valid against a traditional collection measure? There are likely many levels of game elements and mechanics that can be manipulated to balance the enjoyableness of a game with the production of valid results. In *Bubble Trip*, we created three versions of the game representing different levels of game elements and mechanics. We found the “full version” (containing item collection, enemy avoidance) to be just as scientifically and statistically valid as the traditional instrument, while also being preferred over much more lean versions. Our next steps are to test *Bubble Trip* with other kinds of assessments (e.g. measuring academic achievement predictors or cognitive abilities), incrementally add in other game elements (e.g. time pressures, aesthetics, music styles), or modify the survey instrument itself (e.g. different Likert-scale representations, removing questions) to see where the game loses its validity as compared to traditional collection measures. Future research into the effects on scientific validity and reliability that game elements and mechanics present will allow for more engaging and useful scientific game tools and methods.

ACKNOWLEDGMENTS

This work is funded through grants from ACT, Inc. and NSF# 0905127.

REFERENCES

- Allaire, J.C., McLaughlin, A.C., Trujillo, A., Whitlock, L.A., LaPorte, L., Gandy, M. 2013. Successful aging through digital games: Socioemotional differences between older adult gamers and Non-gamers, *Computers in Human Behavior*, Volume 29, Issue 4, Pages 1302-1306.
- Ashby, F.G. and A.M. Isen. 1999. A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*. **106**(3): p. 529.
- Ashton, M.C. and K. Lee. 2009. The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*. **91**(4): p. 340-345.
- Basak, C., et al. 2008. Can Training in a Real-Time Strategy Video Game Attenuate Cognitive Decline in Older Adults? *Psychology and Aging*. **23**(4): p. 765-777.
- EA. 2008. *Boom Blox*. Los Angeles, USA.
- EA. 2011. *Battlefield 3*. Stockholm Sweden.
- Furnham, A. and K. Allass. 1999. The influence of musical distraction of varying complexity on the cognitive performance of extroverts and introverts. *European Journal of Personality*. **13**(1): p. 27-38.
- Göbel, S.M. 2015. Up or down? Reading direction influences vertical counting direction in the horizontal plane – a cross-cultural comparison. *Frontiers in Psychology*, **6**, 228.
- Green, C.S. and D. Bavelier. 2006. Effect of action video games on the spatial distribution of visuospatial attention. *Journal of Experimental Psychology. Human Perception and Performance*. **32**(6): p. 1465-78.
- John, O.P. and S. Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: theory and research*. **2**(1999): p. 102-138.
- Kellaris, J.J. and R.J. Kent. 1992. The influence of music on consumers' temporal perceptions: does time fly when you're having fun? *Journal of Consumer Psychology*. **1**(4): p. 365-376.
- Stenros, J.. 2014. In defence of a magic circle: the social, mental and cultural boundaries of play. *Transactions of the Digital Games Research Association*, **1**(2).
- Levy LM, Moore, J., Liu, R., Way, J., Wilson, J., Johnson, J., Lambeth, A., Solomon, R., and Gandy, M. 2016. Grouches, Extraverts, and Jellyfish: Assessment validity and game mechanics in a gamified assessment. *DiGRA/FDG 2016: 1st Joint International Conference of DiGRA and FDG*, August 1-6, 2016, Dundee, Scotland, UK
- Levy, LM., Solomon, R., Moore, J., Way, J., Liu, R., and Gandy, M. 2015. Actions Speak Louder Than Words: An exploration of game play behavior and results from traditional assessments of individual differences. *Proceedings of the 10th International Conference on the Foundations of Digital Games (FDG 2015)*, June 22-25, Pacific Grove, CA
- Levy, LM, Solomon, RES, Gandy, M., & Catrambone, R. 2015. The Rhythm's Going to Get You: Music's effects on gameplay and experience. *Paper presented at CHI Play 2015, London, U.K.*
- Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S. J., & C.D., F. 2000. Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences of the United States of America*. **97**(8), 4398-4403.
- McLaughlin, A. C., Gandy, M., Allaire, J. C., & Whitlock, L. A. 2012. Putting fun into video games for older adults. *Ergonomics in Design: The Quarterly of Human Factors Applications*, **20**, 13-22.
- McMahan, R.P., Ragan E.D., Leal, A., Beaton, R.J., Bowman, D.A. 2011. Considerations for the use of commercial video games in controlled experiments, *Entertainment Computing*, Volume 2, Issue 1, Pages 3-9, ISSN 1875-9521
- Noice, H. and T. Noice, What studies of actors and acting can tell us about memory and cognitive functioning. *Current Directions in Psychological Science*, 2006. **15**(1): p. 14-18.
- Sedikides, C. 1992. Changes in the valence of the self as a function of mood. *Review of Personality and Social Psychology*. **14**: p. 271-311.
- Shute, V.J. 2011. Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*. **55**(2): p. 503-524.
- Shute, V.J. and F. Ke. 2012. Games, learning, and assessment, in *Assessment in Game-Based Learning*. Springer. p. 43-58.
- Shute, V.J., et al. 2009. Melding the power of serious games and embedded assessment to monitor and foster learning, in *Serious Games: Mechanisms and Effects*. Routledge/LEA Philadelphia, PA. p. 295-321.
- Tekofsky, S., Spronck, P., Plaat, A., van den Herik, J., Broersen, J. 2013. Personality Assessment Through Gaming Behavior. *Foundations of Digital Games*. Greece.