IMPROVING THE DATA QUALITY OF ATTENTION-BASED COGNITIVE TASKS
THROUGH THE USE OF GAMES

A Thesis Submitted to the
College of Graduate and Postdoctoral Studies
In Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy
In the Department of Computer Science
University of Saskatchewan
Saskatoon

By

KATELYN WILEY

© Copyright Katelyn Wiley, March 2024. All rights reserved.
Unless otherwise noted, copyright of the material in this thesis belongs to the author
Permission to Use

In presenting this thesis in partial fulfilment of the requirements for Degree of Doctor of Philosophy in the Department of Computer Science, University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor who supervised my thesis work, Dr. Regan Mandryk, or in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Computer Science
176 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada

OR

Dean College of Graduate and Postdoctoral Studies
116 Thorvaldson Building, 110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9 Canada
Abstract

Cognitive tasks have many uses as scientific tools, from helping us understand how the human mind works to informing mental health assessments. But these tools are not perfect—they face problems of attrition for research and with patient cooperation for clinical use. Participants often exert suboptimal effort, leading to poor-quality data. To address these problems, researchers have turned to games. When cognitive tasks are gamified, users may find the previously boring tasks become interesting, and feel more willing to engage with the tasks and motivated to put forth their best effort. However, there are still many unanswered questions about how game-based assessments work. In this dissertation, we ask the question: How do games for the assessment of attention affect the quality of gathered data, in terms of both the experience of participants and their performance?

We systematically reviewed the literature to identify how game-based assessments of attention are made and measured. Our results provided an overview of the field, further research questions, and a guide for how to approach future studies. We found that many assessment games are not sufficiently evaluated to determine the quality of data they produce. We then conducted two studies to examine how game elements affect performance on and enjoyment of an attention-based cognitive task. These studies provided guidance on which game elements are most likely to improve data quality, in terms of increasing participant enjoyment (and therefore engagement with the task) and performance (how quickly and accurately they respond to the task). Finally, we examined how the game elements identified in the previous studies affect the reliability of an attention-based cognitive task.

To advance the use of game-based assessment, we need robust literature that examines individual game elements and their effects of all aspects of the assessment—the user experience, the behavioural performance of users, and the psychometric properties of the tasks. Our studies add to these structural building blocks of knowledge. We also offer models for how to design studies that adequately measure both user experience and performance, and ultimately the data quality of these tasks.
Acknowledgements

I owe many thanks to my supervisor, Regan Mandryk. I went to my first meeting with Regan thinking that I would take a couple of Computer Science courses as a part-time thing, and left the meeting feeling inspired to pursue a PhD. I’ve learned so much about writing, conducting research, thinking critically, and being a good person from Regan, and she has been endlessly patient and supportive as I’ve navigated the ups and downs of life.

I would also like to thank the members of my committee: Carl Gutwin, Cody Phillips and Steven Prime, and my external examiner, Karyn Moffatt. They have been instrumental in improving my work.

Everyone in the University of Saskatchewan Interaction Lab was delightful to work with. I would like to give particular thanks to those who directly collaborated with me in this work, including Raquel Robinson, Sarah Vedress, Phaedra Berger and Maxmilian Friehs.

NSERC and SWaGUR provided funding throughout my studies, for which I am very grateful.

Thank you to my family: my mom, my sisters, my wonderful in-laws. Thank you to my husband, Jake Ursenbach, and Picard (my very good dog). Words aren’t enough.

Finally, this dissertation is dedicated to my dad, who was with me every step of the way.
Table of Contents

Permission to Use ........................................................................................................... i
Abstract ............................................................................................................................ ii
Acknowledgements .......................................................................................................... iii
Table of Contents .............................................................................................................. iv
List of Tables .................................................................................................................... viii
List of Figures .................................................................................................................... ix
List of Abbreviations ....................................................................................................... x

1 Introduction .................................................................................................................. 1
  1.1 Research Problem ..................................................................................................... 1
      1.1.1 Cognitive Tasks ................................................................................................. 1
      1.1.2 Gamification of Cognitive Tasks ......................................................................... 2
      1.1.3 Benefits of Gamification .................................................................................... 2
      1.1.4 Problems with Gamification .............................................................................. 3
      1.1.5 Theories of Gamification ................................................................................... 4
  1.2 Research Questions ................................................................................................... 5
  1.3 Overview of this Dissertation .................................................................................. 6
      1.3.1 Contribution to Each Manuscript ...................................................................... 7

2 Manuscript A: The Making and Evaluation of Digital Games Used for the Assessment of
Attention: Systematic Review ......................................................................................... 8
  2.1 Introduction to Manuscript A .................................................................................... 8
  2.2 Abstract .................................................................................................................... 9
  2.3 Introduction .............................................................................................................. 10
      2.3.1 Attention ........................................................................................................... 10
      2.3.2 Assessment of Attention .................................................................................. 11
      2.3.3 Digital Games for the Assessment of Attention ............................................... 11
      2.3.4 The Present Research .................................................................................... 13
  2.4 Methods ................................................................................................................... 14
      2.4.1 Eligibility Criteria ............................................................................................ 14
      2.4.2 Information Sources and Search Strategy ....................................................... 15
      2.4.3 Study Records and Data Management ............................................................. 16
      2.4.4 Data Items and Synthesis .............................................................................. 16
  2.5 Results ..................................................................................................................... 17
      2.5.1 Search Results ................................................................................................ 17
      2.5.2 Summary of Included Studies ........................................................................ 18
          2.5.2.1 Publishing Formats and Dates ................................................................. 18
          2.5.2.2 Study and Participant Characteristics .................................................... 20
      2.5.3 Why are digital games used for assessment? .................................................... 20
      2.5.4 How are digital games used for assessment made? .......................................... 21
          2.5.4.1 Development process ............................................................................. 21
             2.5.4.1.1 Gamification ....................................................................................... 21
             2.5.4.1.2 Custom Games ................................................................................. 22
             2.5.4.1.3 Commercial Games ......................................................................... 22
          2.5.5 How are digital games used for assessment measured? ................................ 23
3.4.2 Gamification of Cognitive Tasks for Assessment
3.4.2.1 Points and Feedback
3.4.2.2 Sound Effects
3.4.2.3 Graphics
3.4.2.4 Thematic Elements
3.4.3 On Choosing a Cognitive Task and Game Elements
3.4.3.1 Choosing a Cognitive Task
3.4.3.1.1 Gamification of the Dot Probe Task
3.4.3.2 Choosing Game Elements
3.4.4 The Present Research

3.5 Study 1
3.5.1 Methods
3.5.1.1 Tasks
3.5.1.2 Stimuli
3.5.1.3 Procedure
3.5.1.4 Analysis
3.5.2 Results
3.5.2.1 Performance
3.5.2.2 Experience
3.5.3 Discussion of Study 1

3.6 Study 2
3.6.1 Methods
3.6.1.1 Procedure
Contribution to this Dissertation ................................................................. 61

4 Manuscript C: Measuring the Reliability of a Gamified Stroop Task: Quantitative Experiment ................................................................. 62

4.1 Introduction to Manuscript C ...................................................................... 62

4.2 Abstract ........................................................................................................ 62

4.3 Introduction .................................................................................................. 64

4.3.1 Background .......................................................................................... 64

4.3.2 Cognitive Task Gamification ................................................................. 64

4.3.2.1 Enjoyment and Motivation ............................................................... 65

4.3.2.2 Performance ...................................................................................... 66

4.3.3 Psychometric Properties of Gamified Tasks ........................................... 67

4.3.4 This Study ............................................................................................... 68

4.3.4.1 The Stroop Task ............................................................................ 68

4.4 Methods ....................................................................................................... 69

4.4.1 Tasks ........................................................................................................ 69

4.4.2 Participants ............................................................................................. 70

4.4.3 Statistical Analysis .................................................................................. 71

4.5 Results ......................................................................................................... 72

4.5.1 Participants ............................................................................................. 72

4.5.2 Intrinsic Motivation Inventory ............................................................... 73

4.5.3 Reaction Time and Error Rate Data ..................................................... 73

4.5.4 Internal Consistency .............................................................................. 78

4.5.4.1 Reaction Time ................................................................................ 78

4.5.4.2 Error Rate ........................................................................................ 79

4.5.5 Test-Retest Reliability .......................................................................... 80

4.5.5.1 Reaction Time ................................................................................ 80

4.5.5.2 Error Rate ........................................................................................ 82

4.6 Discussion .................................................................................................... 84

4.6.1 Summary and Explanation of Findings .................................................... 84

4.6.1.1 Performance .................................................................................... 84

4.6.1.2 Enjoyment ....................................................................................... 86

4.6.2 Limitations and Future Work ................................................................. 86
List of Tables

Table 1. Included and excluded cognitive functions for the eligibility criteria in our review..... 15
Table 2. References to all included papers by publication field.............................................. 19
Table 3. Results of the statistical tests for Study 1 and 2. ......................................................... 48
Table 4. Descriptive statistics for reaction time and error rates, at Time 1 and 2 for each task type.................................................................................................................................... 75
Table 5. ANOVA Summary Table for Reaction Time Distribution............................................ 76
Table 6. ANOVA Summary Table for Reaction Time................................................................ 77
Table 7. ANOVA Summary Table for Error Rate........................................................................ 77
Table 8. Repeated Measures ANOVA Summary Table for Reaction Time and Error Rate ....... 78
List of Figures

Figure 1. Flowchart of included and excluded records through our review process. .................... 18
Figure 2. Number of included papers by publication year and field. ....................................... 19
Figure 3. Top left: Facial stimuli, with and without the art filter. Top right: Points only condition, after a correct response. Bottom left: Points and theme condition, after a correct response. Bottom right: Part of the introduction to the theme condition .................................................. 45
Figure 4. Estimated marginal means of enjoyment ratings, before, during and after the task. .... 52
Figure 5. Game version of the task, after a correct response was entered ................................. 70
Figure 6. Histograms of reaction time by time point and task type, for each type of trial condition. .............................................................................................................................................. 76
Figure 7. Internal consistency of reaction time cost for each time point and task type ............ 79
Figure 8. Internal consistency of error rate cost for each time point and task type ................. 80
Figure 9. Test-retest reliability of reaction time cost as the number of trials increases, for each task type. .......................................................................................................................................... 81
Figure 10. Test-retest reliability of reaction time as the number of trials increases, for each trial type and task type ........................................................................................................................................ 82
Figure 11. Test-retest reliability of error rate cost as the number of trials increases, for each task type ........................................................................................................................................ 83
Figure 12. Test-retest reliability of error rate as the number of trials increases, for each trial type and task type. ........................................................................................................................................ 84
### List of Abbreviations

(in alphabetical order)

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABM</td>
<td>Attentional Bias-Modification</td>
</tr>
<tr>
<td>ADHD</td>
<td>Attention-Deficit/Hyperactivity Disorder</td>
</tr>
<tr>
<td>CI</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>HCI</td>
<td>Human Computer Interaction</td>
</tr>
<tr>
<td>ICC</td>
<td>Intraclass Correlation Coefficient</td>
</tr>
<tr>
<td>IMI</td>
<td>Intrinsic Motivation Inventory</td>
</tr>
<tr>
<td>MMSE</td>
<td>Mini Mental State Examination</td>
</tr>
<tr>
<td>MOT</td>
<td>Multiple Object Tracking</td>
</tr>
<tr>
<td>MTurk</td>
<td>Amazon Mechanical Turk</td>
</tr>
<tr>
<td>NASA TLX</td>
<td>NASA Task Load Index</td>
</tr>
<tr>
<td>PANAS</td>
<td>Positive and Negative Affect Schedule</td>
</tr>
<tr>
<td>PXI</td>
<td>Player Experience Inventory</td>
</tr>
<tr>
<td>RT</td>
<td>Reaction Time</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SDT</td>
<td>Self-Determination Theory</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Research Problem

1.1.1 Cognitive Tasks

Imagine sitting in front of the computer, with simple instructions on the screen: “For this Go-No-Go task, if you see an image of a green object, press the spacebar as quickly as you can. If you see a red object, do nothing.” The task begins and you respond to the first image. Another image appears. You respond. Another image appears. You respond. Ten minutes later, you have responded to a total of 300 images. Did you enjoy those ten minutes? Were you focused on the task the entire time? Did you always respond as quickly as you could? Would you be willing to do the task again?

Cognitive tasks, like the Go-No-Go task described above, have a long history of being used as scientific tools. In 1890, James Cattell introduced the term ‘mental tests’ to refer to a battery of ten tests designed to measure aspects of human cognition like memory, reaction time, and sensation and perception [1]. He proposed administering these tests to a large number of people and using the results to understand mental processes on both a population and individual level. Indeed, what we now call ‘cognitive tasks’ are used for furthering the field of psychology and understanding cognition, along with developing population norms, informing mental health assessment, and even treating disorders [2].

As an example, Cattell’s work led to the creation of the Stroop task, one of the best-known paradigms in cognitive psychology [3]. Cattell first observed that people take much longer to name images and their properties than they do to read corresponding words [4]. Building on this knowledge, John Ridley Stroop created a task that contrasted naming the colour of coloured rectangles with incongruently coloured words (e.g., the word ‘blue’ printed in yellow ink) [5]. Subsequent studies and methodological innovations [6–8] have led to the Stroop task as we know it today, which has been used to study attention [9] and in clinical contexts [10,11].
The Stroop task has also been further modernized for administration via computer. Computerized versions of tasks are especially valuable, as they allow for remote testing and automatic scoring, making it faster and easier to collect data. Digital capabilities also provide opportunities for further innovation of cognitive tasks [12].

1.1.2 Gamification of Cognitive Tasks

One innovation has been the gamification of cognitive tasks. Gamification, the use of game elements in non-game contexts [13], has been used to motivate behaviour in contexts like education, physical fitness, and organizational settings. Both extrinsic and intrinsic motivation can be targeted through gamification [14], with the goal of improving user engagement with less appealing tasks. For example, gamification has been used to motivate users and improve performance on microtasks, like image labeling [15–17].

1.1.3 Benefits of Gamification

A recent article in *Nature* called for the widespread adoption of gamified tasks by behavioural science [18]. Applying game elements to cognitive tasks is an appealing idea, largely because cognitive tasks are inherently boring and repetitive, and well-designed games are fun and novel. Traditional cognitive tasks face problems of attrition for research and with patient cooperation for clinical use [19], and participants often exert suboptimal effort, leading to poor-quality data [20,21]. When cognitive tasks are gamified, users may find the previously boring tasks become interesting, and feel more willing to engage with the tasks and motivated to put forth their best effort.

Naturally engaging tasks will appeal to more diverse and larger populations. Traditionally, researchers in Human-Computer Interaction (HCI) and psychology have recruited participants with “WEIRD” demographics (participants who are Western, Educated, Industrialized, Rich and Democratic) [22], or even more narrowly, participants who are psychology undergraduate students [23]. Incentives to participate in lab-based research can be limited, whereas games can be inherently motivating—as an example, 4.3 million people played a game called ‘Sea Hero Quest’, informing research on dementia and navigation [24,25]. Games can also facilitate
engagement with children, who may lack the ability to endure the length and repetition associated with traditional tasks [26].

The ecological validity of cognitive tasks can also be improved through games. For example, while studying language, researchers found that participants engaged in a game produced more natural dialogue, and thus higher-quality data [27]. When Vanden Abeele et al. [28] compared two games designed to measure psychoacoustic thresholds in preschoolers, they found that the more fully developed and motivating game was able to detect lower thresholds.

1.1.4 Problems with Gamification

While games have been offered as a solution to many of the pitfalls of traditional cognitive tasks [18], the utility of gamification has long been questioned. In 1999, Amy Bruckman compared the gamification of educational software to ‘chocolate-covered broccoli’, noting that it often is neither educational nor fun [29]. The same problem has been discussed in the context of cognitive tasks. Levy et al. [30] note that often, when using games as scientific tools, we see either enjoyable games that fail to produce reliable, useable data or scientifically useful games that are not appealing to play.

Cognitive task development is already a sensitive process. The ‘Standards for Educational and Psychological Testing’ [31] present factors that should be considered through test development and delivery. The interpretations of task scores need to be valid for their intended use, and task scores also need to be consistent across replications. Even minor changes to a task, like slight changes to the stimuli used, must be rigorously studied before the task can be used [32].

Gamifying tasks may be especially complex as game elements will inherently affect aspects of cognition. For example, adding a point scoring system that is displayed on the screen increases the amount of information that a user must attend to during a cognitive task. Increasing narrative suspense has been linked to narrowed attentional focus [33]. Even framing a task as a game, without any functional changes to the task itself, may influence how people respond to a task [34].
1.1.5 Theories of Gamification

Gamified tasks need to be scientifically robust, and they are also often expected to offer a more engaging user experience. One prominent theoretical framework for why gamification can be motivating is Self-Determination Theory (SDT), which posits that games can address the core needs of competence, autonomy, and relatedness [35]. When a task offers the opportunity to meet these needs, users will be intrinsically motivated to engage with it. SDT differentiates intrinsic motivation from extrinsic motivation, in which users are motivated to engage due to separate outcomes, like rewards of money or praise.

Both extrinsic and intrinsic motivation can be targeted through gamification. Nicholson uses the categories of meaningful gamification (intrinsically motivating) and reward-based gamification (extrinsically motivating) [14]. One type of gamification is not necessarily better than the other—it depends on the context of use. Meaningful gamification is more likely to lead to long-term change of behaviours and beliefs, whereas reward-based gamification is only appropriate for immediate and short-term changes.

Different game elements will facilitate intrinsic versus extrinsic motivation. Points, levels, leaderboards, badges, and other achievements all target extrinsic motivation [14]. These elements are easily implemented in any system, leading to their popularity for gamification [36]. Intrinsic motivation can be targeted through game design that allows for play, exposition, choice, information, engagement, and reflection [14].

Cognitive tasks are typically gamified through reward-based gamification, which is usually appropriate given their context of use. Assessments using cognitive tasks are often given at a single point in time; for example, following a brain injury to assess impairment. Some assessments may be repeated over time, but still in a narrow capacity. As well, there is limited value to enhancing intrinsic motivation for these tasks. The value of completing cognitive tasks is already extrinsic in nature; for example, to access medical care.

However, just as gamification may change the psychometric properties associated with cognitive tasks, game elements may also change the user experience in unintended ways. In a literature
review, Seaborn and Fels [37] found that while intrinsic and extrinsic motivation are frequently used as theoretical bases for gamification, they are rarely empirically studied in that context. Further, when studies are conducted, they often combine multiple game elements, which makes it difficult to parse out the effects of individual elements on user motivation [17]. Different users may also respond to gamification differently; for example, older adults, children, and people with more and less gaming experience may all have different expectations around games.

Overall, the problem is that while researchers continue to look towards games to address the challenges of cognitive assessment, there are many questions that need to be answered before these gamified tasks can be widely implemented as scientific tools.

1.2 Research Questions

My research focuses on tasks used for the assessment of attention, though cognitive tasks are used in various contexts and can target different cognitive domains.

I focus on assessment as there are unique considerations when gamifying cognitive tasks for assessment, as opposed to training and intervention. Much of the research on the use of gamification for cognitive tasks has concentrated on training and intervention [38]. In these contexts, users repeatedly practice these tasks to improve their cognitive abilities or modify their behaviours. Gamifying tasks for training and intervention is different from gamifying tasks for assessment. With assessment, the goal is to capture an individual’s ‘true’ ability, often within a single cognitive domain. Training tasks are not held to the same psychometric standards and are free to target multiple cognitive domains, which allows for easier gamification. More research is required to ensure that gamified assessment tasks accurately estimate specific cognitive functions.

I also specifically focus on the assessment of attention. Cognitive tasks have been developed to assess many cognitive domains, such as memory, executive functions, decision making, processing speed, perception, and attention. Attention is an important function to assess, as an accurate measure of attention can help diagnose many common disorders, such as attention-deficit/hyperactivity disorder (ADHD) and dementia; however, using games to assess attention
may be associated with unique problems. Games inherently manipulate attention through elements like sound effects, graphics, and rewards. One study found that using gamification normalized the performance of participants with ADHD [39].

My research seeks to understand how using games for the assessment of attention affects data quality. I focus on the following questions, centered around improving the quality of data captured by cognitive tasks:

1. How are digital games used for the cognitive assessment of attention made and measured? (Manuscript A)
2. How do game elements affect performance on, and enjoyment of, an attention-based cognitive task? (Manuscript B)
3. How do game elements affect the reliability of an attention-based cognitive task? (Manuscript C)

Overall, I seek to answer the question: How do games for the assessment of attention affect the quality of gathered data, in terms of both the experience of participants and their performance?

1.3 Overview of this Dissertation

For Manuscript A, I conducted a systematic review to answer the question: How are digital games used for the cognitive assessment of attention made and measured? This review provided an overview of the field and a guide for further research questions. Ultimately, we found that many assessment games are not sufficiently evaluated to determine the quality of data they produce, which is a significant barrier to progress in the field. Thus, for the remainder of my dissertation research, I focused on evaluating and improving the quality of data captured by cognitive tasks.

In Manuscript B, I examined how game elements affect performance on, and enjoyment of, an attention-based cognitive task. This study provided guidance on which game elements are most likely to improve data quality, in terms of increasing participant enjoyment (and therefore engagement with the task) and performance (how quickly and accurately they respond to the task).
In Manuscript C, I examined how the game elements identified in Manuscript B affect the reliability of an attention-based cognitive task. High reliability is fundamental to data quality, but any changes to a task (like adding game elements), can affect participant performance and therefore the reliability of the data.

1.3.1 Contribution to Each Manuscript

Each of the manuscripts presented in this dissertation are products of multiple authors; as such, I use ‘we’ throughout each manuscript to discuss the work, and I will clarify my personal contributions here.

I am the first author on all three manuscripts. I lead the research under the supervision of Dr. Regan Mandryk. With her feedback, I developed the research questions and study designs. I was primarily responsible for the data analysis and writing of each manuscript, again with the helpful guidance of Regan Mandryk. Raquel Robinson provided a second evaluation of the records identified for the systematic analysis in Manuscript A. Sarah Vedress helped build the game in Manuscript B and Phaedra Berger helped build the game in Manuscript C. Maxmilian Friehs assisted with the analysis in Manuscript C. All co-authors provided general feedback and edited the final manuscripts.
2 Manuscript A: The Making and Evaluation of Digital Games Used for the Assessment of Attention: Systematic Review\textsuperscript{1}

2.1 Introduction to Manuscript A

When exploring the literature on the gamification of cognitive tasks, we found review papers on gamified tasks for training and intervention, but less systematic exploration of the literature on gamified assessment tasks.

Cognitive tasks for assessment are heavily researched and standardized. Gamified tasks need to be subject to the same rigour before they can be widely used. Well-established cognitive tasks are implemented in ways that are replicable across studies. For example, there are standardized stimuli for tasks, fixed times for trials, and careful control of the experimental setting. Currently, the research on assessment games is diverse and lacks uniform tools and methods.

We conducted a systematic review to map the diverse ways assessment games are developed and evaluated. We focused on the assessment of attention both to scope our review, and because game elements inherently manipulate attention and thus properly evaluating the assessment capabilities of such games is important.

We aimed to answer the question: How are digital games used for the cognitive assessment of attention made and measured? In contrast to previous systematic reviews, we looked beyond gamification to other approaches for developing assessment games. We also focused on the process and methodology of developing and evaluating these games, rather than just the end product. Our findings allowed us to summarize existing methods for developing and evaluating assessment games, and to identify gaps for future research.

