GAMIFIED LOW-PREVALENCE VIGILS

A Dissertation by

Andrew T. Miranda

M.A. Wichita State University, 2012
B.A. Kansas State University, 2009

Submitted to the Department of Psychology
and the faculty of the Graduate School of
Wichita State University
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

December 2015
© Copyright 2015 by Andrew Miranda

All Rights Reserved
GAMIFIED LOW-PREVALENCE VIGILS

The following faculty members have examined the final copy of this dissertation for form and content, and recommend that it be accepted in partial fulfillment of the requirement for the degree of Doctor of Philosophy.

________________________
Evan Palmer, Committee Chair

________________________
Barbara Chaparro, Committee Member

________________________
Alex Chaparro, Committee Member

________________________
Joe Keebler, Committee Member

________________________
Nicole Rogers, Committee Member

Accepted for the College of Liberal Arts & Sciences

________________________
Ron Matson, Dean

Accepted for the Graduate School

________________________
Kerry Wilks, Interim Dean
DEDICATION

To my wife, Beth
Do not undertake a scientific career in quest of fame or money. There are easier and better ways to reach them. Undertake it only if nothing else will satisfy you; for nothing else is probably what you will receive. Your reward will be the widening of the horizon as you climb. And if you achieve that reward you will ask no other.

Cecilia Payne-Gaposchkin (1900-1979)
ACKNOWLEDGEMENTS

I would like to thank my adviser, Evan Palmer, for his many years of continued support, guidance, and encouragement. I would also like to extend appreciation to the committee, Barbara Chaparro, Alex Chaparro, Joe Keebler, and Nicole Rogers, for their helpful comments and suggestions on this project.

I was fortunate to spend time with a number of exceptional colleagues during my time in graduate school, to which I owe immense gratitude. To Chris Brown and Nichole Morris, who both taught me what it took to be an effective researcher. To Rondell Burge, Bobby Nguyen, Mikki Phan, and Jen Teves, who inspired me to get the most out of my time in graduate school. To Duy Nguyen and Tiffany Leverenz, whose mere presence made those moments enjoyable. And most of all to Samantha Jansen and Sarah Fouquet for being the best friends I could ever ask for during some of the best years of my life.

I would also like to thank my family for all of their support. To my in-laws, John and Joan, for always listening and providing encouragement. To my parents, who showed me the importance of both a strong work ethic and a sense of humor. Lastly, to my wife, who is my favorite person to be around, no matter what.
ABSTRACT

The vigilance decrement has been a well-studied human factors problem since World War II (Mackworth, 1948). A more recently discovered and fundamentally similar problem is the low-prevalence effect in visual search (e.g., Lau & Huang, 2010). Both circumstances typically produce substantial misses of critical signals, whether from a vigilance activity requiring sustained attention to a single information source for a prolonged period of time (See, Howe, Warm, & Dember, 1995), or from a visual search for rare targets in difficult displays (Wolfe, Horowitz, & Kenner, 2005).

Furthermore, efforts have been made to apply some of the practices from the gaming industry towards non-gaming environments to increase task engagement and improve performance, a practice known as gamification (Deterding, Dixon, Khaled, & Nacke, 2011; Hamari, Koivisto, & Sarsa, 2014). The present project investigates the potential benefits of gamification in a low-prevalence vigil.

A simulated inspection task featuring images of round metal washers as search stimuli was created for this study. Five total experiments were conducted: the initial four each assessed the impact of an individual gaming element on performance and experience (badges, a points-based challenge during a brief burst of high prevalence, storytelling, and points-based feedback throughout the task). Badges and points-based feedback throughout the task were ineffective at improving performance while the points-based challenge and storytelling were. The final experiment assessed gamified training as a practical way to integrate gamification in a low-prevalence vigil. Gamified training did not produce any significant benefits compared to traditional or no-intervention training. Practical and theoretical implications are discussed.
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. LITERATURE REVIEW</td>
<td>1</td>
</tr>
<tr>
<td>Signal Detection Theory</td>
<td>3</td>
</tr>
<tr>
<td>Vigilance</td>
<td>5</td>
</tr>
<tr>
<td>Arousal Theory</td>
<td>6</td>
</tr>
<tr>
<td>Expectancy Theory</td>
<td>8</td>
</tr>
<tr>
<td>