\textsuperscript{1} The manuscript in this chapter was published as: Katelyn Wiley, Raquel Robinson, and Regan Lee Mandryk. The Making and Evaluation of Digital Games Used for the Assessment of Attention: Systematic Review. JMIR Serious Games 2021;9(3):e26449
2.2 Abstract

**Background:** Serious games are now widely used in many contexts, including psychological research and clinical use. One area of growing interest is that of cognitive assessment, which seeks to measure different cognitive functions such as memory, attention, and perception. Measuring these functions at both the population and individual levels can inform research and indicate health issues. Attention is an important function to assess, as an accurate measure of attention can help diagnose many common disorders, such as attention-deficit/hyperactivity disorder and dementia. However, using games to assess attention poses unique problems, as games inherently manipulate attention through elements such as sound effects, graphics, and rewards, and research on adding game elements to assessments (i.e., gamification) has shown mixed results. The process for developing cognitive tasks is robust, with high psychometric standards that must be met before these tasks are used for assessment. Although games offer more diverse approaches for assessment, there is no standard for how they should be developed or evaluated.

**Objective:** To better understand the field and provide guidance to interdisciplinary researchers, we aim to answer the question: How are digital games used for the cognitive assessment of attention made and measured?

**Methods:** We searched several databases for papers that described a digital game used to assess attention that could be deployed remotely without specialized hardware. We used Rayyan, a systematic review software, to screen the records before conducting a systematic review.

**Results:** The initial database search returned 49,365 papers. Our screening process resulted in a total of 74 papers that used a digital game to measure cognitive functions related to attention. Across the studies in our review, we found three approaches to making assessment games: gamifying cognitive tasks, creating custom games based on theories of cognition, and exploring potential assessment properties of commercial games. With regard to measuring the assessment properties of these games (e.g., how accurately they assess attention), we found three approaches: comparison to a traditional cognitive task, comparison to a clinical diagnosis, and
comparison to knowledge of cognition; however, most studies in our review did not evaluate the game’s properties (e.g., if participants enjoyed the game).

**Conclusions:** Our review provides an overview of how games used for the assessment of attention are developed and evaluated. We further identified three barriers to advancing the field: reliance on assumptions, lack of evaluation, and lack of integration and standardization. We then recommend the best practices to address these barriers. Our review can act as a resource to help guide the field toward more standardized approaches and rigorous evaluation required for the widespread adoption of assessment games.

2.3 Introduction

2.3.1 Attention

From crossing the street to composing a tweet, functioning as a human has always required people to take in information, process it, and respond accordingly. Whether in the lab or in the world, detecting stimuli and responding to them, both consciously and unconsciously, involves many cognitive functions. One of these important cognitive functions is **attention**, which Kahneman describes as “a label for some of the internal mechanisms that determine the significance of stimuli” [40].

These ‘internal mechanisms’ of attention can be broken down into multiple types. Common areas of attention include selective attention (how people attend to relevant information and ignore irrelevant information), divided attention (when people attend to multiple things at once), and sustained attention (the ability to focus on something for a continuous amount of time) [41,42].

There are also models of attentional control, which describe the difference between involuntary and voluntary attention. Attentional control is related to inhibition, shifting, and updating. Inhibition involves preventing irrelevant stimuli from impairing performance, shifting refers to the allocation of attention to whatever is most relevant at the time, and updating is how people take new information into working memory [43].
2.3.2 Assessment of Attention

Measuring and understanding attention and attentional control are important as attention is a major cognitive function that influences human development and mental health. Further, as attention is related to a variety of cognitive deficits (e.g., ADHD [44], dementia [45]), and abilities (e.g., reading [46]), an accurate measure of attention can help assess and diagnose a number of common disorders.

Measuring attention, and other aspects of cognition like memory and perception, is often done using cognitive tasks. A common approach is to present participants with stimuli and ask them to respond in different ways, while measuring their reaction time and accuracy (i.e., how quickly they attend to stimuli and if they respond in the way intended). Research on attention has often relied on specific cognitive tasks, such as the Eriksen flanker task [47] and Posner cueing task [48], which have been fundamental to the study of attention. More recent cognitive tasks also continue to advance knowledge of attention. For example, the dot probe task demonstrates that people with anxiety preferentially attend to threatening stimuli [49], and attentional blink tasks support the idea that attentional resources are limited [50].

Cognitive assessment tasks have specific standards they must meet before they are widely used, especially in clinical settings. They are expected to have certain psychometric properties, such as validity (how well they measure what they claim to measure), reliability (how consistent the test is), sensitivity (how well they identify true positives), and specificity (how well they identify true negatives).

2.3.3 Digital Games for the Assessment of Attention

While cognitive assessment tasks are standardized and highly used, they do have some limitations. They can be expensive, as many require trained experts to administer the tasks [51]. They are boring and repetitive, which can cause difficulty with recruitment for research and with patient cooperation for clinical use [19]. The data collected by these tasks can be unreliable, as participants might not be fully engaged and often exert suboptimal effort [20,21]. They also lack ecological validity [51,52], so may not be indicative of how these cognitive skills affect daily
functioning. In order to address these limitations, researchers have started to integrate elements from computer and video games into cognitive tasks for assessment.

Games have the potential to improve the quality and quantity of collected data, by increasing participants’ engagement in the moment (better data) and by engaging many more people over longer periods of time (more data) [53]. For example, using a game called ‘Sea Hero Quest’ [25], researchers were able to collect spatial navigation data from over 4.3 million people, something that would be near impossible with a traditional paper-based task or even a standard digitized assessment. However, while assessment games can be very successful, they do not always improve participant enjoyment. Vanden Abeele et al. [28] note the important of game quality when developing assessment games. In fact, studies have shown that some game elements are associated with lowered enjoyment compared to traditional tasks [19,53].

Game elements can also hinder the assessment properties of a task. Cognition is complicated, and traditional tasks are heavily studied before researchers can be confident that they measure what they claim to measure in a consistent way (issues of validity and reliability). Even a small change to a task must be studied in order to understand its effects [32].

Using games to measure cognitive processes related to attention poses unique issues. Through their use of graphics, stimuli, and visual feedback, games inherently manipulate the player’s attention, which has been shown to be problematic when using games for assessment. For example, Wiley, Vedress, and Mandryk [53] found that participants responded more quickly, but also less accurately to a dot probe task when points were awarded for faster, correct responses. Other features could also manipulate attention; for example, increasing narrative suspense has been linked to narrowed attentional focus [33]. In Go/No-Go games, using ‘game-like stimuli’ such as cartoon characters has resulted in decreased performance compared to standard tasks, possibly because it is more difficult to differentiate between complicated graphical stimuli than simple coloured shapes [19].

Attention can also interact with games based on individual differences. A study by Delisle and Braun [39] found that game elements can normalize the performance of individuals ADHD.
They designed a task to resemble a fast-paced video game and found that the presence of game elements improved the performance of participants with ADHD more than it improved the performance of non-ADHD participants. This unequal effect on performance implies that a game designed to assess ADHD could, ironically, be rendered unable to effectively discriminate. Other individual differences may also affect the data, such as differences in age, gender, and game playing experience. Studies have shown that action video game players demonstrate visual search advantages [54]; such a difference may be emphasized by using a game to measure attention. Similarly, older adults or people with little gaming experience may perform poorly on a game not because they have lower attention abilities, but because they are less familiar with computers and games.

2.3.4 The Present Research

There has been a large research interest in the use of gamification (i.e., the use of game elements in non-game contexts [55]) on cognitive tasks, which has been synthesized in several review papers [38,56]. However, the focus of much of this synthesis research has been on the gamification of training and intervention (cf, [38]), with less systematic exploration of the efficacy of games for assessment. Although games for training and assessment are often grouped together, recent research suggests that gamification may not be the best approach for assessment [53]. While cognitive tasks are standardized and have been heavily researched, serious games for assessment are diverse and there is no field-wide standard for how they should be developed. Our systematic review seeks to explore the different approaches to using games for assessment, particularly for assessing attention as it can be complex. In comparison to other systematic reviews, we provide two unique contributions: First, we look beyond gamification to other approaches for developing assessment games. Second, we review the methodology of developing and evaluating serious games for assessment, rather than just the end product. We aim to provide a guide for interdisciplinary researchers on how assessment games can be developed and evaluated. Our main research question for this systematic review is: How are digital games used for the cognitive assessment of attention made and measured?
2.4 Methods

2.4.1 Eligibility Criteria

Our eligibility criteria required each included paper to be published before March 1, 2021, a peer-reviewed journal article or conference proceeding, primary research (i.e., not a literature review or background article), and in English.

Additionally, each paper needed to include a digital game used for the assessment of attention-related processes, that could be used remotely. For these criteria, we used the following definitions:

- **Digital game**: Because there are many ways to define a ‘game’, we chose to follow the original researchers’ intentions. If the authors of a paper referred to an assessment as a ‘game’, ‘gamified’, or some other variation, we included the paper.
- **Attention**: We included papers related to attention and attentional control. For a detailed list of the cognitive processes included, see Table 1.
- **Assessment**: We were interested in studies that seek to measure attention, for purposes of either detection/diagnosis, research, or monitoring cognitive changes. Studies focused on treatment, training, or interventions were excluded, as well as studies on educational and work assessments (e.g., assessing for employee selection or how well a concept was learned).
- **Remote**: As the goal of digitizing assessment is often to increase its scale, accessibility, and reach [57], we included only papers where the game could potentially be deployed remotely, using only a computer, tablet, or phone. Studies that required specialized hardware (e.g., a Kinect, gaming controllers, any custom hardware) were excluded, though studies with commonly used devices (e.g., a mouse, headphones, keyboard) were included. Though the ability for studies to be deployed remotely depends on more than available hardware, for this review we did not exclude studies without the requisite software.
Table 1. Included and excluded cognitive functions for the eligibility criteria in our review.

<table>
<thead>
<tr>
<th>Included Cognitive Functions</th>
<th>Excluded Cognitive Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Attention and Attentional Control</strong></td>
<td><strong>Excluded Cognitive Functions</strong></td>
</tr>
<tr>
<td>• Selective attention</td>
<td>• Working memory</td>
</tr>
<tr>
<td>• Divided attention</td>
<td>• Long-term memory</td>
</tr>
<tr>
<td>• Visual attention</td>
<td>• Autobiographical memory</td>
</tr>
<tr>
<td>• Sustained attention</td>
<td>• Visuo-spatial memory (including spatial navigation)</td>
</tr>
<tr>
<td>• Inhibition</td>
<td>• Prospective memory</td>
</tr>
<tr>
<td>• Shifting</td>
<td>• General knowledge</td>
</tr>
<tr>
<td>• Updating</td>
<td>• Learning</td>
</tr>
<tr>
<td>• Orienting</td>
<td>• Problem solving</td>
</tr>
<tr>
<td></td>
<td>• Deductive reasoning</td>
</tr>
<tr>
<td></td>
<td>• Decision making</td>
</tr>
<tr>
<td></td>
<td>• Intelligence</td>
</tr>
<tr>
<td></td>
<td>• Creativity</td>
</tr>
<tr>
<td></td>
<td>• Planning</td>
</tr>
<tr>
<td></td>
<td>• Etc.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Perception and Processing Speed</th>
<th>Memory</th>
<th>Executive Functions and Decision Making</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Visual</td>
<td>• Working memory</td>
<td>• General knowledge</td>
</tr>
<tr>
<td>• Auditory</td>
<td>• Long-term memory</td>
<td>• Learning</td>
</tr>
<tr>
<td>• Temporal</td>
<td>• Autobiographical memory</td>
<td>• Problem solving</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Deductive reasoning</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Decision making</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Intelligence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Creativity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Planning</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Etc.</td>
</tr>
</tbody>
</table>

2.4.2 Information Sources and Search Strategy

We searched the titles and abstracts of papers in several databases, chosen for their relevance to games-user research and psychology: ACM Digital Library, IEEE Xplore, PubMed, PsychInfo, Scopus, and Web of Science. Our keywords were used in a search string adapted for the requirements of each database, but generally followed the same format. Additionally, where possible we added requirements that the returned records be articles published in English. We set no lower time limit but did require that included papers be published before March 1, 2021 (when the search was conducted). In order to facilitate other potential systematic reviews, we cast a wide net with our search terms, and included other cognitive processes. As an example, for Scopus, the search string was: TITLE-ABS ((gamif* OR game OR games) AND (cognit* OR
neuropsych* OR assessment OR memory OR executive function OR attention* OR impulse control OR processing speed OR inhibition OR anxiety OR depression)) AND DOCTYPE (ar).

2.4.3 Study Records and Data Management

To screen the final set of records for our inclusion criteria, we used Rayyan, an online system developed for conducting systematic reviews [58].

The first author screened the titles and abstracts of all records for obvious exclusions, such as papers that referenced physical games and sports, as opposed to digital games. The first two authors both screened the remaining papers and made final decisions, resolving conflicts through discussion.

As a final quality check, we also screened the first 100 results from Google Scholar using the search string: ((gamif* OR game OR games) AND (attention*)).

2.4.4 Data Items and Synthesis

Our main research question for this systematic review is: How are digital games used for the cognitive assessment of attention made and measured? We also had follow-up questions related to the evaluation of the game’s efficacy and engagement: How effective are the games at accurate assessment?; and, how effective are they in terms of participant engagement?

To conduct the review, we gathered data related to a list of questions for each paper, using a spreadsheet. We listed what specific area of attentional control the study looked at (e.g., selective, divided), the population the study examined (e.g., children, older adults), and the sample size of the study.

We also listed a number of details about each assessment, including its intended purpose, a general description of the assessment, and if it was focused on a specific disorder (e.g., ADHD, dyslexia, dementia). We were also interested in how each assessment was measured; particularly, how it was evaluated, if it was compared to any traditional cognitive tasks or a clinical diagnosis, and any results from the evaluation.
We listed details about gameplay, giving a general description of each game and listing any game mechanics used (e.g., points, narrative, avatars). We also noted how the game was developed (e.g., gamification of a task, custom game, existing commercial game), the expertise of individuals involved in its development (e.g., healthcare professionals, game designers), any evaluation of the game and the results (e.g., enjoyment, immersion), and the authors’ motivation for using a game.

The first two authors collected data from each paper in a spreadsheet, with each author responsible for half of the papers. The first author then reviewed each paper and the spreadsheet to ensure uniform data collection. This data then informed the qualitative synthesis presented in our results. The wide range of methods used in the included papers precluded any meta-analysis or meaningful quantitative analysis. At most, we provide summary statistics. We intend for this review to be an overview of how research is done in the field and focus on the methodology of each paper.

2.5 Results

2.5.1 Search Results

Our initial search was conducted in December 2019, with an updated search conducted in March 2021 to include any new publications. The two searches resulted in a set of 91,968 records. We used Mendeley Reference Management software, which automatically deleted 38,179 duplicate records. We then manually deleted a remaining 4,424 duplicates, resulting in a final set of 49,365 records to review.

The first author’s initial screening excluded 46,969 records. These exclusions were made quickly based on brief searches through the titles and abstracts. For example, many records referenced ‘the Olympic Games’ or ‘game’ as in animal game.

The first two authors then reviewed the remaining 2,396 papers in more detail and excluded a further 2,326 papers. At this stage, common exclusions included papers that addressed cognitive
processes other than attention (e.g., [59]) papers that used virtual reality or other specialized hardware (e.g., [60]), and papers that focused on interventions and training (e.g., [61]).

A total of 78 papers were selected for analysis; however, we were unable to obtain the full-text of four papers from any digital library, interlibrary loan, or attempting contact with the authors, leaving a final set of 74 papers for the review (see Figure 1).

![Flowchart of included and excluded records through our review process.](image)

2.5.2 Summary of Included Studies

Full details for papers included in this review can be found in Appendix A (general details of each study) and B (assessment and game details of each study).

2.5.2.1 Publishing Formats and Dates

The 74 papers in our review were published in journals from a variety of fields (based on the journal descriptions): 33 papers were from psychology and medical publications, 26 papers were from interdisciplinary publications, 13 papers were from computer science publications, and 2 were from education publications (see Table 2).
The earliest study in our review is from 2000. The majority of papers were published in the late 2010s, with 14 papers from 2018 and 13 from 2019 (see Figure 2). Two papers were published in the first two months of 2021, before the upper time limit for inclusion in this review.

Table 2. References to all included papers by publication field.

<table>
<thead>
<tr>
<th>Field</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interdisciplinary</td>
<td>[19,53,70–79,62,80–85,63–69]</td>
</tr>
<tr>
<td>Computer Science</td>
<td>[119,120,129–131,121–128]</td>
</tr>
<tr>
<td>Education</td>
<td>[132,133]</td>
</tr>
</tbody>
</table>

Figure 2. Number of included papers by publication year and field.
2.5.2.2 Study and Participant Characteristics

The selected papers study a total of 40,154 participants, with sample sizes ranging from five to 16,233 participants.

The majority of studies focused on children (35 papers), with 12 papers focused on older adults and 25 on a general adult population. Two papers had participants from across the human lifespan, with both children and adults of all ages.

While the majority of papers looked at a general population (34 papers), there were three main disorders also studied: 16 papers on dementia and general cognitive impairment, 11 on ADHD, and seven on dyslexia. Six papers examined other disorders (schizophrenia, substance abuse, aggression, multiple sclerosis, down syndrome and zika virus, and Parkinson’s and Huntington’s).

The papers studied many different aspects of attention and attentional control: inhibition (25 papers), sustained attention (23 papers), visual attention (19 papers), selective attention (14 papers), switching (9 papers), updating (7 papers), divided attention (6 papers), orienting (2 papers), and attentional bias (1 paper). Twenty-three papers measured multiple types of attention. Five papers did not specify what type of attention was measured, nor could it be inferred from information in the papers.

2.5.3 Why are digital games used for assessment?

The papers in our review listed several reasons for using a game to assess cognition: to address limitations of traditional tests (32 papers), to increase participant motivation (22 papers), to engage children (18 papers), and because of previous research (7 papers). Seven papers listed multiple reasons. Seven papers did not list a motivation for using a game.

The most common reason given for using a game was to improve motivation and engagement (40 papers total; 22 with a general population, 18 specifically geared towards children). For example, Thirkettle et al. [76] created an app that gamified a battery of cognitive tests, specifically with the goal of encouraging repeated play. Dibbets et al. [92] sought to study task
switching in children but realized that traditional tasks require participants to be literate. They
developed the Switch Task for Children, which does not require a reading response and
presented it as a game to “appeal to young children”.

Thirty-two papers used games to address the limitations of traditional assessments, as they can
be costly in time and resources, require special expertise, lack ecological validity, and cannot be
widely deployed. For example, Brown et al. [63] were able to collect data from over 16,000
users, using a smartphone app that gamified several tasks. Tong et al. [107] created a ‘whack-a-
mole’ game for delirium screening in emergency departments, noting that it would be
particularly useful to have an automated cognitive test given the busy and demanding nature of
emergency rooms.

2.5.4 How are digital games used for assessment made?

2.5.4.1 Development process

Our review found three different approaches to developing games for assessment purposes:
gamifying cognitive tasks (33 papers), creating custom games based on theories of cognition (37
papers) and exploring potential assessment properties of commercial games (4 papers). One
paper uses both a gamified task and a commercial game.

2.5.4.1.1 Gamification

Gamifying a traditional cognitive task is a common approach to making a game for assessment.
This approach involves adding points, graphics, and other game elements to a traditional task.
Typically, these tasks are digitized (if they are not already) and then game elements are layered
over top of the basic task. As an example of gamification from our review, Johann and Karbach
[97] created a series of gamified tasks for children, all based around the story of a magic
kingdom where an evil wizard must be defeated. One task was a Go/No-Go task where the
stimuli were dragons of different shapes and colours. Correct responses advanced a progress bar
and earned participants ‘magic power points’. This example uses points, graphics, and theme to
gamify the task, but any one game element or combination of elements can be used for
gamification. For example, Lumsden et al. [19] created a Stop Signal gamified task using only
points, and another variation using only thematic graphics. On the other end of the spectrum,
Ryokai et al. [74] used many game elements in their multiple object tracking (MOT) game for children, including a theme, graphics, music and other sound effects, dynamically adjusted levels, and feedback.

2.5.4.1.2 Custom Games

Another popular approach is to create a custom game, based on theories of cognition or other previous research. For example, several papers in our review create games to detect dyslexia based on theories of visual-spatial attention [69,119,123,127]. As another example, McKanna, Jimison, and Pavel [70] created ‘21 Tally’, described as ‘blackjack played in two dimensions’, based on theories of divided attention.

The custom games in our review are diverse, ranging in appearance and complexity. Some are fairly similar to gamified tasks in their simplicity and approach. For example, while Chesham et al. [64] did not gamify a specific task, their custom game looks and feels like a gamified task, with very simple designs and mechanics; in fact, they call it a ‘taskified game’. Other custom games, like Anguera et al.’s [86] ‘EVO’ game, for example, are more akin to a commercial game. ‘EVO’ was designed specifically to resemble a commercial action video game, with a focus on “high-level art, music, feedback and storylines”.

These custom games are also diverse in how they are developed. Some games are very literature- and hypothesis-driven, while others offer less rationale for their design. For example, Rauschenberger et al. [127] created ‘MusVis’, a game to detect dyslexia using language-independent methods. They based their design choices on the visual and auditory processing abilities of individuals with dyslexia. On the other hand, some papers in our review did not offer clear rationale for how they designed their custom game and why they expected it to work.

2.5.4.1.3 Commercial Games

Four papers used pre-existing commercial games. In these cases, the games are not created by the researchers; rather, the researchers explore the potential assessment properties of a game. For example, Intarasirisawat et al. [130] used Tetris, Fruit Ninja, and Candy Crush Saga to investigate how touch gestures in popular games might relate to performance on traditional cognitive tasks. In their study, participants were asked to play the games and also complete
traditional paper-based tasks. A bivariate analysis then revealed correlations between the commercial games and the tasks.

As another example, Houghton et al. [66] used Crash Bandicoot to study motor control and sequencing of boys with and without ADHD, under low and high working memory and distractor conditions. The goal of the study was to compare how boys with ADHD performed in an ecologically valid, highly motivating environment (a computer game) compared to a standard laboratory environment.

2.5.5 How are digital games used for assessment measured?

2.5.5.1 Evaluation of the assessment aspect

Because digital games have not traditionally been used for cognitive assessments, they need to be evaluated for their relation to cognition; for example, researchers may want to know if people with and without ADHD display different mouse behaviours [132] or if scores on a gamified Go/No-Go task correlate to correct responses on a traditional Go/No-Go task [65,97,100,106].

Our review found three approaches to evaluating the assessment aspect of games: comparison to a traditional task (31 papers), comparison to a clinical diagnosis (14 papers), comparison to knowledge of cognition (11 papers). An additional four papers compare game results to both a traditional task and a clinical diagnosis, and one paper uses both clinical diagnoses and comparisons to normative data. Of these papers, we further identified four papers that use machine learning for evaluating assessment. Thirteen papers did not evaluate the assessment aspect of the game.

2.5.5.1.1 Comparison to a traditional task

The most common approach to evaluating the assessment aspect of a game is to compare the results of a game to the results of an established cognitive task. For example, if the scores from a game are designed to measure response inhibition, researchers may want to compare those scores to scores from a Go/No-Go task. If the patterns of responses are similar, it is likely that the game is measuring response inhibition in a similar way to a Go/No-Go task. This process is what Chicchi Giglioli et al. [65] followed when evaluating their game, ‘EXPANSE’, which gamified
the Dot Probe Task, Go/No-Go Task, Stroop Task, Trail Making Task, and Wisconsin Card Sorting Test. Participants were asked to complete both the standard tasks, and game-based versions.

While this process offers the most direct comparison between tasks and their gamified components, of the 33 papers that gamified a traditional task, only 14 then evaluated the game by comparing it to the task.

Custom and commercial games may also be compared to traditional tasks. For example, Tong et al. [106] created a ‘whack-a-mole’ game designed to measure response inhibition. While this game was not based on a specific task, they did calculate correlations between performance on the game and performance on a standard Stroop task, which also measures response inhibition. Similarly, Baniqued et al. [87] had participants play 20 commercial casual games as well as a battery of cognitive tasks, and then looked at the relationships between how participants performed on the games and the tasks.

2.5.5.1.2 Comparison to a clinical diagnosis

Another approach to evaluation is comparing the results of a game to a clinical diagnosis or questionnaire. For example, if a game is designed to measure selective attention, researchers may want to look at how scores on the game differ between children with and without ADHD. Because children with and without ADHD typically display different patterns of selective attention, the game can be assessed to see if it discriminates between children with and without ADHD. If game performance is different between the two groups, it may be measuring selective attention. Alternatively, it may be picking up on some other feature of cognition that differs between children with and without ADHD, so this approach needs to be used carefully.

If the intent of the game is to diagnose ADHD, it matters less why it works than if it works. Most of the games in our review used this evaluation approach for situations in which diagnosis was the goal. For example, Peijnenborgh et al. [72] demonstrated that their game, ‘Timo’s Adventure’, had clinical validity by showing that children with and without ADHD had significant performance differences.
As another example, Fukui et al. [95] recruited participants with mild cognitive impairment, participants with Alzheimer’s disease, and age- and gender- matched healthy control participants, and then examined if performance data from their games was able to discriminate between the three groups.

2.5.5.1.3 Comparison to theories of cognition

The results of a game can also be compared to theories of cognition. For example, if a game is designed to look at attentional blink phenomena, we can look at literature on patterns of attentional blink to see if the results from the game make sense. These comparisons can happen in different ways—for example, comparisons to normative data or to a specific theory. Brown et al. [63] use a set of gamified tasks. They do not compare those games to a task version directly, but they do compare their results to the literature on those tasks, finding that their games produced ‘canonical results’. Thirkettle et al. [76] compared their results based on demographic effects, through grouping participant results by gender and age, and looking for a replication of known effects (for example, increases in age correlating with increases in reaction time). Similarly, Ryokai et al. [74] compare their results to known MOT limits.

2.5.5.1.4 Machine learning

As an additional finding, we also found four papers that use machine learning in their evaluation. These papers compare the results of a game to the results of a cognitive task or clinical diagnosis, and then use machine learning to build a classification model for the game results. For example, Jung et al. [68] used machine learning to classify game scores by comparing them with Mini Mental State Examination (MMSE) scores, and Mwamba et al. [71] classified game data based on children with and without a diagnosis of ADHD.

2.5.5.2 Evaluation of the game aspect

When using digital games for assessment, it is important to also evaluate the games themselves. As discussed above, many papers in our review discussed using games as assessment tools with the motivation to increase participant engagement with testing. To that end, researchers may want to know if a game is more enjoyable than a traditional task (e.g., [100,120]) or if participants find the game too difficult to play (e.g., [128]).
In our review, only 25 studies evaluate some aspect of game play. The majority of studies do not evaluate game features at all (41 papers). Eight papers do not formally report an evaluation of the game, but suggest that some evaluation was done (e.g., a sentence in the discussion section that indicates most participants enjoyed the game [62] and mentioning that participants were asked if they enjoyed the game without reporting the results [87]).

Of the 25 studies that evaluated game play, most used a short questionnaire to assess enjoyment and difficulty, though these measures vary in complexity. For example, Gaggi et al. [119] asked children two simple questions about their game experience: “Do you like the game?” and “Is the game difficult to play?” with simple choices for answers: “Yes, a lot; Yes; Not so much; No” and “Easy; Medium; Hard” respectively. On the other end of the spectrum, Szalma et al. [75] used a more detailed battery of the NASA Task Load Index (NASA TLX) (a questionnaire that measures perceived workload) and other questionnaires to measure stress, task engagement, and difficulty.

Game behavior was also used as a metric of enjoyment; for example, Thirkettle et al. [76] noted that 1400 participants played their game more than ten times, and Godwin et al. [120] found that their ‘Monster Mischief’ game was three times as popular with children as the MOT task it was based on.

Another way to assess game elements is to directly compare multiple versions of a game/task, with different elements included in each version. Miranda and Palmer [102] compared three versions of a visual search game: with points and sound, with only points, and with only sound. Lumsden et al. [100] also took this approach, comparing three versions of a gamified task (a non-game task, a gamified task with points, and a gamified task with a theme).

2.5.6 How effective are assessment games?

While our review was mostly concerned with methodology; that is, how assessment games are made and evaluated, we also wanted to explore how effective these games are. How effective are they at accurate assessment, and how effective are they at participant engagement?
Unfortunately, the general lack of evaluation done by the papers in our review precludes us from answering these questions. In addition, the wide range of methodologies makes it difficult to compare even the limited studies that do include some evaluation. For example, Miranda and Palmer [102] gamified a visual search task and used it to investigate how participants respond to a task with points and sound effects included. Similarly, Lumsden et al. [100] took the same approach with a Go/No-Go task, to look at the effects of points and theme. However, the evidence from these studies is difficult to compare, despite their similar methodologies. Miranda and Palmer compare versions of their game with only points, with only sound effects, and with both points and sound effects. They do not compare it to a control version of the task. Lumsden et al. compare versions of their game with only points, with only theme, and with a control version of the task, but do not have a version that combines points and theme. These differences in evaluation are found across all the studies in our review.

2.6 Discussion

2.6.1 Principal Results

We identified and reviewed 74 papers that use a digital game to measure cognitive functions relating to attention. We sought to answer the question: How are digital games used for the cognitive assessment of attention made and measured?

We found three different approaches to making assessment games: gamifying cognitive tasks, creating custom games based on theories of cognition, and exploring potential assessment properties of commercial games.

Games for assessment have two aspects that can be evaluated: the assessment properties (e.g., how accurately attention is measured), and the game properties (e.g., how fun the game is). The papers in our review that evaluated the assessment properties used three approaches: comparison to a traditional cognitive task, comparison to a clinical diagnosis, and comparison to knowledge of cognition; however, the majority of studies did not evaluate the game properties.
From our review, we see three barriers to progress in using games for cognitive assessment. We propose recommendations to address these barriers and offer ideas for further research.

2.6.1.1 Barriers to Progress in the Field

There are three barriers to making substantial progress in using games for cognitive assessment. The first barrier we identified is that the literature currently perpetuates assumptions about how users interact with assessment games. Second, there is a lack of evaluation of these games. Third, there is not a clear standard of integration across the field.

2.6.1.2 Assumptions

While the papers in our review did not explicitly state that assumptions about games informed their choices, our results did reveal some patterns. Thirteen papers did not evaluate the assessment aspect of the game. Of these papers, the vast majority were about gamified tasks (9 papers). In addition, of the 33 papers that gamified a traditional task, only 14 then evaluated the game by comparing it to the task. This lack of evaluation for gamified tasks may be due to the assumption that adding simple elements like points and graphics will not interfere with performance on the basic task.

Forty papers chose to use a game for assessment due to a potential increase in participant engagement and enjoyment; however, this choice was rarely followed up with an evaluation of how enjoyable participants found the game. In fact, out of these papers, only eighteen evaluated the game play. This assumption that a game of any type or quality will be engaging and yield better results than a traditional task is pervasive across the literature, despite evidence to the contrary. For example, Wiley, Vedress, and Mandryk [53] looked at the effects of including points and theme in a gamified task. They found that while points increased participants’ experiences of enjoyment, challenge and meaning, adding a theme actually lowered these experiences. After a theme-based introduction to task, enjoyment was temporarily higher, but dropped after play, likely because the basic game play failed to live up to participants’ heightened expectations. Different game experiences will influence enjoyment and engagement in ways that cannot always be predicted.
2.6.1.3 **Evaluation**

More than half of the papers in our review did not evaluate any aspect of game play (41 out of 74 papers), and 13 papers did not evaluate any assessment properties. This lack of evaluation is problematic for the advancement of cognitive assessment games. In order for the games in our review to be seriously considered as assessment tools in the way that standard cognitive tasks are, they need to be evaluated and validated in the same way.

Only 21 papers from our review evaluated both assessment and game properties. As Levy et al. [30] note, “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”. Each assessment game needs to be evaluated based on both its assessment value and its game value.

2.6.1.4 **Integration and Standardization**

A final barrier to progress is integration across the field. Currently, there is little guidance on how assessment games should be made, evaluated, and used. Every project will be different (e.g., will use different tasks, themes, or gamification approaches), but for the field to advance there needs to be integration at the level of game structure. In referring to the structural level, we mean that researchers should develop a clear understanding of how different game mechanics in assessment games (e.g., points, theme, rewards, feedback, procedures, rules, game input, narrative elements) interact with user performance and experience. Some research is being done in this regard; for example, multiple studies have shown a classic speed-accuracy trade-off when points are included in an assessment game [53,77]. This type of work should be expanded to other game mechanics, and reviews and meta-analyses should integrate the findings to develop standards within the field. The majority of the included studies were published in psychology or medical venues, compared to computer science venues, or interdisciplinary venues, so it may be that research teams lack formal training in the design and deconstruction of games (e.g., [134]).
2.6.1.5 Recommendations to Address Barriers

In order to help advance the field, we identified some ‘best practices’ from the papers in our review. These include using clearly defined goals to guide the development of a game, ensuring robust evaluation, and working with interdisciplinary teams.

2.6.1.5.1 Motivation and Purpose

**Describing a clearly-stated motivation for using a game over a traditional assessment can justify the use of games in this context.** For example, the motivation may be to increase engagement for children or reduce drop-out rates for long-term monitoring. The game can then be evaluated to see if it meets the motivation (e.g., does it engage children? Does it reduce drop-out rates?). This evaluation is important, as the field is new enough that results are often not generalizable.

**Clearly articulating a clear long-term goal can set guideposts for the development and evaluation of games for assessment.** For example, if the goal is to use a game as a neuropsychology tool, then researchers can focus their efforts on making a standardized game that meets robust standards for validity and reliability, and that has a large normative data set [51]. If the goal is to create a game for wide-spread dissemination for population-level research, then the focus can be on making the game truly engaging and fun, to encourage natural use ‘in-the-wild’. As an example from our review, McKanna, Jimison and Pavel [70] identified the need for an unobtrusive way to continuously monitor and detect cognitive decline in older adults. This goal guided their development of a computer game that naturally appealed to older adults and targeted divided attention, a function associated with daily activities that often declines with age.

2.6.1.5.2 Evaluation

**Evaluating the assessment capacity of assessment games is necessary for establishing their validity.** Given the new variables that any game elements introduce to an assessment, moving forward, a focus on robust evaluation needs to be prioritized for any cognitive assessment game. We can look to cognitive psychology for best practices when evaluating tasks. For example, when developing a new ‘Parametric Go/No-Go’ task, Langenecker et al. [135] took care to measure its sensitivity, construct validity, and test-retest reliability. From our review, Chesam et
al. [64] measured the validity of their Search and Match Task, a puzzle game designed to assess visual search. To do this, they looked at correlation analyses between performance on the game and performance on traditional tasks.

**Measuring the player experience within assessment games is necessary for justifying the use of games over traditional assessment tasks.** In our review, we identified three approaches to evaluating assessment games: comparison to a task, comparison to a clinical diagnosis, and comparison to theories of cognition. Comparison to a task and theories of cognition can help determine construct validity. Comparison to a clinical diagnosis may measure sensitivity. However, there are many other issues that need to be evaluated. For example, does the game work for people from different cultures or with different educational backgrounds? Are there practice effects that may interfere with repeated use? Some of the studies in our review did address questions such as these; for example, Rello et al. [121] assessed their game for both English- and Spanish-speakers, but every study in our review answered different evaluation questions.

It may be useful to develop best practices around how and what should be evaluated when developing an assessment game. This list should include evaluating the assessment aspects in ways similar to cognitive psychology methods, but it also needs to include evaluating the game aspects. Knowledge on how to evaluate games can come from games-user research.

2.6.1.5.3 Interdisciplinary Work

**Integrating knowledge across the disciplines of psychology, clinical sciences, game design, or user experience will help to ensure robust results.** No matter the goals of an assessment game, the best result will often come from an interdisciplinary team. Levy et al. [30] note that, “…the design of scientifically robust games is often at odds with accepted game design practices”. We need to draw from both knowledge of game design and knowledge of cognitive testing. Interdisciplinary work will be key in developing robust, enjoyable games, and it will also be useful in knowing how to evaluate these games, as discussed above. Experts in cognitive psychology, neuropsychology, game design and games-user research can all contribute to this field. As an example from our review, Bottiroli et al.’s [90] ‘Smart Aging’, a game platform
designed to measure various cognitive functions, was developed in collaboration with “neurologists, psychologists, neuropsychologists, bioinformatics, designers, and ICT engineers”.

2.6.2 Limitations

There are some limitations to our systematic review. We did not search every database for papers; for example, we ruled out using Springer Link because we could not search by title and abstract. We picked our databases based on relevance to the field and focused on using a mix of computer science and psychology-related databases. Similarly, we tried several combinations of search terms. Some yielded too many results to feasibly review. We aimed to search as comprehensively as we could realistically manage. With 49,365 records to review after duplicates were removed, we feel like we struck an appropriate balance. We also used cross-referencing and Google Scholar to check for additional papers that met our criteria; however, it is still possible that we missed some papers.

Our review also only addressed papers that used games to assess cognitive processes related to attention, in order to keep this review to a manageable scope. There are many other papers that assess memory and other cognitive functions using games. We intend to cover memory in another review, and future work should cover games that are used to assess more complicated cognitive functions, like decision making.

Another limitation of our selection criteria is our reliance on authors to define their work as a game or task. Our search criteria depended on the word ‘game’ being included in the title or abstract of a paper. We may have missed papers in our review that used what authors defined as a ‘task’, but still implemented game elements, like points. In addition, we may have included papers that use what authors defined as a ‘game’ but could be considered a task.

Finally, our review only covers published work, and thus carries the risk of publication bias. Because our review is focused on the methodologies used by studies and not the outcomes, this issue is not a large concern.
2.6.3 Conclusions

We conducted a systematic review to answer the question: How are digital games used for the cognitive assessment of attention made and measured? We searched a wide range of databases to identify an initial set of 49,365 papers, which we then narrowed to a set of 74 papers that we reviewed in full. From these papers, we identified three unique approaches to developing a game for assessment. We also identified that across the field, the focus tends to be on development rather than evaluation. Assumptions about how the application of games to cognitive tasks should improve assessment are widespread, but perhaps not widely demonstrated. Our review can act as a resource to help guide the field towards the more standardized approaches and rigorous evaluation required for the widespread adoption of assessment games.

2.7 Contribution to this Dissertation

This review provided an overview of the field for my dissertation and pointed to a significant area for further research: understanding how game elements affect data quality. It also emphasized the importance of evaluating both the assessment properties and game properties in any study I conduct.

Our findings highlighted the importance of interdisciplinary work in this space. We need to use methods from psychology to develop scientifically rigorous tasks and evaluate cognitive data. We also need to draw from HCI and games user research to develop motivating games and evaluate the user experience. In the following studies, I use knowledge and tools from both fields.
3 Manuscript B: How Points and Theme Affect Performance and Experience in a Gamified Cognitive Task

3.1 Introduction to Manuscript B

In Manuscript A, we identified the ways in which games for the assessment of attention are made and measured. Ultimately, we found that many assessment games are not sufficiently evaluated. Studies often do not measure both the assessment capabilities and user experience of games. Many studies do not evaluate participant enjoyment, focusing solely on performance. We also found that few studies isolate individual game elements to understand their effects.

Thus, for Manuscript B, we sought to determine which game elements are most likely to improve the data quality of a cognitive task, in terms of both participant performance and enjoyment. We looked at the effects of both points and thematic game elements, individually and in combination.

3.2 Abstract

Cognitive tasks are increasingly being gamified in an attempt to leverage the motivational power of games; however, they are sensitive to manipulation and literature is divided on how adding game elements affects participant performance and experience. We applied two popular gamification approaches (points/feedback and theme/narrative) to a typical cognitive task (the dot probe) and measured performance and experience in two studies (N1=287, N2=321). Similar to prior work, we confirm in Study 1 that points increase reaction time and error rate, and positive affect. We replicated these results in Study 2, and expanded our analysis to investigate participant experience. Our findings suggest that theme creates expectations of an interesting game, which gamified tasks fail to deliver, whereas points maintain enjoyment better throughout the task itself. Important for the development of gamified cognitive tasks, our findings suggest that novel approaches to gameful assessment may be better than the status quo.

---

3.3 Introduction

Cognitive tasks are widely used in science, for collecting information about population norms, for conducting research in understanding cognition, and for informing mental health assessment [2]. High-quality, high-quantity data from these tasks, comprised of information from large cohorts over time, is crucial for developments in psychology and related fields [136]. Traditionally, capturing large-scale datasets was challenging, time-consuming, and expensive; however, advancements in technology have led to digital strategies in which tasks can be deployed remotely and automatically scored and evaluated [137,138].

While it is now possible to collect large amounts of data, it can still be difficult to recruit participants due to the boring and repetitive nature of cognitive tasks. Further, recruited participants may not be fully engaged with the task, due to lack of interest. People often exert suboptimal effort on cognitive tasks [20,21], making their collected data unreliable for interpretation [139]. As well, cognitive tasks are associated with high attrition rates [19], making data difficult to collect on a large-scale or over the long-term.

In attempts to facilitate engagement with cognitive tasks and collect quality data on a large-scale, researchers have turned to games. Over 164 million Americans of all ages, genders, and ethnicities play video games [140]—an attractive sample for psychological research. This broad demographic of gamers extends participant sampling to those who generally have not been involved in research (e.g., moving beyond the convenience of undergraduate student samples). Further, games are highly motivational to play and have been shown to satisfy human needs [35]. This motivational power of games has been leveraged in many contexts [141], such as in education [142] and health [143], giving rise to the phenomenon of gamification, in which game elements are used in non-game contexts [13]. Making cognitive tasks more interesting through gamification holds the potential to increase attention in the moment (leading to better data) and facilitate the long-term engagement of many people (leading to more data) [144].

However, cognitive tasks are sensitive to manipulation; even basic tasks are heavily studied to understand the effects of making small changes [32]. With such sensitive tasks, gamification could affect performance and experience in unexpected ways. Though many tasks have been
gamified [56], few studies isolate individual game elements to understand their effects on data quality; rather, tasks are often gamified by applying several elements at once. Further, studies that do isolate individual game elements have shown mixed results [56]; there is little agreement on how typical gamification approaches affect performance on, and engagement with, cognitive tasks.

In this paper, we investigate how two popular gamification approaches (points/feedback and theme/narrative) affect performance on, and enjoyment of, a typical cognitive task (the dot probe). We created four versions of the dot probe task—the basic task plus versions that included points and theme individually and in combination. In two studies, we found that points enhance experience, but that theme harms enjoyment and player experience. We further explored why theme decreases enjoyment, and provide evidence suggesting that experiencing a theme may raise player expectations, on which the gamified task fails to deliver.

Our results send a cautionary message about gamifying cognitive tasks. It is not always enough to layer game elements on an existing task and expect more engagement. We recommend using reward-based game elements (e.g., points and feedback) to increase engagement with existing cognitive tasks, but also recommend developing new methods of meaningful gameful assessment. Engaging games can be created if they are based on the underlying principles of existing cognitive tasks, but perhaps not on the tasks themselves, allowing for the true potential of games to be leveraged in large-scale cognitive assessment.

3.4 Background

We discuss theories of gamification, gamification of cognitive assessment tasks, and the effects of gamification. We also justify our choice of cognitive task and game.

3.4.1 Gamification

According to Deterding et al. [13], gamification is “the use of game design elements in non-game contexts”. They also suggest ‘gamefulness’, a term that refers to play structured by rules, goals, and competition, and ‘gameful design’. Gamification is then the design strategy of using
game elements, whereas gameful design is the goal of designing for gamefulness. These concepts do not always go together; a task can be gamified but not lead to gameful behavior. Nicholson [14] proposes two types of gamification: reward-based gamification, which targets extrinsic motivation (i.e., the motivation to engage in an activity for a desirable but separate benefit [145]), and meaningful gamification, which targets intrinsic motivation (i.e., the motivation to engage in an activity because it is inherently interesting [145]). Reward-based gamification is implemented by adding elements like badges, achievements, and points. In contrast meaningful gamification uses six types of game elements to target intrinsic motivation: play, exposition (i.e., narrative), choice, information, engagement, and reflection.

These suggestions for meaningful gamification are based on Self-Determination Theory, which states that competence—the experience of mastery over challenges, autonomy—the experience of volition and choice, and relatedness—the experience of connectedness and camaraderie, all contribute to build intrinsic motivation [146]. Games have been shown to offer rich environments for building mastery, autonomy and relatedness, which helps explain why playing them is intrinsically motivating [35]. Due to these motivational features, gamification has been used in a variety of contexts, such as learning (e.g., [147]), behavior change (e.g., [148,149]), persuasion to change beliefs (e.g., [150]), and cognitive assessment. Cognitive tasks for assessment have been gamified with the intention of leveraging the power of games to capture better data through increased engagement in the moment and to capture more data through improved adherence, facilitated by task enjoyment [19].

3.4.2 Gamification of Cognitive Tasks for Assessment

Although there is a large body of work on gamifying cognitive tasks for training and treatment (e.g., [151,152]), we focus our review on cognitive tasks for assessment, which measure cognitive functions and are not considered diagnoses. People are required to trace patterns, memorize sequences, and respond to visual cues, among other tasks. Data from these tasks are used to approximate a measure of a cognitive domain, such as memory or attention [153]. The performance of an individual is then compared to the results from a normative sample group (norms). Differences between an individual’s performance and norms may inform research or indicate mental health issues [154].
However, there are problems associated with cognitive tasks. The interpretation of cognitive task data depends on the assumption that individuals are putting forth their best effort and are focused on the task, but cognitive tasks are often repetitive and boring, and suboptimal effort on these tasks is a common problem [20,21]. Data from an individual who is not fully engaged will not represent their true ability. This discrepancy can lead to inaccurate interpretations [139]. To improve individuals’ engagement, researchers have looked to games [155,156], with Aeberhard et al. [136] stating that “leveraging gamification to repeatedly obtain behavioral samples paves the way for next-generation high-throughput psychometric toolset”. For example, the game ‘Sea Hero Quest’, designed to assess spatial navigation and develop a large normative data set [25], has been played by over 4.3 million people, collecting far more data than would be possible with traditional methods [24]. Furthermore, as identified by Lumsden et al. [56], gamified tasks may be used to increase motivation, usability for various demographics, long-term engagement, ecological validity, and suitability for different disorders.

Previous work has typically taken a cognitive task, digitized it, and then gamified it by layering points and graphics over the task. In a systematic review of cognitive task gamification, Lumsden et al. [56] identified 2D graphics, sound effects, score and theme as prominent tactics for assessment gamification (the gamification of training tasks often uses more complex elements). These gamified tasks may create the appearance of a game through the application of game elements in a non-game activity, but they do not always lead to gamefulness [13]. Further, the inclusion of game-based elements has not always shown an improvement in data quality for cognitive assessment. We next present prior research on different game elements and their effects, limiting our review to papers that explicitly investigated specific game elements for assessment-based cognitive tasks; although gamification is an often-mentioned design intention, the effects of individual game elements are not always extracted or reported.

3.4.2.1 Points and Feedback

Many gamified cognitive tasks include ‘points’, which provide feedback and rewards to a player. Feedback can also be non-point based, but points always give feedback. Often, adding points increases enjoyment [19,153,157]. Points may lead to the experience of competence and thus
intrinsic motivation [158,159], but other research echoes Nicholson’s [14] idea that reward-based gamification (i.e., points) is less effective than meaningful gamification, suggesting that points increase motivation due to extrinsic incentives [17], rather than feelings of competence or intrinsic motivation.

Adding points to a task may also lead to improved performance (e.g., faster reaction times) [56]. In some cases, this improvement in performance can be detrimental to the task’s purpose. While gamification can help increase engagement for a target demographic, it can also pose difficulties for certain populations. For example, gamified tasks can normalize the performance of individuals with ADHD, meaning that the gamified cognitive task no longer differentiates between people with and without ADHD [39].

3.4.2.2 Sound Effects

Research on the use of sound effects in gamification is limited; however, one study found that sound effects (e.g., a xylophone note) resulted in attentional capture and shortened reaction times of a visual search task. Sound effects alone did not increase enjoyment [157].

3.4.2.3 Graphics

Graphics may make a task seem more game-like, but are also used as stimuli in cognitive tasks, and changing them can affect performance. Lumsden et al. [153] used cowboy characters in a Go/No-Go task and saw performance worsen relative to a control condition in which people differentiated green and red objects, possibly because the stimuli were not as easy to discriminate. Birk et al. [160] saw performance drop in a Go/No-Go task with graphic characters, again likely because it is harder to discriminate characters than shapes. Even in a Stop-Signal task in which the stimuli were similar (coloured fruit versus coloured circles), the graphics version resulted in lower performance [19].

Graphics may increase enjoyment of a task. In the cowboy game, while the graphics were detrimental to performance, they increased enjoyment [153]. However, it is difficult to separate out the effect of graphics-based stimuli from the presence of a narrative or theme. Lumsden et al.’s cowboy game not only had cartoon characters, but also a narrative that contextualized the
graphics. In another study, however, graphics without narrative actually decreased enjoyment of a task relative to a control condition [19].

It should be noted that graphics may not always harm performance. It is difficult to draw conclusions as there are only a few studies that examine the isolated effect of graphics, but there are many studies in which cognitive tasks are gamified in multiple ways, including graphics, and which find acceptable performance scores [56].

3.4.2.4 Thematic Elements

Graphics are often added to allow for premise, backstory, character, and narrative—dramatic elements in games that can be described as ‘theme’. Research on including theme shows mixed results for performance and enjoyment.

Participants did enjoy the theme of Lumsden et al.’s [153] cowboy Go/No-Go game, in which they played a sheriff rescuing hostages in a saloon. They had to ‘shoot’ criminals but inhibit response to hostages. The narrative was well-received and raised enjoyment, but as noted, the graphics worsened performance. Birk et al. [160] also found that theme increased the number of hostile words used in an ambiguous word completion task. In the same study, which added premise and backstory to cognitive tasks, participants unexpectedly reported greater enjoyment, relatedness, autonomy and immersion in control tasks compared to tasks with theme. It is unclear why these differences occurred; however, the authors speculated that the packaging of a task as a game may have raised expectations that were then unmet by a thematic version of a cognitive task. Again, it is difficult to draw conclusions on the isolated effects produced by theme, as most studies combine theme with other game elements, such as points [56].

Taken together, the effects on performance and enjoyment of applying game elements to cognitive tasks are mixed, and are further muddied by the dearth of studies that isolate and analyze individual gamification elements, rather than confounding them into a single ‘gamified’ design.
3.4.3 On Choosing a Cognitive Task and Game Elements

In addition to the choice of which gamification elements to apply, research findings are complicated by the use of different cognitive tasks, such as the Go/No-Go [153,160], stop-signal task [19], or an N-back task [160]. When choosing a task, researchers generally are most interested in what the task is meant to assess. In our research, we chose to study how gamification elements, when applied in isolation, affect performance and enjoyment of a standard, well-known, and highly-studied task—the dot-probe task.

3.4.3.1 Choosing a Cognitive Task

In the dot-probe task, individuals are presented with two sets of stimuli, one relevant to an area of interest, and one neutral (see Methods for additional details). Individuals may demonstrate an attentional bias to certain stimuli [161]; for example, individuals with social anxiety respond preferentially to images of threatening faces, in that they are both faster to engage with the threatening stimuli and slower to disengage with it [162,163]. The dot probe task has been used to study individuals with anxiety, post-traumatic stress disorder, insomnia, and other disorders [162,164–166], and research has validated the use of a remote dot probe task delivered via the internet [167]. As a typical cognitive task that is already digital, validated for online use and heavily studied, the dot probe task is suitable for us to explore the effects of gamification on participant engagement and data quality.

3.4.3.1.1 Gamification of the Dot Probe Task

The dot probe task has not previously been gamified for the purposes of assessment. However, it has been gamified for training. Attentional-Bias Modification (ABM) uses the dot probe task to train users to attend towards preferential stimuli and/or away from undesirable stimuli [168–170]. For example, individuals with alcohol addiction problems demonstrate an attentional bias towards alcohol-related stimuli. An ABM dot probe task can train this bias away by presenting two stimuli, one alcohol-related and one neutral stimulus, and then requiring individuals to respond to the neutral stimulus. While at first alcoholic individuals will be distracted by the alcohol-related stimuli, over time they become faster at responding to the neutral stimuli [171]. ABM tasks have also been used for other addictions [172,173], anxiety [170,174], depression [175], and more [176,177].
Dennis and O’Toole [178] created a gamified version of an ABM dot probe task, designed for use as a mobile app. They incorporated animated characters, points, sound effects, and a swiping motion into a typical dot probe task. When played for over 45 minutes, this game reduced threat bias in anxious individuals. Another gamified version of an ABM task, called ‘Shots’ was designed by Boendermaker et al. [179] to influence heavy-drinking undergraduate students. This game, while functionally identical to the control ABM dot probe task, included rewards (in-game coins and points), a graphical theme, sound effects, and animations. While participants’ motivation to train decreased in all conditions, it decreased even more in the game condition. The authors suggest this effect could be due to participants’ expectations; they expect a game to be more than just a point-system with graphics, and thus experience disappointment and decreased motivation when the game fails to be fun. As well, only the regular training condition was successful in reducing attentional bias. Boendermaker et al. [179] conclude that the game elements distracted participants, hence the lack of improvement.

Game elements can also be added outside of the actual task; for example, Birk and Mandryk [180] added avatar customization to an ABM dot probe task. Participants were either assigned an avatar, or created one themselves, and these avatars were then used as stimuli by giving them angry or neutral expressions. Avatar customization increased both engagement and training efficacy, even though the task itself was identical in both conditions.

### 3.4.3.2 Choosing Game Elements

For assessment-based cognitive tasks, typical gamification strategies can be grouped into two categories: reward/feedback features, and narrative/theme features. Reward/feedback features include points and rewarding feedback (e.g., visual effects, bonus points, high scores). Narrative/theme features involve cohesive graphics, premise, characters, and story. We chose to focus on these two groups of game elements due to their popularity as gamification strategies in our review of previous literature.
3.4.4 The Present Research

We add to previous work on gamified cognitive assessments by examining the effects of reward/feedback features and narrative/theme features on participant performance and enjoyment in a typical cognitive assessment tool—the dot probe task. We developed several versions of the dot probe task, which individually layered different game elements, and in two studies, we measured their effects on reaction time, error rates, and various measures of enjoyment and experience.

3.5 Study 1

The goal was to understand the effects of points and theme on performance and explore effects on experience.

3.5.1 Methods

We manipulated the inclusion of points and feedback (which we call points) and narrative and thematic elements (which we call theme) to develop four versions of a dot probe task: 1) a basic task as a control (points: off; theme: off); 2) a task with points/feedback (points: on; theme: off); 3) a task with narrative/theme game elements (points: off; theme: on); and 4) a task with both points/feedback and narrative/theme elements (points: on; theme: on).

3.5.1.1 Tasks

See Figure 3 for images of the tasks.

**Basic Control:** This task was designed as a basic dot probe task. Following standard protocols, we presented a fixation point (+) in the center of the screen for 500ms, followed by two images simultaneously displayed on the left and right side of the screen, also for 500ms. After this, a probe stimulus (a large black dot) was displayed on either the left or right location, and remained until participants made a response of either the left or right arrow keys. Participants were instructed to select the arrow key that corresponded with the probe’s location.

**Theme:** This version of the task added narrative and theme game elements to the basic control task. Participants played as a racoon who was recruited by another racoon, ‘Bijou’ to find their
family’s stolen paintings. The instructions for the task describe the story of an art thief, ‘Zabat’, who had stolen the raccoon family’s art collection and hidden them in a cave. To mask his trail, Zabat mixed the authentic paintings with inauthentic ones, and the only way to distinguish them was to look for a red seal underneath the paintings. After an introduction to this story, participants then proceeded to the dot probe task, with the images as the paintings and the probe as the red seal.

**Points:** This version of the task added reward and feedback game elements to the basic control task. The dot probe design was the same, though participants now earned and lost points, were given a score, and could see their reaction times to each trial. Faster responses were rewarded with more points. When participants indicated the dot location, colorful confetti animated on the screen for a correct response, and they lost points for an incorrect response.

**Points and Theme:** Finally, this version of the task combined the Points task and the Theme task. For this task, participants were presented with the same narrative about Zabat, and also were awarded points and given feedback. When participants correctly indicated the seal location, in addition to colorful confetti, a smiling Bijou appeared (and a frowning Bijou for incorrect responses).
3.5.1.2 Stimuli

The stimuli shown in all conditions were images of faces expressing different emotions; either non-threatening stimuli (neutral or happy expressions) or threatening stimuli (angry or sad expressions). Images were shown in non-threatening/threatening pairs (i.e., non-threatening on the left side and threatening on the right, or vice versa).

We took pictures of 14 student actors (9=male, 5=female). The actors made four expressions (angry, sad, neutral, happy) after being coached on how to do so (e.g., “purse your lips and furrow your eyebrows”). We then added art filters (to match the story of our theme conditions) to each of the images, as shown in Figure 3. We conducted a pre-study to validate the images’ emotional valence when filtered, along with filtered images from the Ekman set of faces [181]. Raters scored each image using the Self-Assessment Manikin (SAM), a pictorial scale used to assess the arousal and valence of a stimulus [182]. Participants also rated each image for happiness, anger, and sadness on a 7-point Likert scale from low (1) to high (7). In general, the art filters increased the arousal of images, but maintained the same valence when compared to the unfiltered image. We eliminated images that had too great a deviation in valence from their
corresponding baseline image, resulting in 14 images for the study (3 from our actors, 11 from the Ekman set of faces). We used the filtered pictures for all the conditions to maintain the consistency of the stimuli and reduce any problems with graphic elements changing the nature of the task, as seen in previous studies [153].

3.5.1.3 Procedure

We used Amazon Mechanical Turk (MTurk) to recruit participants. MTurk is commonly used in human-computer interaction research to conduct studies [183], and has been shown to yield reliable data when precautionary methods around data gathering and analysis are used (e.g., [184,185]).

The study used a between-subjects design, with each participant being shown one of the four versions. Prior to the task, participants were surveyed on demographic information and game-playing habits. Participants also completed the Positive and Negative Affect Schedule (PANAS) before and after the task, in order to explore experience. PANAS measures the two primary dimensions of mood valence by asking participants to rate their agreement with different sentences (e.g., “I feel attentive”, “I feel irritable”) on a five-point scale, which translate into a measure of positive affect and a measure of negative affect [186]. The experiment took 15 minutes to complete on average and we compensated participants with $3.00 USD.

3.5.1.4 Analysis

We received 349 responses. Participants who did not meet threshold quality criteria were removed, according to best practices [184,185]. We removed participants who completed the questionnaires too quickly (faster than 1.5 seconds per item) (n=40), those who showed zero variance within the questionnaires (e.g., answered every question with ‘disagree’) (n=9), and finally, those who incorrectly responded to the dot probe too frequently (total incorrect responses > 1 standard deviation (SD) above the mean number of incorrect responses) (n=13). Our final sample consisted of 287 people (122 female, 161 male, 4 non-binary, mean age=37.5, SD=12.03).
We conducted two-way ANOVAs with points (on/off) and theme (on/off) as between-subjects factors on reaction time (RT) and error rate data, and two-way ANCOVAs on post-play positive and negative affect, controlling for pre-play affect. To ensure a normally distributed data set for RT (which tends to be positively skewed), we took the log value of RT, as suggested when working with reaction time data [161,187]. Our required level of significance was p<.05.

3.5.2 Results

See Table 3 for means and standard deviations, and results of the statistical tests.
Table 3. Results of the statistical tests for Study 1 and 2.

<table>
<thead>
<tr>
<th></th>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 2, P&lt;sub&gt;i&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Error</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td></td>
<td>Rate</td>
<td>Affect</td>
<td>Affect</td>
</tr>
<tr>
<td>No Theme</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Points</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.21</td>
<td>1.33</td>
<td>13.38</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>1.53</td>
<td>7.15</td>
</tr>
<tr>
<td>Points</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.02</td>
<td>4.90</td>
<td>13.08</td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>4.72</td>
<td>6.81</td>
</tr>
<tr>
<td>Theme</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Points</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.20</td>
<td>1.24</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>1.92</td>
<td>6.56</td>
</tr>
<tr>
<td>Points</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.04</td>
<td>4.44</td>
<td>12.28</td>
</tr>
<tr>
<td></td>
<td>0.18</td>
<td>3.39</td>
<td>6.99</td>
</tr>
<tr>
<td>Statistical Tests</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Points</td>
<td>F</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>71.51</td>
<td>&lt;.001*</td>
<td>81.63</td>
</tr>
<tr>
<td>Theme</td>
<td>F</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.56</td>
<td>0.43</td>
</tr>
<tr>
<td>Interaction</td>
<td>F</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.40</td>
<td>0.25</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>.527</td>
<td>.615</td>
<td>.614</td>
</tr>
</tbody>
</table>
3.5.2.1 **Performance**

We found significant main effects of points on RT and error rates. There were no significant differences in RT or error rate as a result of including theme. There was further no interaction between points and theme on RT or error rate.

3.5.2.2 **Experience**

In terms of experienced negative affect, there were no significant effect of points or theme, and no significant interaction. In terms of positive affect, participants rated their mood more positively when points were included. There was no significant effect of theme and no significant interaction; see Table 3 for results of statistical tests.

3.5.3 **Discussion of Study 1**

Our results show that the inclusion of points affects performance by increasing reaction time and error rates. Points also increase affect; however, this same increase is not seen for theme. These findings echo previous research on thematic elements. Typically, researchers have included theme expecting increases in enjoyment; however, in studies that isolate the effects of theme, enjoyment has not increased [19,160]. In fact, Birk et al. [160] even found that participants’ experienced higher immersion, enjoyment, and relatedness after playing a basic task compared to a task with thematic elements (premise, backstory, and corresponding graphics). Our results, similar to these results in the literature, seem counter-intuitive. Nicholson [14] suggests using exposition (i.e., story and theme) as part of meaningful gamification. Villani et al. [188] suggest that games can improve emotional regulation and mood, in part because narrative elements immerse the player in the game and allow them to take on another identity. Following these guidelines, we included theme elements that created a story and allowed participants to take on a new identity of a wronged raccoon, eager to recover their stolen art. We anticipated increased affect from these elements, but instead saw results that mirror what others have seen—that thematic elements do not necessarily increase experience. Prior work, and our results that suggest how theme is not helpful, prompted us to conduct Study 2. We designed Study 2 to confirm that theme does not improve experience and further investigate why it is detrimental when applied.
3.6 Study 2

The goal of Study 2 was to replicate the performance findings observed in Study 1; specifically, that including points decreases RT and increases error rate, and to understand why theme is detrimental to experience.

3.6.1 Methods

Study 2 used the same design, tasks and stimuli as Study 1, with additional questionnaires to measure enjoyment.

3.6.1.1 Procedure

In Study 2, we followed the same procedure as in Study 1, but also included player experience measures. We added an enjoyment questionnaire that was given before, during (after the introduction sequence but before play), and after the task. Comprised of five statements—“I am enjoying playing this game”; “Playing this game is fun”; “This game is interesting”; “I am thinking about how much I am enjoying playing”; “This game is not holding my attention” (reverse-coded)—participants rated their agreement on a 7-pt scale. Items were worded for each time point, e.g., “I will enjoy playing this game”, “I am enjoying playing this game”, and “I enjoyed this game”. After the task, they completed the Player Experience Inventory (PXI), which measures ten constructs: meaning, mastery, immersion, autonomy, curiosity, ease of control, challenge, progress feedback, audiovisual appeal, and goals and rules [189]. Finally, they were optionally asked an open-ended question: “Please explain why you did or did not enjoy this game.”

3.6.1.2 Data Analyses

We followed the same analyses as Study 1. We received 391 responses. After filtering out participants to ensure data quality, our final sample consisted of 321 people (143 female, 177 male, 1 other, mean age=37.3, SD=10.98). Again, we used the log value of RT [161,187]. We conducted two-way ANOVAs on RT and error rate data, and two-way ANCOVAs on post-play positive and negative affect (controlling for pre-play affect), identical to Study 1. For enjoyment ratings, we conducted a repeated-measures ANOVA with time (before, during, after)
as the within-subjects factor and points and theme as the two between-subjects factors. As the PXI has 10 subscales, which share variance, we conducted a MANOVA with points and theme as between-subjects factors and the 10 subscales as dependent measures. We did not conduct a qualitative analysis of the open-ended question; however, we include some quotes in our discussion as examples.

3.6.2 Hypotheses and Research Questions

For this chapter, we made the following hypotheses and research questions:

**H1. RT will be faster when points are included.** Based on the results of Study 1, we expected the same result.

**H2. Error rate will be higher when points are included.** We expected the same result as seen in Study 1.

**H3. When controlling for incoming affect, post-play positive affect will be higher when points are included.** Based on Study 1, we expected the same result.

**H4. Immersion in the game will be lower when theme is included.** We expected this due to previous work, which found lower immersion with thematic elements similar in style and narrative to our own [160].

**H5. Enjoyment of the game will be lower when theme is included.** We expected this due to the results of Study 1 on affect, combined with previous work which shows lowered enjoyment with thematic elements [19,160].

**RQ1. If enjoyment is lower when theme is included, why?** If the drop in enjoyment is due to raising expectations with theme and not meeting them, then we should see higher ratings following the introduction of the theme, and a significant drop post-game. However, if the drop in enjoyment is due to disliking the game’s design, then we should see lower ratings following the introduction of the theme, and a smaller drop post-game.
RQ2. How is player experience changed by the inclusion of points and theme? We included the PXI to understand how the inclusion of points and theme affect the experience of the task along constructs of player experience.

3.6.3 Results

See Table 3 for means and standard deviations, and results of the statistical tests.

3.6.3.1 Performance

As expected, we found significant main effects of points on RT and error rates. There were no significant differences in RT or error rate as a result of including theme. There was further no interaction between points and theme on RT or error rate. See Table 3 for results of the statistical tests.

3.6.3.2 Experience

PANAS: In terms of experienced negative affect, there were no significant effect of points or theme, and no significant interaction. In terms of positive affect, participants rated their mood more positively when points were included. There was no significant effect of theme and no significant interaction; see Table 3.

Figure 4. Estimated marginal means of enjoyment ratings, before, during and after the task.
**Enjoyment:** There were no main effects of points ($F_{1,317}=2.155$, $p=.143$), theme ($F_{1,317}=.001$, $p=.975$), or interaction on enjoyment ($F_{1,317}=.075$, $p=.784$). There was a main effect of time ($F_{1,317}=305.487$, $p<.001$), with interactions between points and time ($F_{1,317}=19.114$, $p<.001$) and theme and time ($F_{1,317}=12.686$, $p<.001$). The main effect of time showed decreases in enjoyment at each time point (all $p<.001$). The interaction of points and time show that there was no difference in enjoyment before play ($p=.964$) or during play (after the game introduction but before the actual task/game; $p=.269$), but that points increased enjoyment ratings after play had completed ($p<.001$). The interaction of time and theme showed that there was no difference in enjoyment before play ($p=.906$), but theme increases enjoyment during play ($p=.005$). After play, enjoyment ratings were marginally lower ($p=.057$) as a result of theme. There was no three-way interaction ($F_{1,317}=2.232$, $p=.108$). See Figure 4. Thus, we provide evidence for H5 and answer RQ1.

### 3.6.3.3 Player Experience Inventory (PXI)

See Table 3 for means and standard deviations. As the MANOVA results were significant for points ($F_{10,308}=50.69$, $p<.001$, $\eta^2=0.62$), theme ($F_{10,308}=6.12$, $p<.001$, $\eta^2=0.17$), and the interaction ($F_{10,308}=2.02$, $p=.031$, $\eta^2=0.06$), we report the univariate results (see Table 3).

**Meaning:** There was a significant effect of both points and theme, with points yielding higher meaning and theme yielding lower meaning. There was also a significant interaction. Pairwise comparisons revealed that that the loss of meaning from adding theme is only present when points are not included ($p=.003$). With points, the negative effect of adding theme disappears ($p=.999$); i.e., points seem to protect against the harm to meaning from adding theme.

**Immersion:** There was a significant effect of theme, in which theme lowered immersion. There were no significant effects of points or interaction. The results from this construct provide evidence to accept H4.

**Autonomy:** There was a significant effect of points, with points associated with higher autonomy. There were no significant effects of theme or interaction.
Ease of control: There was a significant effect of theme, with theme associated with higher ease of control. There were no significant effects of points or interaction.

Challenge: There was a significant effect of both points and theme, with points associated with higher challenge and theme associated with lower challenge. There was no significant interaction effect.

Progress and feedback: There was a significant effect of points, with points yielding higher progress and feedback. There were no significant effects of theme or interaction.

Audiovisual appeal: There was a significant effect of points, with points yielding higher audiovisual appeal. There were no significant effects of theme or interaction.

Goals and rules: There was a significant effect of both points and theme, with both points and theme yielding higher goals and rules. There was no significant interaction. There were no significant effects for Mastery or Curiosity.

3.6.4 Discussion of Study 2

Like Study 1, Study 2 found an increase in RT and error rate with points. Theme did not affect performance. Points increased positive affect, and while enjoyment suffered across time points for all conditions, the decrease was smallest when points were included. Enjoyment was lowest when theme was used alone, even lower than the basic task. The PXI further found that theme was associated with lower meaning, immersion, and challenge, though higher ease of control and goals and rules. In contrast, points increased meaning, autonomy, challenge, progress and feedback, audiovisual appeal, and goals and rules.

The results confirmed our hypotheses. As expected, the addition of points decreased participant reaction time (H1), and increased error rates (H2), and positive affect (H3), whereas immersion (H4) and enjoyment (H5) were lower with the addition of theme. We also addressed why theme lowered enjoyment. Enjoyment ratings collected during the game (after the introduction
sequence) were higher when theme was included; however, after game play, the theme-only condition had the lowest rating of all. These results suggest that the lower ratings were not because participants disliked the theme itself, but rather that the game play was dissatisfying after having been introduced to the theme. This is likely due to the heightened expectations created by the introductory exposition, and then disappointment when the game play failed to meet those expectations.

3.7 Discussion

We summarize our results, make design recommendations, and discuss limitations and future work.

3.7.1 Summary of Results

Our objective was to evaluate how adding game elements to a dot probe task affect task performance and player experience. Gamification of cognitive tasks often conflates many game elements into a single ‘gamified’ design that is compared to a control task [56], leading to little knowledge on how single elements affect performance and experience. Together, our studies make several important findings:

1. Points decreased reaction time and increased error rates (Study 1, Study 2)
2. Points increased positive affect (Study 1, Study 2)
3. Theme lowered enjoyment and immersion (Study 2)
4. Enjoyment was actually higher after a theme-based introduction to the task, but dropped after play when the game failed to deliver on player expectations (Study 2)
5. Enjoyment of the game was higher when points were included (Study 2)
6. Points increased challenge and meaning, whereas theme lowered challenge and meaning (Study 2)
7. Both points and theme increased the clarity of goals and rules (Study 2)

3.7.2 Explanation of Results

We explain our results in regard to the effects of points and theme on performance and experience.
3.7.2.1 Performance

**Points:** Our analysis of reaction time data found that when points and feedback are embedded into the dot probe task, participants are faster. This increase in speed suggests that adding points may manipulate attention. In this case, participants may have been motivated to concentrate on the location of the probe so that they could score points, see animations, and feel a sense of accomplishment as they got faster. Without points, there were no incentives to pay attention. Cognitive tasks rely on engaged participants. If a participant is daydreaming, distracted, and failing to pay attention, the data does not represent true cognitive ability. Increasing motivation is one way to increase data quality.

It should be noted that while our participants were faster when points were included, they were also more error prone. Higher engagement may not be worth the loss of accuracy. A speed-accuracy trade-off is generally expected and has been found with other gamified tasks [190]. The increased RT and error rates may interfere with a task’s interpretation and ability to discriminate.

**Theme:** Theme did not affect performance in our studies. Previous research has found thematic elements detrimental to performance when the graphics interfere with the task. Given these results, for our studies we were careful to maintain consistent stimuli across conditions, to better isolate the effect of theme (versus the effect of graphics).

3.7.2.2 Experience

**Points:** The inclusion of points increased positive affect in both studies. Additionally, points yielded higher meaning, autonomy, challenge, progress and feedback, audiovisual appeal, and goals and rules. These findings reflect research on the mood repairing effects of games. In-game success is linked to higher moods and need satisfaction [188,191]. While participants were equally ‘successful’ in all versions of the task, the addition of points made the task seem more challenging, and success more meaningful. Correct trials corresponded to positive feedback, experienced as higher score communicated in visual effects. Self-Determination Theory (SDT) suggests that challenges and positive feedback contribute to feelings of competence, and thus help satisfy needs [159]. Further, points anecdotally led to gameful behavior [13] as some participants described trying to beat a high score: “*i enjoyed it because i was able to challenge*
myself to get better reaction times, so i was able to immerse myself in the game and had fun doing it” [points/theme].

Even when participants acknowledged the limitations of the game, they were able to find ways to engage with the game: “There wasn't a whole lot to keep me focused besides wanting to beat my score. Luckily I'm a sucker when it comes to silly goals” [points/no theme].

While these quotes are anecdotal, they reflect potential reasons for why RT was faster when points were included, and link the speed increase to the experience of challenge, feedback, and enjoyment.

**Theme:** We anticipated that adding theme would enhance post-play affect; however, we did not see this in Study 1 and confirmed in Study 2 that theme was detrimental to enjoyment ratings post-play, even though enjoyment ratings were higher after the theme-based introduction. As others [179] have suggested, this may be due to expectations. Expectancy-disconfirmation theory suggests that individuals’ satisfaction with an experience can be explained by differences between their expectations and what actually happens [192]. Digital games are supposed to be fun and are often elaborate, with incredible graphics, deep stories, and interesting mechanics. In our study, the theme elements outlined a story and introduced characters. When that depth failed to carry through to the game play, participants may have felt more disappointment and disconnection from the game than if they were never introduced to the theme. When we call a task a game but then simply add a basic story, it fails to meet player expectations, leading to disappointment. For example, one participant in the theme-only condition said, “I thought it was cute and fun at first, but then there were too many trials and it became monotonous. I would have liked more interaction during the game.”

This theory may also explain why feelings of immersion were lower when theme was included. Jennett et al. [193] describe immersion as “a lack of awareness of time, a loss of awareness of the real world, involvement and a sense of being in the task environment”. The discrepancy between participants’ expectations and the reality of the game play may have interfered with their engagement in the task.
3.7.3 Implications for Design

Our results point to several important considerations when designing gamified cognitive tasks.

3.7.3.1 Considerations about Performance

First, there are trade-offs with gamification: adding reward and feedback game elements increased engagement and RT. However, these elements also resulted in higher error rates. Any speed-accuracy trade-off needs to be investigated to ensure data from the task can be interpreted correctly. In our study, the interpretation of the dot probe task relies on the difference in RT between types of stimuli (a ‘bias’ score), rather than the overall RT. In order to fully examine the effects of theme, we obscured our stimuli with art filters, so we did not expect to gather interpretable data and cannot draw conclusions about how increased RT and error rates affect interpretation. Other studies have managed these changes in performance; for example, Tong and Chignell’s [190] gamified task led to higher RT and error rates, leading them to derive an overall performance score that better correlated with cognitive abilities than assessing speed and accuracy separately. The speed-accuracy trade-off could also be manipulated in a game’s design. Miranda and Palmer [157] suggest that point schemes can be allocated based on the goals; for example, in prioritizing speed over accuracy, and thus shaping behavior.

Performance may also suffer due to game elements. Boendermaker et al. [179] suggest that the game elements in their ABM game distracted participants from the goals of the game. In our study, if we had added many more reward and feedback game elements, our results may have been too error-prone to provide quality data. Each gamified task should be validated to ensure changes in performance can be accounted for when interpreting the data.

3.7.3.2 Considerations about Experience

Our results suggest that it may be important to design high-quality games that meet expectations. Our simple story, cartoon characters and graphics were insufficient to increase enjoyment. Vanden Abeele et al. [28] discuss the importance of game quality. They compared an existing game designed to measure psychoacoustic thresholds in preschoolers, with a new game which was designed by dyslexia researchers and game designers. The addition of game researchers to
the development team resulted in a game that children vastly preferred and enjoyed, and was able to measure lower thresholds than the original game.

Sometimes, the inherent properties of cognitive tasks may not lend the task to meaningful gamification, even for the most skilled game designer. Current cognitive tasks rely on uninterrupted, repeated trials, in distraction-free contexts. Slight changes to a task can affect data quality. Even non-gamified versions of tasks must be carefully controlled. For example, a difference of 500ms in exposure to a stimulus can affect results [194]. Tasks this sensitive to manipulation may be difficult to modify in meaningful ways.

Tong and Chignell [190] suggest that designers should “design game components to reflect psychometric properties of existing neuropsychological tasks” by looking to validated tasks and incorporating relevant features. Due to the nature of many cognitive tasks, designing high-quality games for assessment may require moving away from glorified digital copies of already-existing tasks, and instead design games that target the underlying principles of cognitive tasks in a new way. We should still look to validated tasks, but then also beyond them to the new capabilities that games and digitization offer.

Some research has started to move in this direction. For example, while Boendermaker et al.’s [179] gamified dot probe task did not work for ABM, Notebaert et al. [195] were able to create a successful ABM game, by targeting the underlying principles of a dot probe task. Their game displayed several stimuli/faces at once, moving around the screen, with instructions to track the face with a certain emotional expression. At times, the current target face would take on a new expression, while another face simultaneously took on the target expression. Participants then had to make the switch to tracking the new face as quickly as possible. At the end of each block, participants were given various scores based on how well they did and were encouraged to beat their current high score.

Another approach is to start from existing successful games. Commercial games already have mechanisms to collect vast amounts of data on measures like RT, working memory and more. What we do not yet understand is how to interpret that data given the game context. It may also
be possible to embed specific tasks, meant for assessment, within an enjoyable game. For example, many Triple-A games include embedded mini games that do not interact or interfere with the overall performance or narrative of the story (e.g., lockpicking in Skyrim, hacking puzzles in Bioshock). Some games are made almost entirely of mini games in which players complete memory, reaction time, and gambling tasks (e.g., Mario Party), which are similarly simple and repetitive as cognitive tasks. Tasks embedded as mini-games in complex high-quality games may be less likely to yield lower enjoyment through raised expectations.

Adhering strictly to copies of cognitive tasks overlaid with game elements may be useful at times, but also limits the potential of games. Based on our results and synthesis of the literature, we recommend when to consider employing the two approaches to gameful assessment, described by Nicholson [14]. Reward-based gamification: Fairly simple game elements like points seem to improve enjoyment without drastically changing the underlying task. Theme often complicates stimuli and obfuscates tasks, and even if it does not, it hinders enjoyment when the task fails to match players’ expectations. Meaningful gamification: In order to fully leverage the power of games for assessment, designers must explore new methods that move beyond existing tasks. Elements like theme and narrative that are associated with meaningful gamification could be useful when paired with game play that adheres to players’ expectations of a well-designed game. This kind of design likely means creating new methods of assessment that use the underlying principles of cognitive tasks in a new way. New methods require rigorous validation, but the potential for enjoyment and more meaningful assessment is higher.

3.7.4 Limitations and Future Work

While we used a typical cognitive task and two popular gamification techniques, there are many tasks and many game elements that may interact differently. Our research provides some general guidelines about potential effects of points and theme, but each task and game iteration will be different. As tasks are gamified, each task needs to be validated to a rigorous standard before it can be used to collect meaningful data. Similarly, because we were focused on investigating isolated game elements, we did not design our dot probe tasks to generate interpretable data (i.e., bias scores). As well, individual differences will further affect how gamified tasks work. For example, gamified elements can normalize the performance of individuals with ADHD [20]. Any
dot probe game meant for collecting data would need to accurately assess attentional bias and ensure that the game elements did not differentially affect individuals. Finally, gamified tasks are not always appropriate for research. While they can assist in collecting large quantities of data, they may not be suitable for more specialized questions; for example, when looking at specific effects in the brain.

Future work should also investigate new methods of gameful assessment, such as using data from existing games, or embedding tasks into more complex games.

3.8 Conclusion

Using gamification has the potential to increase the quality and quantity of data that can be collected from cognitive tasks. However, game elements cannot be used with the expectation that they will always lead to better performance and experience. Research has shown gamification to have a variety of effects on cognitive tasks, especially in regard to theme. Our study investigates why this occurs, and suggests that gamified tasks should align with players’ expectations of a game. In order to achieve this, we encourage the development of new methods for gameful assessment, which use the underlying principles of cognitive tasks while leveraging the full capabilities of digital games.

3.9 Contribution to this Dissertation

In Manuscript A, we found that a primary reason for gamifying cognitive tasks is to increase user enjoyment; however, user enjoyment is rarely evaluated. In Manuscript B, we found that game elements do not always lead to increased enjoyment. We further found that not only does it matter that we evaluate the user experience, but it also matters how we conduct that evaluation. In Study 1, we only used the PANAS questionnaire to assess user experience. We then needed to conduct Study 2 to fully understand the user experience across the duration of the tasks.

We also found that the use of points affects participant performance, by decreasing reaction time and increasing error rates, a speed-accuracy trade-off. We wondered how this speed-accuracy trade-off affected data quality, which led to the design of the study in Manuscript C.
4 Manuscript C: Measuring the Reliability of a Gamified Stroop Task: Quantitative Experiment³

4.1 Introduction to Manuscript C

In Manuscript B, we found that adding game elements to a cognitive task may affect the user experience in unintended ways. Our results, along with the literature, suggest that traditional gamification (i.e., layering game elements onto typical cognitive tasks) may work best with reward-based elements like points. When using elements associated with meaningful gamification, like theme and narrative, we may need to move beyond just layering game elements onto tasks, and further develop the game play to align with players’ expectations.

Developing more complex games that also function as cognitive tasks requires significant resources and may not always be feasible. We began to wonder if there was value in using reward-based gamification beyond the modest gains in user enjoyment. After all, while increasing user enjoyment is a primary motivator for using gamification, often it is coupled with the desire to increase data quality. If users enjoy a task more, they will be more engaged with the task and thus the data will be better quality. However, data quality may not be conditional on user enjoyment. Users may be more engaged with a gamified task even if they do not report enjoying it more.

We designed the following study to examine the data quality of a gamified task, where the goal was to influence user performance, rather than user enjoyment.

4.2 Abstract

**Background:** Few gamified cognitive tasks are subjected to rigorous examination of psychometric properties, despite their use in experimental and clinical settings. Even small manipulations to cognitive tasks require extensive research to understand their effects.

---

³ The manuscript in this chapter was published as: Katelyn Wiley, Phaedra Berger, Maximilian Aachim Friehs, and Regan Lee Mandryk. Measuring the Reliability of a Gamified Stroop Task: Quantitative Experiment. JMIR Serious Games 2024;12:e50315.
**Objective:** With this study, we seek to research how game elements can affect the reliability of scores on a Stroop task. We specifically investigate within and across session performance consistency.

**Methods:** We created two versions of the Stroop task, with and without game elements, and then tested each task with participants at two time points. The gamified task used points and feedback as game elements. In this paper, we report on the reliability of the gamified Stroop task, in terms of internal consistency and test-retest reliability, compared to the control task. We used a permutation approach to evaluate internal consistency. For test-retest reliability, we calculated the Pearson correlation between each time point, as well as Intraclass Correlation Coefficients to indicate the degrees of consistency and agreement between each time point. We also descriptively compare the reliability of scores on a trial-by-trial basis, with consideration of the different trial types.

**Results:** At the first time point, the Stroop effect was reduced in the Game condition, indicating an increase in performance: Participants in the Game condition had faster reaction times ($p = .005$) and lower error rates ($p = .047$) than those in the Basic Task condition. Further, the Game condition led to higher measures of internal consistency at both time points for error rates as well as reaction times, which indicates a more consistent response pattern. Test-retest reliability analysis revealed a distinctive performance pattern depending on the trial type that may be reflective of motivational differences between the task versions. In short, especially in the incongruent trials where cognitive conflict occurs, the performance in the Game condition reaches peak consistency after 100 trials, while performance consistency drops after 50 trials for the basic version and only catches up to the game after 250 trials.

**Conclusions:** Even subtle gamification can impact task performance, albeit not only in a direct difference in performance between conditions. People playing the game reach peak performance sooner, and their performance is more consistent within and across sessions. We advocate for a closer look at the impact of game elements on performance. Further, given the increased reliability of game-like tasks, they may be especially suitable for assessing populations that are not able to perform the task for an extended time period.
4.3 Introduction

4.3.1 Background

In 1886, James Cattell observed that it takes people longer to name colours and pictures of objects than it does for them to read the corresponding word [4]. This experiment, along with others, paved the way for the development of what Cattell would call “mental tests” and what we now call “cognitive tasks”. Based on these and other results, J.R. Stroop developed a test of cognitive ability, which had study participants read the color but not the meaning of a color-word aloud [5]. Results revealed an interference effect if the word color and word meaning did not match. Typical cognitive tasks require people to respond to such visual or auditory cues while data about their responses, often reaction time and accuracy, are collected. These data can then be used for studying human cognition, creating population norms, and informing medical decisions, such as dementia diagnoses [2].

Cognitive tasks are most useful when they can collect high-quality, high-quantity data. This process can be challenging. Traditionally, capturing large datasets has been time-consuming and expensive, requiring highly trained professionals to administer and score tasks with individual participants. With technological advancement, tasks can now be administered via computer, deployed remotely, and automatically scored [137,138]. This automation makes it easier to collect large quantities of data but raises new concerns around data quality. There are many factors that influence cognitive test performance over and above cognitive capacity, such as motivation, stereotype threat and fatigue [196,197]. Cognitive tasks are often repetitive and boring, leading to high attrition rates [19] and suboptimal effort from participants [20,21].

In attempts to improve the quality of data collected by such tasks, researchers have increasingly turned to gamification, with the hope that tasks can be made more engaging through the addition of game elements like points and graphics.

4.3.2 Cognitive Task Gamification

Deterding et al. [55] define gamification as “the use of game design elements in non-game contexts”. In the context of cognitive tasks, this process typically involves layering game
elements over an already existing task. For example, the Go/No-Go task has commonly been gamified by adding points [198], narrative elements [65], and fun graphics [114] to the basic task.

4.3.2.1 Enjoyment and Motivation

Usually, tasks are gamified with the intent to increase participant enjoyment and motivation. Nicholson [14] notes that gamification can target both extrinsic and intrinsic motivation, depending on the game elements used. Reward-based elements—like points, achievements, and badges—target extrinsic motivation, whereas elements like play, exposition and choice target intrinsic motivation. Through targeting motivation, researchers aim to combat attrition and encourage repeated, prolonged play [76,199,200].

However, there is little examination of whether participants do experience increased enjoyment when tasks are gamified. In a systematic review of gamified attention tasks, only 25 of 74 studies reported results from an evaluation of the game play [199]. When enjoyment is measured, research shows mixed results. Some studies do find that gamification increases motivation; for example, participants in a Stop-Signal task study experienced higher enjoyment and more flow-like experiences when in the gamified condition (as opposed to the basic task) [80].

Other studies have found that certain game elements, especially thematic or narrative elements, can have a negative effect on self-reported enjoyment of cognitive tasks [19,53,160], possibly due to the “chocolate-covered broccoli” effect [29]. Tasks can only be gamified so much and still retain the important elements of a task. When participants are led to expect a fun game, and still must complete a repetitive cognitive task, they may experience even lower enjoyment than if they expected a boring task [53]. Game elements can also introduce other emotions. For example, Levy et al. [30] found that some older Jewish participants were uncomfortable with their cooking-themed game, as it required making recipes containing pork products.

Do these mixed findings mean researchers should move away from gamifying tasks? Not necessarily—participants might not enjoy assessment games more than a control task, but the data they produce may still be higher-quality.
4.3.2.2 Performance

Groening and Binnewies [36] note that enjoyment is only one way to operationalize motivation—one closely linked to intrinsic motivation. They found that adding achievement-based game elements to a series of simple tasks did not improve self-reported motivation, but did improve persistence—when participants could earn achievements, they engaged with a Stroop task for longer before voluntarily switching tasks, compared to when no achievements were available. Similarly, Mekler et al. [17] found that when they gamified an image annotation task, participants generated significantly more annotations, despite no reported differences in intrinsic motivation or competence need satisfaction when compared to the basic task.

Adding game elements to a task may improve performance (without affecting enjoyment) in various ways. For example, Jung et al. [158] compared the performance of participants who were given a numeric goal (i.e., generate twenty-two ideas) to those who were asked to “do their best”. Participants who were given the specific goal generated higher quantity and higher quality responses. When completing cognitive tasks, participants are often instructed to respond, “as quickly and accurately as possible”. This nebulous goal can be clarified and reinforced through game elements that provide immediate feedback, like scoring points for fast reactions or losing points for incorrect responses.

When designing gamified tasks for research and assessment purposes, it may even be beneficial to focus on influencing performance, rather than enjoyment. Levy et al. [30] note that changes in emotions can influence cognitive abilities, which may interfere with collecting valid and reliable data when using games as scientific tools. When Vanden Abeele et al. [28] compared two games designed to measure psychoacoustic thresholds in preschoolers, they found that the more fully developed and motivating game was able to detect lower thresholds. As another example, Delisle and Braun [39] found that changing a task to resemble a fast-paced videogame actually normalized the performance of participants with ADHD, meaning that participants with and without ADHD performed similarly on the gamified task (but differently on the standard task). In some cases, such an effect may be desired, but it depends on why the task is being used and gamified.
4.3.3 Psychometric Properties of Gamified Tasks

Tasks may also be gamified with the goal of improving the psychometric properties of a task, such as validity (how well a task measures what it claims to measure) and reliability (how consistent the measurement obtained by the task is) [31]. There are also different types of evidence for reliability that need to be considered when gamifying cognitive tasks. Internal consistency refers to how stable the task data is within an assessment; for example, how similar a participant’s reaction time at the beginning of a task is to their reaction time at the end of the task. Test-retest reliability refers to how stable the task data is over time; for example, how similar a participant’s score on a task is at one time point compared to their score on the task a month later.

Typical cognitive tasks are boring, repetitive, and long partly because of the issue of reliability. From one trial to the next, people will perform quite differently, so multiple trials are needed to decrease measurement noise [201]. Adding game elements to a task may change how reliable its measures are. Participants may be engaged enough that their performance is more stable over time; for example, perhaps only twenty trials are needed for a reliable measure, instead of two hundred. Friehs et al. [80] found that response variability in a gamified Stop Signal task was lower compared to the non-game version. Shorter tasks would require fewer resources to administer and reduce the burden on participants, which would be particularly beneficial for clinical and pediatric populations.

Game elements also offer the ability to guide participant performance. Most cognitive tasks use measures of reaction time and accuracy, which leads to classic speed-accuracy trade-offs. The faster a participant responds, the less accurate they will be, and vice versa. Individual participants will also favor speed or accuracy differently than one another [201]. By encouraging participants to emphasize speed, for example through game elements like feedback and points for faster responses, this problem may be addressed [157,201].
4.3.4 This Study

Few gamified cognitive tasks are subjected to rigorous examination of psychometric properties [199], despite their use in experimental and clinical settings. Parsons, Kruijt and Fox [202] note that psychology lacks a standard practice of reporting the reliability of cognitive task measurements. This problem is exacerbated when tasks are adapted, like they are through gamification. Even small manipulations to cognitive tasks require extensive research to understand their effects [32].

With this study, we seek to research how game elements can affect the reliability of scores on a cognitive task; specifically, the Stroop task. As a typical cognitive task that demonstrates robust experimental effects in general populations [3], the Stroop task is well-suited for this research.

4.3.4.1 The Stroop Task

Building on Cattell’s 1886 work with cognitive tasks [4], in 1935, John Ridley Stroop conducted an experiment where he asked participants to either name the colours of coloured rectangles or name the colours of mismatched words (for example, the word ‘blue’ printed in red ink). Participants responded much slower when naming incongruent coloured words, a paradigm we now call the Stroop effect [5].

Since Stroop’s first experiment and subsequent development of the experimental protocol [6–8], the Stroop task has become one of the most widely used tasks in both cognitive and clinical psychology [3,203]. Recently, the Stroop task has been gamified for both experimental and clinical use. For example, Groening and Binnewies [204] used the Stroop task to investigate the effects of game elements on participant motivation and performance. They found that when points and story elements were added to the task, participants were more persistent (they engaged with the task for longer before switching to a new task) and reported higher motivation. Gomez-Tello et al. [205] used gamified tasks as part of a battery of tests for the neuropsychological screening of children, and found evidence of the Stroop effect in a gamified version of the task. However, previous work has not considered reliability of the Stroop effect in
a gamified task, either in terms of internal consistency or test-retest reliability. Thus, we have little guidance for when gamified tasks can or should not be used in assessment.

We created two versions of the Stroop task, with and without game elements, and then tested each task with participants at two time points. In this paper, we report on the reliability of the gamified Stroop task, in terms of internal consistency and test-retest reliability, compared to the control task. We also compare the reliability of scores on a trial-by-trial basis. Our objective is to demonstrate how game elements can affect the reliability of scores on a Stroop task.

4.4 Methods

4.4.1 Tasks

The control task was designed using the basic computerized Stroop task described by Macleod [3] and Hedge, Powell and Sumner [206] as models. Participants were shown words in the middle of their screen, in various colours (red, blue, green, or yellow). The word could be the same as the font colour (congruent condition), a non-colour word (lot, ship, cross, or advice) (neutral condition), or a non-matching color word (e.g., the word ‘blue’ shown in green) (incongruent condition). After each word, participants were asked to press a key corresponding to the font colour (z key for red, x key for blue, n key for green, and m key for yellow). Participants first completed a training exercise to learn each key mapping. The task then consisted of 240 trials in each condition (congruent, neutral, and incongruent), for a total of 720 trials.

The gamified version was designed to increase reliability by manipulating the speed-accuracy trade-off [201] and improving engagement through game elements. Based on prior research, which demonstrated increased enjoyment from points and decreased enjoyment from theme added to a gamified task [53], we focused on adding points-based game elements to the Stroop task. Points-based elements also target extrinsic motivation (rather than intrinsic), which may be more effective at influencing participant performance [36]. We followed the Feedback category of the Gameful Design Heuristics, from Tondello, Kappen, Ganaba and Nacke, which states that
the system should offer users clear and immediate feedback, actionable feedback, and graspable progress [207].

Using feedback also allowed us to manipulate the speed-accuracy trade-off, by preferentially awarding points for faster (but still correct) answers. In the game version of our task, participants saw their response time for each trial and if they answered correctly. A record of their fastest response time was also displayed in the corner of the screen. They lost 5 points for any incorrect answer, gained 5 points for any correct answer, and were rewarded with a bonus 25 points for responses that broke their previous ‘fastest time’ record. A progress bar at the bottom of the screen kept track of the points (see Figure 5).

![Figure 5. Game version of the task, after a correct response was entered.](image)

4.4.2 Participants

Participants were recruited through Prolific, an online platform for recruiting research participants. Online platforms are commonly used in human-computer interaction research to conduct studies [183], and have been shown to yield reliable data when precautionary methods around data gathering and analysis are used (e.g., [184,185]). Each participant completed either the control task or the gamified task, at two time points three weeks apart (Time 1 and Time 2). Participants signed a consent form, were given instructions and training for the task, and then completed the task. After completion, they answered questionnaires collecting demographic
information, including information about their experience of the task (Intrinsic Motivation Inventory (IMI) [208]), their general gaming behaviour, and self-reported attentional control (Attentional Control Scale [209]).

The study design was between-subjects, with half the participants completing the control version of the task, and the other half completing the points version. Participants were randomly assigned to a condition. The study took around 40 minutes to complete, and participants were given £6 in compensation at each time point.

Our analyses were based on methods used by Parsons, Kruijt and Fox [202] and Hedge, Powell and Sumner [206]. Both of these papers used the same datasets, which had data from 47 (Study 1) and 56 (Study 2) participants for the Stroop task. Thus, we aimed to have around 50 participants for each condition.

We only analyzed data from participants who completed both sessions. We also set quality thresholds and removed any participants who did not meet them at either time point. Finally, we also removed outlying data points, such as individual trials that were much slower than average for each participant, to reduce noise in the data.

4.4.3 Statistical Analysis

We conducted two-way ANOVAs with task type (basic/game) and trial condition (congruent/neutral/incongruent) for reaction time and error rate data. We used one-way ANOVAs to compare the effect of task type on the skewness and kurtosis of the distribution of reaction time data for each participant. Additionally, we conducted three-way repeated measures ANOVAs (task type x trial type x time) for reaction time cost and error rate cost data. We also created groups representing low and high attentional control, based on the median (51.0) of our participants, and then conducted three-way repeated measures ANOVAs (task type x attention x time) for reaction time cost and error rate cost data.

For measuring and reporting reliability, our analysis followed the recommendations from Parsons, Kruijt and Fox [202]. To evaluate internal consistency, we used a permutation
approach, which involves repeatedly randomly splitting the data, calculating the reliability estimate, and then averaging all the estimates. This approach provides a more stable estimate, independent of how trial stimuli and conditions are presented [202]. To evaluate test-retest reliability, we calculated the Pearson correlation between each time point. We also used Intraclass Correlation Coefficients (ICCs) to indicate the degrees of consistency and agreement between each time point. Based on Parson’s recommendations, we use the ICCs labelled ICC(3,1) and ICC(2,1) as described by Shrout and Fleiss [210]. Finally, we plotted test-retest reliability as the number of trials increases. To do this, we followed the method used by Hedge, Powell and Sumner [206].

4.5 Results

4.5.1 Participants

For the first round of data collection (Time 1), we received 135 responses, followed by 78 responses for Time 2.

All participants met the criteria for questionnaire speed of completion (participants needed to spend an average of 1.5 seconds per item) and variance (participants needed to show some variance across items). Thirteen participants were removed for too frequently providing an incorrect response on the Stroop task (total incorrect responses > 1 SD above the mean number of incorrect responses) and/or for responding to trials too slowly (mean reaction time > 3SD above the group mean reaction time). Before calculating the group mean reaction time, we also removed any individual trials that were slower than average for each participant (reaction time > 3SD above individual mean reaction time), as well as any remaining outlier trials that were slower than 2000ms. At Time 1, we removed 1,667 trials (out of 50,400). At Time 2, we removed 1,976 trials (out of 49,680). Notably, both at Time 1 and Time 2, significantly fewer trials needed to be removed from the Game condition as compared to the basic version; 38.6% of the removed trials were in the Game condition at Time 1, and at Time 2, 32.9% were in the Game condition.
After exclusions, 65 participants remained (50 women, 13 men, 1 non-binary, 1 prefer not to disclose; mean age = 23.91, SD = 4.64), with 31 participants in the Basic Task condition and 34 participants in the Game condition. Our sample had a high proportion of women due to the online platform we used [211]. Participants had a mean score of 51.8 (SD = 7.54) on the Attentional Control Scale.

4.5.2 Intrinsic Motivation Inventory

At both time points, the Basic Task and Game conditions showed no significant differences for any of the IMI subscales (interest, competence, effort, and pressure).

4.5.3 Reaction Time and Error Rate Data

We averaged reaction times and error rates across participants, and then analyzed each measure by task type and trial condition at each time point. We also calculated the reaction time and error rate costs (mean incongruent trials – mean congruent trials). Descriptive statistics for each measure are shown in Table 4.

Histograms of reaction time for all participants are presented in Figure 6, by task type and time point. One-way ANOVAs revealed no significant effects of task type on the skewness and kurtosis of the distribution of reaction time data for each participant (see Table 5).

The 2-way ANOVAs for reaction time and error rate demonstrated evidence of the Stroop effect at both time points (significant differences between incongruent trials and both congruent and neutral trials). Further, congruence sequence effect analysis revealed the expected adaptive control effect but no effect of task condition, time, or an interaction between the two emerged. There were also significant differences between task conditions at Time 1: participants in the Game condition had faster reaction times and lower error rates than those in the Basic Task condition. There were no significant differences at Time 2. (See Table 6 and Table 7.)

Two-way repeated measures ANOVAs (task type x time) for reaction time cost and error rate cost data showed no significant interaction effects (see Table 8). The three-way repeated
measures ANOVAs (task type x trial condition x time) for reaction time and error rate data showed no significant interaction effects (see Table 8). Based on grouping our participants into low and high attentional control categories, we found a significant three-way interaction between time, task type and attention category for error rate (see Table 8). Participants who scored low in attentional control and were in the Basic Task condition had a lower error rate cost at Time 1 than Time 2. In the Game condition, participants who scored low in attentional control had a higher error rate cost at Time 1 than Time 2. Error rate cost for participants who scored high in attentional control showed the opposite pattern. There were no significant simple two-way interactions of task type and attention category at either time point.
Table 4. Descriptive statistics for reaction time and error rates, at Time 1 and 2 for each task type.

<table>
<thead>
<tr>
<th></th>
<th>Time 1 mean (SD)</th>
<th>Time 2 mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Basic Task</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Congruent reaction time</strong></td>
<td>678 ms (103)</td>
<td>659 ms (104)</td>
</tr>
<tr>
<td><strong>Neutral reaction time</strong></td>
<td>671 ms (94.0)</td>
<td>656 ms (94.7)</td>
</tr>
<tr>
<td><strong>Incongruent reaction time</strong></td>
<td>796 ms (124)</td>
<td>758 ms (118)</td>
</tr>
<tr>
<td><strong>Reaction time cost</strong></td>
<td>118 ms (50.9)</td>
<td>98.8 ms (39.8)</td>
</tr>
<tr>
<td><strong>Congruent % correct</strong></td>
<td>96.0 % (2.86)</td>
<td>96.1 % (2.52)</td>
</tr>
<tr>
<td><strong>Neutral % correct</strong></td>
<td>96.7 % (2.33)</td>
<td>96.8 % (2.43)</td>
</tr>
<tr>
<td><strong>Incongruent % correct</strong></td>
<td>93.1 % (5.46)</td>
<td>93.6 % (4.36)</td>
</tr>
<tr>
<td><strong>Error rate cost</strong></td>
<td>2.86 % (4.53)</td>
<td>2.55 % (3.23)</td>
</tr>
<tr>
<td><strong>Game Task</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Congruent reaction time</strong></td>
<td>638 ms (94.5)</td>
<td>645 ms (95.3)</td>
</tr>
<tr>
<td><strong>Neutral reaction time</strong></td>
<td>628 ms (84.1)</td>
<td>631 ms (79.1)</td>
</tr>
<tr>
<td><strong>Incongruent reaction time</strong></td>
<td>753 ms (112)</td>
<td>730 ms (103)</td>
</tr>
<tr>
<td><strong>Reaction time cost</strong></td>
<td>115 ms (48.8)</td>
<td>85.3 ms (42.3)</td>
</tr>
<tr>
<td><strong>Congruent % correct</strong></td>
<td>94.6 % (3.70)</td>
<td>95.5 % (2.50)</td>
</tr>
<tr>
<td><strong>Neutral % correct</strong></td>
<td>96.0 % (2.53)</td>
<td>96.0 % (2.79)</td>
</tr>
<tr>
<td><strong>Incongruent % correct</strong></td>
<td>92.1 % (3.90)</td>
<td>93.0 % (4.80)</td>
</tr>
<tr>
<td><strong>Error rate cost</strong></td>
<td>2.52 % (4.71)</td>
<td>2.53 % (4.18)</td>
</tr>
</tbody>
</table>
Figure 6. Histograms of reaction time by time point and task type, for each type of trial condition.

Table 5. ANOVA Summary Table for Reaction Time Distribution

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>1</td>
<td>.316</td>
<td>1.863</td>
<td>.177</td>
<td>.029</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>1</td>
<td>.003</td>
<td>.001</td>
<td>.978</td>
<td>.000</td>
</tr>
<tr>
<td><strong>Time 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skewness</td>
<td>1</td>
<td>.317</td>
<td>1.852</td>
<td>.178</td>
<td>.029</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>1</td>
<td>1.159</td>
<td>.358</td>
<td>.552</td>
<td>.006</td>
</tr>
</tbody>
</table>
Table 6. ANOVA Summary Table for Reaction Time

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Type</td>
<td>1</td>
<td>85185.015</td>
<td>8.107</td>
<td>.005</td>
<td>.041</td>
</tr>
<tr>
<td>Condition</td>
<td>2</td>
<td>317396.780</td>
<td>30.205</td>
<td>&lt;.001</td>
<td>.242</td>
</tr>
<tr>
<td>Task Type x Condition</td>
<td>2</td>
<td>49.167</td>
<td>.005</td>
<td>.995</td>
<td>.000</td>
</tr>
<tr>
<td><strong>Time 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Type</td>
<td>1</td>
<td>23700.032</td>
<td>2.402</td>
<td>.123</td>
<td>.013</td>
</tr>
<tr>
<td>Condition</td>
<td>2</td>
<td>201201.515</td>
<td>20.394</td>
<td>&lt;.001</td>
<td>.178</td>
</tr>
<tr>
<td>Task Type x Condition</td>
<td>2</td>
<td>788.555</td>
<td>.080</td>
<td>.923</td>
<td>.001</td>
</tr>
</tbody>
</table>

Table 7. ANOVA Summary Table for Error Rate

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Type</td>
<td>1</td>
<td>.005</td>
<td>4.012</td>
<td>.047</td>
<td>.021</td>
</tr>
<tr>
<td>Condition</td>
<td>2</td>
<td>.024</td>
<td>18.301</td>
<td>&lt;.001</td>
<td>.162</td>
</tr>
<tr>
<td>Task Type x Condition</td>
<td>2</td>
<td>.000</td>
<td>.148</td>
<td>.862</td>
<td>.002</td>
</tr>
<tr>
<td><strong>Time 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Type</td>
<td>1</td>
<td>.002</td>
<td>1.945</td>
<td>.165</td>
<td>.010</td>
</tr>
<tr>
<td>Condition</td>
<td>2</td>
<td>.018</td>
<td>15.402</td>
<td>&lt;.001</td>
<td>.140</td>
</tr>
<tr>
<td>Task Type x Condition</td>
<td>2</td>
<td>.010</td>
<td>.022</td>
<td>.978</td>
<td>.000</td>
</tr>
</tbody>
</table>
Table 8. Repeated Measures ANOVA Summary Table for Reaction Time and Error Rate

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reaction Time Cost</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reaction</td>
<td>1</td>
<td>880.934</td>
<td>1.105</td>
<td>.297</td>
<td>.017</td>
</tr>
<tr>
<td>Effect Size</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Type x Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Reaction Time</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial Type x Task</td>
<td>2</td>
<td>317.106</td>
<td>.616</td>
<td>.491</td>
<td>.010</td>
</tr>
<tr>
<td>Type x Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention x Task</td>
<td>1</td>
<td>1325.711</td>
<td>1.665</td>
<td>.202</td>
<td>.012</td>
</tr>
<tr>
<td>Type x Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Error Rate Cost</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Type x Time</td>
<td>1</td>
<td>&lt;.001</td>
<td>.106</td>
<td>.746</td>
<td>.002</td>
</tr>
<tr>
<td><strong>Error Rate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trial Type x Task</td>
<td>2</td>
<td>.000</td>
<td>.615</td>
<td>.541</td>
<td>.010</td>
</tr>
<tr>
<td>Type x Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention x Task</td>
<td>1</td>
<td>39.218</td>
<td>5.493</td>
<td>.022</td>
<td>.083</td>
</tr>
</tbody>
</table>

a due to the interaction violating the assumption of sphericity (p<.001), p values are derived using the Greenhouse-Geisser statistic

4.5.4 Internal Consistency

We estimated the internal consistency of the basic task using a permutation-based split-half approach [202] with 5000 random splits. Internal consistency ranges between 0 and 1, with higher numbers representing more consistency across an individual’s complete set of trials.

4.5.4.1 Reaction Time

When using reaction time cost, the (Spearman-Brown corrected) split-half internal consistency for the Basic Task at Time 1 was $r_{SB} = 0.78$, 95\% confidence interval (CI) [0.64,0.89]. At Time 2, it was $r_{SB} = 0.64$, 95\% CI [0.40,0.81].

For the Game condition at Time 1, the split-half internal consistency was $r_{SB} = 0.83$, 95\% CI [0.71,0.91]. At Time 2, it was $r_{SB} = 0.76$, 95\% CI [0.60,0.88].
Descriptively, the internal consistency values were higher at both Time 1 and Time 2 for the Game condition (see Figure 7); however, converting the correlations to Fisher’s z-scores indicated no significant differences between groups at each time point.

![Figure 7. Internal consistency of reaction time cost for each time point and task type.](image)

### 4.5.4.2 Error Rate

When using error rate cost, the (Spearman-Brown corrected) splithalf internal consistency for the Basic Task at Time 1 was $r_{SB} = .79$, 95% CI [.66, .89]. At Time 2, it was $r_{SB} = 0.6$, 95% CI [0.34,0.79].

For the Game condition at Time 1, the splithalf internal consistency was $r_{SB} = 0.76$, 95% CI [0.62,0.87]. At Time 2, it was $r_{SB} = 0.74$, 95% CI [0.58,0.86].

Descriptively, the internal consistency values were higher at Time 2 for the Game condition at Time 2 (see Figure 8); however, similar to the reaction time data, converting the correlations to Fisher’s z-scores indicated no significant differences between groups at each time point.
4.5.5 Test-Retest Reliability

4.5.5.1 Reaction Time

Using reaction time cost data, for the Basic Task, the Pearson correlation between each time point indicated a test-retest reliability of .68, 95% CI = [.43, .84]. This correlation was significant, t(29) = 5.04, p < .001. For the Game condition, we found a test-retest reliability of .58, 95% CI = [.31, .77]. This correlation was also significant, t(32) = 4.07, p < .001.

We also estimated test-retest reliability between Time 1 and Time 2 with ICCs using the `psych` package in R [212]. ICCs are used to measure the reliability of a measure between two time points. The value of an ICC can range from 0 to 1, with higher values indicating higher reliability. We report the results of two-way mixed-effects models for absolute agreement, ICC(2,1), and consistency, ICC(3,1).

Using reaction time cost data, for the Basic Task, the estimated agreement was .61, 95% CI = [.36, .78], and the estimated consistency was .66, 95% CI = [.46, .80]. For the Game condition, the estimated agreement was .48, 95% CI = [.16, .69], and the estimated consistency was .58, 95% CI = [.35, .74].
Typically, cognitive tasks need many trials to reduce measurement noise. We plotted how ICC(3,1) changes as the number of trials increases, to see if a more stable estimate could be determined with fewer trials when using game elements. Figure 9 shows how the reliability of the Stroop effect (reaction time cost) changes with an increasing number of trials.

Figure 9. Test-retest reliability of reaction time cost as the number of trials increases, for each task type.

To investigate why the Game condition shows lower test-retest reliability, we also plotted how the reliability of reaction time changes over time for each trial type (neutral, congruent, and incongruent trials) (see Figure 10). Comparing the plots suggests that the Game condition reaches a higher level of consistency sooner for incongruent trials, compared to both the neutral and congruent conditions. The Basic Task shows similar patterns of consistency across all the trial types.
Figure 10. Test-retest reliability of reaction time as the number of trials increases, for each trial type and task type.

4.5.5.2 Error Rate

Using error rate cost data, for the Basic Task, the Pearson correlation between each time point indicated a test-retest reliability of 0.55, 95% CI = [0.24, 0.76]. This correlation was significant, \(t(29) = 3.56, p = .001\). For the Game condition, we found a test-retest reliability of 0.62, 95% CI = [0.35, 0.79]. This correlation was also significant, \(t(32) = 4.45, p < .001\).

Using error rate cost data, for the Basic Task, ICC(2,1) (estimated agreement) was .53, 95% CI = [.28, .71], and ICC(3,1) (estimated consistency) was .53, 95% CI = [.28, .71]. For the Game condition, ICC(2,1) was .62, 95% CI = [.42, .77], and ICC(3,1) was .62, 95% CI = [.41, .77].

Again, we plotted how ICC(3,1) changes as the number of trials increases, to see if a more stable estimate could be determined with fewer trials when using game elements. Figure 11 shows how the reliability of the Stroop effect using error rate cost changes with an increasing number of trials.
Like with reaction time, we also plotted how the reliability of the number of errors changes over time for each trial type (neutral, congruent, and incongruent trials) (see Figure 12). The Basic Task shows similar patterns of consistency across all the trial types, whereas in the Game condition, only the neutral and congruent conditions are similar—the reliability of the incongruent trials keeps increasing over time.

**Figure 11. Test-retest reliability of error rate cost as the number of trials increases, for each task type.**
4.6 Discussion

4.6.1 Summary and Explanation of Findings

4.6.1.1 Performance

Both versions of the task demonstrated the Stroop effect, meaning that the effect is robust to the addition of certain game elements. Gamification can affect the validity of cognitive tasks; for example, adding graphics (especially those that change the stimuli participants respond to) can worsen performance compared to a control task [19,160,198]. For the current study, in the Game condition, reaction times and a progress bar were perpetually displayed on the screen. Graphics indicating gained or lost points also appeared between stimuli. These elements did not interfere with the validity of the Stroop task.

There were no significant differences in performance-based measures between the Basic Task and Game conditions, with one exception: Participants in the Game condition had significantly
faster reaction times and lower error rates than those in the Basic Task condition, but only at Time 1. There may be several reasons for these results.

Points, functioning as extrinsic motivators, have been shown to improve performance on cognitive tasks [17]; however, this effect may be short-lived. Nicholson [14] notes that reward-based game elements can drive immediate spikes in engagement, but only so long as continuous rewards are provided. In our Game condition, participants were continually awarded points for accurate responses; however, for reaction time, they were only awarded bonus points for responses that broke their previous ‘fastest time’ record. There is a physical limitation to how quickly participants can react to stimuli—once that threshold is met, it will be near-impossible to improve further, and the motivating influence of the bonus points may be diminished.

In the Game condition, participants may also learn faster and reach their ‘peak performance’ sooner. Participants are quickly incentivized to put forth their best effort. This effect may be particularly pronounced when the cognitive demands of the task are higher. When we plotted the reliability of reaction time and error rate as the number of trials increased, the incongruent trials showed an improved pattern of consistency only in the Game condition. Specifically, after about 50-100 trials the reaction time stays consistent in the game, whereas there is significant variation in the basic version, with a noticeable drop after 50 trials. A similar pattern emerges for error rates. For the Basic Task, the plots of all three trial types showed similar patterns across both performance measures. This is especially noteworthy because incongruent trials are arguably the most important trials in the Stroop task, as they are the trials wherein cognitive conflict needs to be resolved. Improved performance on the incongruent trials also explains why the reliability of the Stroop effect (reaction time cost) appeared lower in the Game condition—participants in that condition performed better and more consistently in the incongruent trials.

The differences between the Basic Task and Game conditions may be emphasized by the incongruent trials because they are more cognitively demanding than the congruent and neutral trials. There is evidence to suggest that game elements can differentially affect cognition, depending on how participants experience the demands of the task. For example, using gamification can normalize the performance of participants with ADHD [39].
Another indication of improved performance consistency comes in the form of a significantly smaller number of outlier trials needing to be removed from the game condition as compared to the basic version. Approximately twice as many far-out outlier trials were removed from the basic task. These trials are not considered valuable data and were essentially lost time for both researcher as well as participant. By reducing the number of trials needed to be removed from performances, the time-investment for participants is reduced. Further, this means that the previous results are a conservative estimate of the game’s reliability advantage because the most egregious outliers were already removed from analysis.

4.6.1.2 Enjoyment

There were also no differences in self-reported measures of motivation between the Basic Task and Game conditions. These results align with other studies, which have found points and achievement-based game elements are only effective in promoting performance, and not motivation [17,36].

Levy et al. [30] note how carefully games must be designed in order to appropriately function as scientific tools, and highlight the importance of using the research and data collection goals to inform the choice of game design. For our current study, we specifically chose game elements that we thought would influence performance, rather than enjoyment. Gamified tasks may be more successful if the game elements are just “good enough” to achieve the goals of the study, without interfering with the validity of the task [30]. Because we wanted to improve participant performance, irrespective of enjoyment, we did not add extraneous game elements, even if those elements would have made the game more fun.

4.6.2 Limitations and Future Work

One limitation to our study was the small sample size. Our two task conditions were designed with fairly subtle differences, in the form of points and feedback. While this design was intentional, we also had a relatively small sample size, which may not have been powerful enough to reveal small effects of our slight manipulations. We recruited 135 participants for Time 1, with the intent of having at least 50 participants per condition. However, only 78
participants returned for Time 2. Incentivizing participants to return to an online study is difficult. Future studies may find significant effects with a larger sample.

Another limitation is that our sample was also heavily skewed towards young adult female participants. We recruited our participants through an online platform, Prolific. At the time of our study, a young woman made a video describing her side hustle as a participant on the platform. Her video went viral on TikTok, resulting in an influx of new signups to Prolific, most of whom were, like the creator, females in their 20s [211]. However, given the fundamental nature of the research, this sampling bias is unlikely to have influenced the results.

Adding points and feedback is one simple approach to gamification. Other game elements may produce different results. As discussed, we had theoretical and practical reasons for using points, but even within the category of points and achievement-based game elements, we could have made different design and mechanic choices. For example, adding a leaderboard system may have influenced participant behaviour, due to the increased competition. Other game elements should be investigated in future work.

4.6.3 Implications

With this study, we show that the Stroop effect is robust to the addition of simple points-based game elements. Adding points to a Stroop task does initially increase participant reaction time, but this gamification may be most effective in the short term. Our results also suggest that game elements may differently influence parts of a cognitive task, like the more cognitively demanding incongruent trials.

We also provide an example of reporting psychometric data for a gamified task. Despite a long history of cognitive task gamification, the field lacks standard practices around how these tasks are made and measured [199]. Any advancement in how these tasks are designed and used requires a stronger base of knowledge of how individual game elements affect cognitive-behavioural measures [17,202]. One of the most cited reasons for gamifying tasks is to address the limitations of standard neuropsychological testing [199]; however, these games will never be
acceptable replacements for traditional tests if they are not subjected to the same rigorous standards of reliability and validity.

The present results hint towards a potential advantage of using game-like tasks to assess cognitive functioning, especially for difficult to reach populations or individuals that cannot be subjected to prolonged testing. For example, gamified tasks have been shown to provide a more engaging environment which creates a more captivating setting that may aid in collecting data from populations with a lower attention span, such as children or groups of patients with concentration or attention deficits [26].

Our results suggest that the Game condition may provide a faster onboarding to true performance and improved consistency, as demonstrated descriptively through the lower proportion of outlier trials removed, the reaction time distributions, the split-half internal consistency values for reaction time and error rate, and reaction time cost by trial number charts. This faster onboarding is also supported by the significantly faster reaction times and lower error rates in the Game condition at Time 1. However, these trends do not result in significant performance differences between the Basic Task and Game conditions in analyses of reaction time cost, and also do not influence test-retest reliabilities, suggesting that the game elements we included neither significantly improve—nor compromise—performance in a gamified Stroop task.

4.7 Contribution to this Dissertation

In Manuscript C, we explored a novel approach to cognitive task gamification. In Manuscript A, we found that many studies report using gamification to increase user enjoyment. However, in Manuscript B, we found that gamification can affect user enjoyment in unintended ways—adding game elements does not always equate to a better experience for the user. This study prompted the exploration of other strategies, apart from increased enjoyment, to improve data quality. Thus, in Manuscript C we explored a novel approach to cognitive task gamification, where our goal was to influence user performance, rather than user enjoyment.
5 Overall Discussion

5.1 Review of the Work in this Dissertation

In this section, we will summarize each manuscript included in this dissertation.

5.1.1 Summary of Manuscript A

In Manuscript A, we conducted a systematic review to answer the question: How are digital games used for the cognitive assessment of attention made and measured? We searched for papers that described a digital game used to assess attention that could be deployed remotely without specialized hardware, and ultimately identified 74 papers for inclusion in our review. We found three approaches to making game-based assessments: gamifying cognitive tasks, creating custom games based on theories of cognition, and exploring potential assessment properties of commercial games. We also looked at how each study evaluated the assessment properties of the games (e.g., how accurately they assess attention) and found three approaches: comparison to traditional cognitive tasks, comparison to clinical diagnoses, and comparisons to knowledge of cognition. Most studies did not evaluate the game in regard to participant enjoyment.

Based on our review, we identified three barriers to advancing the use of game-based assessments. The first barrier we identified is that the literature currently perpetuates assumptions about how users interact with assessment games. Second, there is a lack of evaluation of these games. Third, there is not a clear standard of integration across the field. These barriers point to significant limitations in our understanding of how game elements affect the data quality of assessments. For example, 40 out of 74 papers in the review used game elements with the intent of improving participant experience (and thus data quality), but only 25 of these studies evaluated the resulting experience, and often with simple yes/no questions. Without robust evaluation practices, we cannot conclude that games do improve how participants experience these assessments.

We also identified some ‘best practices’ from the papers in our review: First, describing a clearly stated motivation for using a game over a traditional assessment can justify the use of games in this context. Clearly articulating a clear long-term goal can set guideposts for the development
and evaluation of games for assessment. Second, evaluating the assessment capacity of assessment games is necessary for establishing their validity. Measuring the player experience within assessment games is necessary for justifying the use of games over traditional assessment tasks. Third and finally, integrating knowledge across the disciplines of psychology, clinical sciences, game design, or user experience will help to ensure robust results.

This review provided an overview of the field for my dissertation and pointed to a significant area for further research: understanding how game elements affect data quality. It also emphasized the importance of evaluating both the assessment properties and game properties in our own research.

5.1.2 Summary of Manuscript B

In Manuscript B, we applied two popular gamification approaches to a typical cognitive task (the dot probe). We used points/feedback and theme/narrative, both individually and in combination. In Study 1, we found that points increase reaction time and error rate, and reported positive affect. Interestingly, we did not find the same increase in positive affect associated with the use of theme/narrative-based game elements. To investigate this effect, we conducted Study 2.

In Study 2, we replicated our results from Study 1 and expanded our analysis to further investigate participant experience, through additional measures of participant enjoyment, taken before, during and after the task. Our findings suggest that using thematic elements creates expectations of an interesting game, which gamified tasks fail to deliver, whereas points maintain enjoyment better throughout the task itself.

5.1.3 Summary of Manuscript C

In Manuscript C, we researched how game elements can affect the reliability of scores on a Stroop task. We specifically investigated within and across session performance consistency. We created two versions of the Stroop task, with and without game elements, and then tested each task with participants at two time points. In the gamified version of the task, we added points and feedback-based game elements.
Ultimately, we did not find significant performance differences between the traditional cognitive task and the gamified version, in analyses of reaction time cost and test-retest reliabilities. However, descriptive trends in the data suggest that the gamified task may provide a faster onboarding to true performance and improved consistency, and warrants further research.

5.2 Methodology

In this section, we discuss the methods we used in our studies and the lessons we learned.

5.2.1 Online Experiments

We conducted all our experiments online. We used Amazon Mechanical Turk (MTurk) for the studies conducted in Manuscript B, and Prolific for Manuscript C. At the time of conducting Manuscript C, we chose to switch from MTurk to Prolific in hopes of reducing the number of bot responses. Studies have since shown that compared to MTurk, Prolific yields higher quality data [213,214].

Online studies have both advantages and limitations when compared to in-person methods [215,216]. HCI research frequently uses online data collection methods. Online studies allow for large subject pools of potential participants. These subject pools are also diverse. If we conducted our research in-person, our participants would have been recruited via university resources, and therefore largely undergraduate students at the University of Saskatchewan. Online participant pools are typically far more diverse in age, ethnicity, gender, and socioeconomic status.

We should note that our study in Manuscript C deviated from this pattern—our participants were largely young females, despite being recruited online via Prolific. Coincidentally, when we ran our experiment, a young woman made a video describing her side hustle as a participant on the platform. Her video went viral on TikTok, with over four million views. The video resulted in an influx of new signups to Prolific, most of whom were, like the creator, females in their 20s [211]. Since then, Prolific has implemented features that allow for easier balancing of participants across demographic features. They also offer a representative sample feature, which recruits a sample reflecting either the US or UK populations.
One criticism of online experiments is that participants are not directly observed, and it can be difficult to know if they are paying sufficient attention to the task at hand. This issue is especially relevant for cognitive task research, which depends on performance metrics and assumes that participants are not multi-tasking. We filtered our data in an attempt to remove participants who may have been distracted or not exhibiting their best efforts. For example, in Manuscript C we used the following process for data cleaning:

1. We removed participants who completed questionnaires too quickly (total time was faster than 1.5 seconds per item)
2. We removed participants who showed zero variance within the questionnaires (e.g., answered every question with ‘disagree’)
3. We removed participants who had too many incorrect responses to task trials, compared to other participants (total incorrect responses > 1 SD above the mean number of incorrect responses)
4. We removed participants who responded to task trials too slowly, compared to other participants (mean reaction time > 3SD above the group mean reaction time.
5. Before calculating the group mean reaction time, we also removed any individual trials that were slower than average for each participant (reaction time > 3SD above individual mean reaction time), as well as any remaining outlier trials that were slower than 2000ms.

It should be noted that choices around data cleaning may affect the results of any analyses. In Manuscript C, we found that we removed around thirty percent fewer trials from the game condition than the basic task, which in itself is an interesting result.

Overall, online studies have been shown to yield reliable data when precautionary methods around data gathering and analysis are used (e.g., [184,185]).

As well, our systematic review found that the top reason researchers cite for using gamified tasks is to address limitations of traditional tasks. These limitations include capturing large-scale datasets. Learning how to conduct studies online and produce adequate data is part of addressing such limitations.
5.2.2 Interdisciplinary Work

Our work drew upon methods used in the fields of HCI and psychology. In our systematic review, we were careful to query datasets across multiple disciplines, such as the ACM Digital Library and PsychInfo (as well as broad datasets like Scopus and Web of Science). We found research published in the fields of psychology, medicine, computer science, education, and interdisciplinary journals. While an exploration of how papers in different journals differed from one another was outside the scope of our systematic analysis, we did notice some trends. Very few papers evaluated the game aspect of the tasks (e.g., how enjoyable users found the game), but papers published in computer science journals were more likely to include such an evaluation. Of the papers from computer science journals, 54% evaluated the game, compared to 35% of interdisciplinary papers and 27% of psychology and medical papers. While improvements could be made across all fields, we can look to the fields of computer science and HCI as examples of how to evaluate the user experience.

In our studies, we learned the value of thoroughly investigating the user experience. In the first study of Manuscript B, we gave participants the PANAS questionnaire before and after the tasks. The results initially seemed counter-intuitive, as we saw player affect increase with the inclusion of points, but not with the inclusion of thematic elements. We then needed to conduct the second study to understand these results. By more thoroughly investigating player enjoyment through additional questionnaires (an enjoyment questionnaire given before, during and after the tasks, as well as the PXI and free-form comments), we were able to show that participants likely experienced heightened expectations created by the introductory exposition, and then disappointment when the game play failed to meet those expectations.

In Manuscript C, we drew upon methods from psychology to investigate the psychometric properties of a gamified task more thoroughly. Very few studies in our systematic review reported on the reliability of their tasks, despite the most cited reason for gamifying tasks being to address the limitations of standard neuropsychological testing. Games will not be able to be used in the same way as traditional tests unless they meet the same rigorous standards of reliability and validity.
As an inherently interdisciplinary field, HCI is well-positioned to meet the challenges of developing and evaluating gamified tasks. Levy et al. [30] offer guidelines for developing games that function as scientific tools. They emphasize the importance of trans-disciplinary co-creation, noting that “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”. Vanden Abeele et al. [28] discuss the benefits of this kind of collaboration. They partnered with researchers in psychoacoustics, who had previously created a game-based assessment for children. By applying their knowledge of game design and using a player-centered design process, they were able to create a new assessment. This new game was preferred by participants, better captured their attention, and even detected lower psychoacoustic thresholds.

In our studies, we tried to match the methods we used with our research goals. For example, in Manuscript B we focused heavily on the game design, as our intent was to investigate the effects of individual game elements. In Manuscript C, our focus was on measuring the reliability of a “just good enough” gamified task.

5.3 Contribution of this Dissertation

This dissertation asked the overall question: **How do games for the assessment of attention affect the quality of gathered data, in terms of both the experience of participants and their performance?** To investigate this issue, our three manuscripts each focused on these questions, respectively:

1. How are digital games used for the cognitive assessment of attention made and measured?
2. How do game elements affect performance on, and enjoyment of, an attention-based cognitive task?
3. How do game elements affect the reliability of an attention-based cognitive task?

In our review of the literature, we found repeated calls for more systematic research on game-based assessments, such as investigating the effects of individual game elements [17]. We noted a lack of integration in the field. There is little guidance on how assessment games should be made, evaluated, and used. Every project will be different (e.g., will use different tasks, themes, or gamification approaches), but for the field to advance there needs to be integration at the level
of game structure. In referring to the structural level, we mean that researchers should develop a clear understanding of how different game mechanics in assessment games (e.g., points, theme, rewards, feedback, procedures, rules, game input, narrative elements) interact with user performance and experience.

Thus, we designed the studies in Manuscript B to examine two popular gamification approaches (points/feedback and theme/narrative) at this structural level, particularly regarding user experience. Our results found that points were associated with greater enjoyment across the duration of the task, and also a speed-accuracy trade-off in terms of reaction time and error rate. Based on these results, we designed Manuscript C to further examine the effects of points on data quality.

Our studies add to the structural building blocks of knowledge necessary to advance the use of game-based assessments. We also offer models for how to design studies that adequately measure both user experience and performance, and ultimately the data quality of these tasks.

5.4 Limitations and Future Work

In addition to the limitations discussed in each manuscript, our work was limited in its overall scope. We will discuss these limitations and present ideas for future work.

In Manuscript A, we identified three approaches to developing game-based assessments: gamifying cognitive tasks, creating custom games based on theories of cognition, and exploring potential assessment properties of commercial games. Our subsequent experimental studies only examined gamified cognitive tasks. This approach was useful as it allowed us to make direct comparisons between established tasks and our gamified versions. We could easily isolate the effects of game elements on user experience and performance. However, other approaches may have different effects and should be studied as well. In particular, we found mixed results for the effect of gamification on enjoyment. Other approaches, like using commercial games, likely have better potential for maximizing enjoyment.
Similarly, the results of Manuscript B presented us with a choice. We found that using thematic elements creates expectations of an interesting game, which gamified tasks fail to deliver, whereas points maintain enjoyment better throughout the task itself. From here, the choice was to further explore meaningful gamification and try to create an experience that met players’ expectations, or to further explore the benefits and limitations of points-based game elements. Ultimately, for Manuscript C we chose the latter. Other studies have shown the efficacy of meaningful gamification [28], but studies that thoroughly report on the reliability of cognitive tasks are rare, even within the field of psychology [202]. However, future work should expand these types of analyses to all approaches of developing game-based assessments.

In both Manuscript B and Manuscript C, we created original gamified tasks. By developing our own games, we were able to specify and isolate the elements we wanted to study. However, replication is a crucial part of developing scientific tools. Traditional cognitive tasks are extensively studied and then standardized. Especially when used for assessment, these tasks must be administered in a consistent way. For example, Macleod [3] lays out a set of standard ‘ingredients’ for a Stroop experiment—consistent stimuli (colours and words), with random and equal presentation of colour-word combinations, presented in a typical way (duration, centered on the screen, in lower case, etc.) and so forth. A Stroop experiment is instantly recognizable and easily compared to other Stroop experiments. There is much less consistency with gamified tasks. Two gamified Stroop tasks may use points and feedback as game elements but present those elements in completely different ways, making comparative work and meta-analyses difficult. As the field moves forward, it may be useful to begin building on previous work in more substantial ways, and start developing more standardized presentations of game elements.
6 Conclusion

Cognitive tasks have many uses as scientific tools, from helping us understand how the human mind works to informing mental health assessments. But these tools are not perfect—they face problems of attrition for research and with patient cooperation for clinical use [19]. Participants often exert suboptimal effort, leading to poor-quality data [20,21]. When cognitive tasks are gamified, users may find the previously boring tasks become interesting, and feel more willing to engage with the tasks and motivated to put forth their best effort.

However, there are still many unanswered questions about how game-based assessments work. In this dissertation, we ask the question: **How do games for the assessment of attention affect the quality of gathered data, in terms of both the experience of participants and their performance?**

We systematically reviewed the literature to identify how game-based assessments of attention are made and measured. Our results provided an overview of the field, further research questions, and a guide for how to approach future studies. We found that many assessment games are not sufficiently evaluated to determine the quality of data they produce.

We then conducted two studies to examine how game elements affect performance on and enjoyment of an attention-based cognitive task. These studies provided guidance on which game elements are most likely to improve data quality, in terms of increasing participant enjoyment (and therefore engagement with the task) and performance (how quickly and accurately they respond to the task). Finally, we examined how the game elements identified in the previous studies affect the reliability of an attention-based cognitive task.

To advance the use of game-based assessment, we need robust literature that examines individual game elements and their effects on all aspects of the assessment—the user experience, the behavioural performance of users, and the psychometric properties of the tasks. Our studies add to these structural building blocks of knowledge. We also offer models for how to design studies that adequately measure both user experience and performance, and ultimately the data quality of these tasks.
7 References

4. Cattell JM. The time it takes to see and name objects. Mind. 1886;os-XI(41).


24. Sea Hero Quest - Gaming to Help Study Dementia [Internet].


43. Eysenck MW, Derakshan N. New perspectives in attentional control theory. Pers Individ


56. Lumsden J, Edwards EA, Lawrence NS, Coyle D, Munafò MR. Gamification of Cognitive


75. Szalma JL, Schmidt TN, Teo GWL, Hancock PA. Vigilance on the move: Video game-based measurement of sustained attention. Alluisi Becker, Boot, Caggiano, Cain, Carter, Davies, Deaton, Drury, Feng, Finomore, Fitts, Funke, Gibson, Gibson, Gill, Granic, Green, Grubb, Gunn, Hancock, Hancock, Hancock, Hancock, Hancock, Hancock, Hancock, Hancock, Hancock, Hart, A, editor. Ergonomics. Szalma, J.L., James.Szalma@ucf.edu: Taylor & Francis; 2014;57(9):1315–1336.


80. Friehs MA, Dechant M, Vedress S, Frings C, Mandryk RL. Effective gamification of the stop-signal task: Two controlled laboratory experiments. JMIR Serious Games. 2020;8(3).


of learners’ game interactions to identify their dyslexia level. 2020 IEEE 20th Int Conf

85. Song H, Yi D-J, Park H-J. Validation of a mobile game-based assessment of cognitive

Characterizing cognitive control abilities in children with 16p11.2 deletion using adaptive

What cognitive abilities are tapped by casual video games? Ackerman Ball, Basak,
Bennett, Bentler, Boot, Boot, Brehmer, Cattell, Chooi, Colom, Conway, Diehl,
Ekstrom, Fan, Fukuda, Gopher, Gray, Green, Green, Green, Halford, Heitz, Holmes,
Jaeggi, Jaeggi, Jausovec, Kaiser, Kane, Kane, Kirchner, Klingberg, K A, editor. Acta
Psychol (Amst). Baniqued, Pauline L.: Beckman Institute, 405 N. Mathews Ave., Urbana,
IL, US, 61801, banique1@illinois.edu: Elsevier Science; 2013;142(1):74–86.

Cognitive Functions of Aged Patients Using Video Games. Proc 4th Work ICTs Improv
Patients Rehabil Res Tech. New York, NY, USA: Association for Computing Machinery;

al. The use of mobile games to assess cognitive function of elderly with and without
cognitive impairment. Ahmadi-Abhari Costa, Creavin, Dawson, Emmons, Folstein, Hilti,
Johnson, Kim, Knols, Lampit, Miller Jr, Mioshi, Neri, Park, Raz, Ruopp, Tam, Toril, Tsoi,
Anatomy, Biomechanics and Organogenesis (LABO), CP 619, Lennik Street 808,
Brussels, Belgium, 1070, bbonnech@ulb.ac.be: IOS Press; 2018;64(4):1285–1293.

aging platform for evaluating cognitive functions in aging: A comparison with the MoCA
in a normal population. Aalbers Boletsis, Cernich, Charisky, Chaytor, Christiansen,
Ciesielska, Davies, Davison, Faul, Feldstein, Fernandez-Montenegro, Flynn, Fortin,
Gamberini, Ghihm, Ismail, Iverson, Jack, Kang, Logan, Manera, McGee, Nasreddine,
Oliveira, Pazzi, Riva, Rizzo, R A, editor. Front Aging Neurosci. Bottiroli, Sara,


103. Angelim Pereira VF, Sanches Valentin LS. The MentalPlus (R) Digital Game Might Be an Accessible Open Source Tool to Evaluate Cognitive Dysfunction in Heart Failure with Preserved Ejection Fraction in Hypertensive Patients: A Pilot Exploratory Study. Int J Hypertens. 2018;


107. Tong T, Chignell M, Tierney MC, Lee Mark; ORCID: http://orcid.org/0000-0001-8120-6905 JSAI-O http://orcid. org/Chignel. Test-retest reliability of a serious game for delirium screening in the emergency department. Anastasi Bhat, Brouillette, Calamia,


Technology in Behavioral Science; 2019;4(3):234–244.


140. 2019 Essential Facts About the Computer and Video Game Industry [Internet].


159. Przybylski AK, Rigby CS, Ryan RM. A Motivational Model of Video Game Engagement.
166. MacMahon KMA, Broomfield NM, Espie CA. Attention Bias for Sleep-Related Stimuli in Primary Insomnia and Delayed Sleep Phase Syndrome Using the Dot-Probe Task. Sleep. 2006;29(11):1420–1427. PMID: 17162988
171. Schoenmakers TM, de Bruin M, Lux IFM, Goertz AG, Van Kerkhof DHAT, Wiers RW.


181. Pictures of Facial Affect (POFA) - Paul Ekman Group [Internet].


189. Vanden Abeele V, Spiel K, Nacke L, Johnson D. Development and Validation of the Player Experience Inventory: A Scale to Measure Player Experiences at the Level of Functional and Psychosocial Consequences.


195. Notebaert L, Grafton B, Clarke PJ, Rudaizky D, Chen NT, MacLeod C. Emotion-in-
Motion, a Novel Approach for the Modification of Attentional Bias: An Experimental Proof-of-Concept Study. JMIR serious games. JMIR Publications Inc.; 2018;6(4):e10993. PMID: 30487121


211. Charalambides N. We recently went viral on TikTok - here’s what we learned. 2021.


Appendix A. General details of each study in Manuscript A

<table>
<thead>
<tr>
<th>Reference</th>
<th>Year</th>
<th>Field of Journal</th>
<th>Sample Size</th>
<th>Area of Attention</th>
<th>Population Age</th>
<th>Disorder</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>2017</td>
<td>1 = interdisciplinary 2 = psych/medical 3 = computer science 4 = education</td>
<td>1</td>
<td>291</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>2020</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>32</td>
<td>2009</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>33</td>
<td>2019</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>34</td>
<td>2016</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>136</td>
<td>4, 5</td>
</tr>
<tr>
<td>35</td>
<td>2018</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>267</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>2013</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>37</td>
<td>2014</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>28</td>
<td>4</td>
</tr>
<tr>
<td>38</td>
<td>2018</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>13979</td>
<td>2, 4</td>
</tr>
<tr>
<td>39</td>
<td>2014</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>40</td>
<td>2016</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>141</td>
<td>5</td>
</tr>
<tr>
<td>41</td>
<td>2019</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>42</td>
<td>2000</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>30</td>
<td>8</td>
</tr>
<tr>
<td>43</td>
<td>2020</td>
<td>1 = visual 2 = divided 3 = selective 4 = sustained 5 = inhibition 6 = switching 7 = updating 8 = orienting 9 = general/not specified 10 = attentional bias</td>
<td>1</td>
<td></td>
<td>30</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>72</td>
<td>2019</td>
<td>2</td>
<td>75</td>
<td>9</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>73</td>
<td>2013</td>
<td>2</td>
<td>682</td>
<td>8</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>74</td>
<td>2019</td>
<td>2</td>
<td>31</td>
<td>4, 5, 6</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>75</td>
<td>2011</td>
<td>2</td>
<td>54</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>76</td>
<td>2019</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>77</td>
<td>2020</td>
<td>2</td>
<td>81</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>78</td>
<td>2020</td>
<td>2</td>
<td>32</td>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>79</td>
<td>2018</td>
<td>2</td>
<td>76</td>
<td>1, 4, 5, 6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>80</td>
<td>2021</td>
<td>2</td>
<td>37</td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>81</td>
<td>2020</td>
<td>2</td>
<td>55</td>
<td>1, 5, 6</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>82</td>
<td>2020</td>
<td>2</td>
<td>1139</td>
<td>3, 4, 5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>83</td>
<td>2017</td>
<td>2</td>
<td>1086</td>
<td>2, 3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>84</td>
<td>2018</td>
<td>2</td>
<td>25</td>
<td>3, 4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>85</td>
<td>2006</td>
<td>2</td>
<td>292</td>
<td>6</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>86</td>
<td>2001</td>
<td>2</td>
<td>160</td>
<td>6</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>87</td>
<td>2019</td>
<td>2</td>
<td>130</td>
<td>3, 4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>88</td>
<td>2015</td>
<td>2</td>
<td>240</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>89</td>
<td>2012</td>
<td>3</td>
<td>24</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>90</td>
<td>2015</td>
<td>3</td>
<td>27</td>
<td>3, 4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>91</td>
<td>2019</td>
<td>3</td>
<td>64</td>
<td>9</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>92</td>
<td>2019</td>
<td>3</td>
<td>22</td>
<td>1, 4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>93</td>
<td>2019</td>
<td>3</td>
<td>41</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>94</td>
<td>2016</td>
<td>3</td>
<td>60</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>95</td>
<td>2017</td>
<td>3</td>
<td>75</td>
<td>3, 4, 5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>96</td>
<td>2017</td>
<td>3</td>
<td>24</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>97</td>
<td>2018</td>
<td>3</td>
<td>103</td>
<td>9</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>98</td>
<td>2018</td>
<td>3</td>
<td>11</td>
<td>4, 5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>99</td>
<td>2018</td>
<td>3</td>
<td>65</td>
<td>5, 6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Year</td>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>------</td>
<td>-----</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>100</td>
<td>2018</td>
<td>3</td>
<td>178</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>101</td>
<td>2018</td>
<td>3</td>
<td>13</td>
<td>3,4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>102</td>
<td>2012</td>
<td>4</td>
<td>32</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>103</td>
<td>2018</td>
<td>4</td>
<td>247</td>
<td>5,6,7</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>
Appendix B. Assessment and game details of each study in Manuscript A

<table>
<thead>
<tr>
<th>Reference</th>
<th>Assessment Evaluation</th>
<th>Purpose of Game</th>
<th>Game Development</th>
<th>Game Evaluation</th>
</tr>
</thead>
</table>
| 1 = Task comparison  
2 = Clinical comparison  
3 = Theory comparison  
4 = None | 1 = ecological validity  
2 = engage children  
3 = address limitations of traditional tests  
4 = increase motivation  
5 = previous research  
6 = none | 1 = gamification  
2 = custom game based on theory  
3 = commercial game  
4 = other | 1 = yes  
2 = no evaluation  
3 = no formal evaluation |
<p>| 13 | 1 | 4 | 1 | 1 |
| 17 | 1 | 4 | 1 | 1 |
| 32 | 4 | 3 | 2 | 2 |
| 33 | 2 | 3 | 2 | 3 |
| 34 | 2 | 3 | 2 | 2 |
| 35 | 2 | 3 | 2 | 2 |
| 36 | 3 | 2 | 1 | 1 |
| 37 | 3 | 4 | 1 | 1 |
| 38 | 3 | 4 | 1 | 3 |
| 39 | 1 | 3, 4 | 2 | 3 |
| 40 | 1 | 4 | 1 | 1 |
| 41 | 1 | 3, 4 | 2 | 1 |
| 42 | 4 | 2 | 1 | 3 |
| 43 | 1 | 4 | 1 | 1 |
| 44 | 1, 2 | 3 | 1 | 2 |
| 45 | 1, 2 | 3 | 2 | 2 |</p>
<table>
<thead>
<tr>
<th></th>
<th>2, 3</th>
<th>2</th>
<th>2</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>2, 3</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>47</td>
<td>1, 2, 3</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>49</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>51</td>
<td>2</td>
<td>1, 2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>52</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>53</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>54</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>55</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>56</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>57</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>58</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>59</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>60</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>61</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>62</td>
<td>1, 4</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>63</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>64</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>65</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>66</td>
<td>1</td>
<td>2</td>
<td>1, 3</td>
<td>1</td>
</tr>
<tr>
<td>67</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>68</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>69</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>70</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>71</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>72</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>73</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>74</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>75</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>76</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>77</td>
<td>3</td>
<td>1,2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>78</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>79</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>80</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>81</td>
<td>1,2</td>
<td>1,3,4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>82</td>
<td>3</td>
<td>1,2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>83</td>
<td>1</td>
<td>3,4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>84</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>85</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>86</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>87</td>
<td>1,2</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>88</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>89</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>90</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>91</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>92</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>93</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>94</td>
<td>2</td>
<td>2,3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>95</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>96</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>97</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>98</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>99</td>
<td>2</td>
<td>2,3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>101</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Appendix C. Study materials from Manuscript B

Consent Form

**Project Title:** Digital Modeling using Game Data: Phase FourB

**Researcher(s):** Dr. Regan Mandryk, Professor, Department of Computer Science, University of Saskatchewan, 306-966-4888, regan@cs.usask.ca

Student Researcher: Katelyn Wiley, Ph.D. student, Department of Computer Science, University of Saskatchewan, 306-966-2327, katelyn.wiley@usask.ca

**Purpose(s) and Objective(s) of the Research:**
- The purpose of this project is to evaluate a game under development and to statistically model whether different types of players experience the game differently.

**Procedures:**
- In this study, you will be asked to complete a survey, asking you some questions about yourself, including questions about your well-being, mood, and personality. Following the completion of the surveys, you will play a computer game for 20 minutes – there will be a tutorial describing how to play the game. Following gameplay, you will be asked to complete additional questionnaires that ask you questions about your play experience.
- Please feel free to ask any questions regarding the procedures and goals of the study or your role.

**Funded by:** The Natural Sciences and Engineering Research Council of Canada (NSERC).

**Potential Risks:**
- There are no known or anticipated risks to you by participating in this specific research project.
- There is always a risk of minor anxiety or stress associated with participating in an experiment.
- Some of the survey questions may prompt you to reflect on your life and feel emotional discomfort as a result of that process.

**Potential Benefits:**
- Your participation will help us to design games, including serious games designed to promote well-being.

**Compensation:**
- To thank you for participating, we will provide you with a $10 honorarium.
- The entire experiment should take approximately 45 minutes to complete.

**Confidentiality:**
- Confidentiality will be maintained throughout the study. The entire process and data will be anonymized. Data will only be presented in the aggregate and any individual user comments will be anonymized prior to presentation in academic venues.
- Vignettes of system use that may include photographs or video representations of you will only be shown in academic presentations with your explicit written permission, after having been given the opportunity to review the photo or video of the interaction.
• Only the principal researcher and her research assistants will have access to the data to ensure that your confidentiality is protected.

• **Storage of Data:**
  - Data (including survey and interview responses, logs of computer use, and videos of interaction) will be stored on a secure password-protected server for 7 years after data collection.
  - After 7 years, the data will be destroyed. Paper data will be shredded and digital data will be wiped from hard disks beyond any possibility for data recovery.

**Right to Withdraw:**
• Your participation is voluntary and you can answer only those questions that you are comfortable with. You may withdraw from the research project for any reason, at any time without explanation or penalty of any sort.
• Should you wish to withdraw, you may do so at any point without penalty, and we will not use your data; we will destroy all records of your data immediately.
• Your right to withdraw data from the study will apply until the data have been aggregated (one week after study completion). After this date, it is possible that some form of research dissemination will have already occurred and it may not be possible to withdraw your data.

**Follow up:**
• To obtain results from the study, please contact Dr. Regan Mandryk (regan@cs.usask.ca).

**Questions or Concerns:**
• Contact the researcher(s) using the information at the top of the form.
• This research project has been approved on ethical grounds by the University of Saskatchewan Research Ethics Board. Any questions regarding your rights as a participant may be addressed to that committee through the Research Ethics Office ethics.office@usask.ca (306) 966-2975. Out of town participants may call toll free (888) 966-2975.

**Consent:**

Do you give your consent?
○ I consent
○ I do not consent

*Please print a copy of this consent form for your records*
Demographics

Please answer the following questions about yourself.

What is your age? (numeric field, min = 1, max = 120)

Indicate your gender: (drop down menu)
   o Female
   o Male
   o Other

Insert your gender in case you selected ‘other’ in the previous question: (text field)

What is the highest degree or level of school you have completed? If currently enrolled, mark the previous grade or highest degree received: (drop down menu)
   o No schooling completed
   o Nursery school to 8th grade
   o 9th, 10th or 11th grade
   o 12th grade, no diploma
   o High school graduate - high school diploma or the equivalent (for example: GED)
   o Some college credit, but less than 1 year
   o 1 or more years of college, no degree
   o Associate degree
   o Bachelor's degree
   o Master's degree
   o Professional degree
   o Doctorate degree
   o I prefer not to answer

If you are a student, please indicate your subject: (text field)

Please indicate your employment status: (drop down menu)
   o Employed for wages
   o Self-employed
   o Out of work and looking for work
   o Out of work but not currently looking for work
   o A homemaker
   o A student
   o Military
   o Retired
   o Unable to work
   o I prefer not to answer

Please indicate your marital status: (drop down menu)
   o Single, never married
- Married or domestic partnership
- Widowed
- Divorced
- Separated
- I prefer not to answer

Please indicate your household income: *(drop down menu)*
- Less than $10,000
- $10,000 to $25,000
- $25,001 to $45,000
- $45,001 to $65,000
- $65,001 to $85,000
- $85,001 to $100,000
- $100,001 to $150,000
- $150,001 or more
- I prefer not to answer

Please indicate your ethnicity: *(drop down menu)*
- American Indian or Alaskan Native
- Asian
- Native Hawaiian or Other Pacific Islander
- Black or African American
- Hispanic/Latino
- White
- Two or more categories
- I prefer not to answer
Games

Please answer the following questions yourself and video/computer games.

Please indicate how often (on average) you play games: *(drop down menu)*
- Every day
- A few times per week
- A few times per month
- A few times per year

If you played games more or less in the past, please indicate how much you used to play: *(drop down menu)*
- Every day
- A few times per week
- A few times per month
- A few times per year
- Never

How much do you self-identify as a gamer on the following scale: *(slider on continuous scale)*

Not at all *(value 0)* [ ] Gamer *(value 100)*

Please indicate the genres that you enjoy playing (check all that apply): *(checkbox)*
- Action
- Platform games
- First person shooter
- Beat ‘em up
- Adventure
- Role playing games
- Mass Multiplayer Role Playing Games (MMORPG)
- Simulation
- Vehicle simulation
- Strategy
- Music games
- Puzzle games
- Sport games
- Multiplayer Online Battle Arena (MOBA)
- Casual games
- Different genre(s)

Please indicate on which device(s) you play (check all that apply): *(checkbox)*
- Desktop (e.g., Windows, Linux, OS X, etc.)
- Console (e.g., X-Box, PlayStation, etc.)
- Mobile device (e.g., phone, tablet, PS Portable, etc.)
- Different device(s)
On a scale of ‘Strongly Disagree’ to ‘Strongly Agree’, please indicate to what extent you agree with the following statements: *(radio grid, item order randomly shuffled)*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel good about playing games.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I feel that playing games is a meaningful activity.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I feel that playing games is a waste of time.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>I feel that playing games is useless.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
State-Trait Anxiety Inventory

Below is a list of statements dealing with your general feelings about yourself. Please rate each of the following statements.

A number of statements which people have used to describe themselves are given below. Read each statement and then, using the scale below, indicate how you feel right now, that is, at this moment. There are no right or wrong answers. Do not spend too much time on any one statement but give the answer which seems to describe your present feelings best. *(radio grid, item order randomly shuffled)*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel calm.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel secure.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am tense.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel strained.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel at ease.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel upset.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am presently worrying over possible misfortunes.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am satisfied.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel frightened.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel comfortable.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel self-confident.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel nervous.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am jittery.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel indecisive.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am relaxed.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel content.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am worried.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel confused.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel steady.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel pleasant.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A number of statements which people have used to describe themselves are given below. Read each statement and then, using the scale below, indicate how you generally feel. There are no right or wrong answers. Do not spend too much time on any one statement, but give the answer which seems to describe how you generally feel. *(radio grid, item order randomly shuffled)*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel pleasant.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel nervous and restless.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel satisfied with myself.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I wish I could be as happy as others seem to be.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel like a failure.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel rested.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am calm, cool, and collected.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel that difficulties are piling up so that I cannot overcome them.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I worry too much over something that really doesn’t matter.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I have disturbing thoughts.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I lack self-confidence.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I make decisions easily.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel inadequate.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am content.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Some unimportant thought runs through my mind and bothers me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I take disappointments so keenly that I can’t put them out of my mind.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I am a steady person.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel confused.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I feel steady.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I get in a state of tension or turmoil as I think over my recent concerns and interest.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>
Positive and Negative Affect Schedule

This scale consists of a number of words and phrases that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you feel this way right now. *radio grid, item order randomly shuffled*

<table>
<thead>
<tr>
<th></th>
<th>Very Slightly</th>
<th>A little</th>
<th>Moderately</th>
<th>Quite a Bit</th>
<th>Extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Interested</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Excited</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Strong</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Inspired</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Proud</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Enthusiastic</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Determined</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Attentive</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Alert</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Upset</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Irritable</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Guilty</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Scared</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Hostile</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Distressed</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Ashamed</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Nervous</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Jittery</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Afraid</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>
Player Experience Inventory

Reflect on your play experiences and rate your agreement with the following statements. *(radio grid, item order randomly shuffled)*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neutral</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel competent at the game.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I feel very capable and effective when playing.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>My ability to play the game is well matched with the game’s challenges.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>The game provides me with interesting options and choices.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>The game lets you do interesting things.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I experienced a lot of freedom in the game.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I am not impacted emotionally by events in the game.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>The game was emotionally engaging.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I experience feelings as deeply in the game as I have in real life.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>When I accomplished something in the game I experience genuine pride.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Learning the control was easy.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>The game controls are intuitive.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>When I wanted to do something in the game, it was easy to remember the corresponding control.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>
Enjoyment Questions

We added an enjoyment questionnaire that was given before, during (after the introduction sequence but before play), and after the task. Comprised of five statements—“I am enjoying playing this game”; “Playing this game is fun”; “This game is interesting”; “I am thinking about how much I am enjoying playing”; “This game is not holding my attention” (reverse-coded)—participants rated their agreement on a 7-pt scale. Items were worded for each time point, e.g., “I will enjoy playing this game”, “I am enjoying playing this game”, and “I enjoyed this game”.

Reflect on your play experiences and rate your agreement with the following statements. *(radio grid, item order randomly shuffled)*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neutral</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am enjoying playing this game.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Playing this game is fun.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>This game is interesting.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I am thinking about how much I am enjoying playing.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>This game is not holding my attention.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>
Appendix D. Study materials from Manuscript C

Consent Form

**Project Title:** Digital Modeling using Game Data: Phase FourB

**Researcher(s):** Dr. Regan Mandryk, Professor, Department of Computer Science, University of Saskatchewan, 306-966-4888, regan@cs.usask.ca

Student Researcher: Katelyn Wiley, Ph.D. student, Department of Computer Science, University of Saskatchewan, 306-966-2327, katelyn.wiley@usask.ca

**Purpose(s) and Objective(s) of the Research:**
- The purpose of this project is to evaluate a cognitive task under development and to statistically model whether game elements added to the task affect data quality.

**Procedures:**
- In this study, you will be asked to complete a cognitive task for 10 minutes. There will be a tutorial describing how to complete the task. Following the task, you will be asked to complete questionnaires about yourself and your experience of the task.
- Please feel free to ask any questions regarding the procedures and goals of the study or your role.

**Funded by:** The Natural Sciences and Engineering Research Council of Canada (NSERC).

**Potential Risks:**
- There are no known or anticipated risks to you by participating in this specific research project.
- There is always a risk of minor anxiety or stress associated with participating in an experiment.
- Some of the survey questions may prompt you to reflect on your life and feel emotional discomfort as a result of that process.

**Potential Benefits:**
- Your participation will help us to design games, including serious games designed to promote well-being.

**Compensation:**
- To thank you for participating, we will provide you with a $10 honorarium.
- The entire experiment should take approximately 45 minutes to complete.

**Confidentiality:**
- Confidentiality will be maintained throughout the study. The entire process and data will be anonymized. Data will only be presented in the aggregate and any individual user comments will be anonymized prior to presentation in academic venues.
- Vignettes of system use that may include photographs or video representations of you will only be shown in academic presentations with your explicit written permission, after having been given the opportunity to review the photo or video of the interaction.
- Only the principal researcher and her research assistants will have access to the data to ensure that your confidentiality is protected.

**Storage of Data:**
Data (including survey and interview responses, logs of computer use, and videos of interaction) will be stored on a secure password-protected server for 7 years after data collection.

After 7 years, the data will be destroyed. Paper data will be shredded and digital data will be wiped from hard disks beyond any possibility for data recovery.

Right to Withdraw:

- Your participation is voluntary and you can answer only those questions that you are comfortable with. You may withdraw from the research project for any reason, at any time without explanation or penalty of any sort.
- Should you wish to withdraw, you may do so at any point without penalty, and we will not use your data; we will destroy all records of your data immediately.
- Your right to withdraw data from the study will apply until the data have been aggregated (one week after study completion). After this date, it is possible that some form of research dissemination will have already occurred and it may not be possible to withdraw your data.

Follow up:

- To obtain results from the study, please contact Dr. Regan Mandryk (regan@cs.usask.ca).

Questions or Concerns:

- Contact the researcher(s) using the information at the top of the form.
- This research project has been approved on ethical grounds by the University of Saskatchewan Research Ethics Board. Any questions regarding your rights as a participant may be addressed to that committee through the Research Ethics Office ethics.office@usask.ca (306) 966-2975. Out of town participants may call toll free (888) 966-2975.

Consent:

Do you give your consent?
- I consent
- I do not consent

Please print a copy of this consent form for your records
Questionnaires

Demographics

Please answer the following questions about yourself.

What is your age? (numeric field, min = 1, max = 120)

Indicate your gender: (drop down menu)
- Female
- Male
- Non-binary
- I prefer not to disclose

What is the highest degree or level of school you have completed? If currently enrolled, mark the previous grade or highest degree received: (drop down menu)
- No schooling completed
- Nursery school to 8th grade
- 9th, 10th or 11th grade
- 12th grade, no diploma
- High school graduate - high school diploma or the equivalent (for example: GED)
- Some college credit, but less than 1 year
- 1 or more years of college, no degree
- Associate degree
- Bachelor's degree
- Master's degree
- Professional degree
- Doctorate degree
- I prefer not to answer

If you are a student, please indicate your subject: (text field)

Please indicate your employment status: (drop down menu)
- Employed for wages
- Self-employed
- Out of work and looking for work
- Out of work but not currently looking for work
- A homemaker
- A student
- Military
- Retired
- Unable to work
- I prefer not to answer

Please indicate your marital status: (drop down menu)
- Single, never married
- Married or domestic partnership
o Widowed
o Divorced
o Separated
o I prefer not to answer

Please indicate your household income: *(drop down menu)*
  o Less than $10,000
  o $10,000 to $25,000
  o $25,001 to $45,000
  o $45,001 to $65,000
  o $65,001 to $85,000
  o $85,001 to $100,000
  o $100,001 to $150,000
  o $150,000 or more
  o I prefer not to answer

Please indicate your ethnicity: *(drop down menu)*
  o American Indian or Alaskan Native
  o Asian
  o Native Hawaiian or Other Pacific Islander
  o Black or African American
  o Hispanic/Latino
  o White
  o Two or more categories
  o I prefer not to answer
Games

Please answer the following questions yourself and video/computer games.

Please indicate how often (on average) you play games: (drop down menu)
- Every day
- A few times per week
- A few times per month
- A few times per year

If you played games more or less in the past, please indicate how much you used to play: (drop down menu)
- Every day
- A few times per week
- A few times per month
- A few times per year
- Never

How much do you self-identify as a gamer on the following scale: (slider on continuous scale)
Not at all (value 0) — Gamer (value 100)

Which is your dominant hand? (drop down menu)
- Left
- Right

Please indicate the genres that you enjoy playing (check all that apply): (checkbox)
- Action
- Platform games
- First person shooter
- Beat ‘em up
- Adventure
- Role playing games
- Mass Multiplayer Role Playing Games (MMORPG)
- Simulation
- Vehicle simulation
- Strategy
- Music games
- Puzzle games
- Sport games
- Multiplayer Online Battle Arena (MOBA)
- Casual games
- Battle Royale

Please indicate on which device(s) you play (check all that apply): (checkbox)
- Desktop (e.g., Window, Linux, OS X, etc.)
- Console (e.g., XBox, Play Station, Wii, etc.)
- Portable consoles (e.g., PSP, GameBoy, DS, etc.)
- Mobile devices (e.g., phone, tablet, etc.)
- Different device(s)

If there are other types of devices you play on, please list them. (text field)
**Attentional Control Scale**

Here are some different ways that people can feel about working and concentrating. Please indicate how strongly each statement applies to you. *(radio grid, item order randomly shuffled)*

<table>
<thead>
<tr>
<th></th>
<th>Almost Never</th>
<th>Sometimes</th>
<th>Often</th>
<th>Always</th>
</tr>
</thead>
<tbody>
<tr>
<td>It’s very hard for me to concentrate on a difficult task when there are noises around.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>When I need to concentrate and solve a problem, I have trouble focusing my attention.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>When I am working hard on something, I still get distracted by events around me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>My concentration is good even if there is music in the room around me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>When concentrating, I can focus my attention so that I become unaware of what’s going on in the room around me.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>When I am reading or studying, I am easily distracted if there are people talking in the same room.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>When trying to focus my attention on something, I have difficulty blocking out distracting thoughts.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I have a hard time concentrating when I’m excited about something.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>When concentrating I ignore feelings of hunger or thirst.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I can quickly switch from one task to another.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>It takes me a while to get really involved in a new task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>It is difficult for me to coordinate my attention between the listening and writing required when taking notes during lectures.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I can become interested in a new topic very quickly when I need to.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>It is easy for me to read or write while I'm also talking on the phone.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I have trouble carrying on two conversations at once.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>----------------------------------------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>I have a hard time coming up with new ideas quickly.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>After being interrupted or distracted, I can easily shift my</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>attention back to what I was doing before.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>When a distracting thought comes to mind, it is easy for me to</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>shift my attention away from it.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>It is easy for me to alternate between two different tasks.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>It is hard for me to break from one way of thinking about</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>something and look at it from another point of view.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Intrinsic Motivation Inventory

A number of statements which people have used to describe the previous task are given below. Read each statement and indicate your agreement with that statement. There are no right or wrong answers. Do not spend too much time on any one statement. Remember, give the answer which seems to describe how you thought during the game. *(radio grid, item order randomly shuffled)*

<table>
<thead>
<tr>
<th>Statement</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Slightly Disagree</th>
<th>Neutral</th>
<th>Slightly Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I enjoyed this task very much.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I think I am pretty good at this task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I put a lot of effort into this task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>It was important to me to do well at this task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I felt tense while completing the task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I experienced a lot of freedom in the game.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I tried very hard while completing the task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Completing the task was fun.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I would describe this task as very interesting.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I am satisfied with my performance at this task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I felt pressured while completing the task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I was anxious while completing the task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I didn’t try very hard at completing the task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>While completing the task, I was thinking about how much I enjoyed it.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>After doing the task for a while, I felt pretty competent.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>I was very relaxed while completing the task.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Statement</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>-----------------------------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>I am pretty skilled at the task.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This task did not hold my attention.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I couldn’t do this task very well.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Task Instructions

Page 1:
Instructions for the task:
In this task, you will see various words in different “ink” colours. When you see a word, please respond by pressing keys associated with the ink colour. For example, you might see:

SHIP

You would then need to respond to the ink colour (blue), and press the associated key, “x”. The other keys used for this task are “z”, “n” and “m” for red, green and yellow, respectively.

Sometimes the words will be colours. For example, you might see:

YELLOW

Please continue to respond to the ink colour (red), and press the associated key, “z”.

Press the space bar for more instructions.

Page 2:
The task will begin with a training session, so that you can learn which ink colours are associated with each key. These are the ink colours and their associated keys:

XXX = z
XXX = x
XXX = n
XXX = m

Press the space bar for more instructions.

Page 3:
For each word, please respond with the associated key press as quickly and accurately as possible.

Press the space bar to begin the training sessions.

After the training session:
Thank you for completing the training session. Now the task will begin. Again, for each word, please respond with the associated key press as quickly and accurately as possible.

You will receive points for accurate responses, and even more points for fast responses!

Press the space bar to begin the trial.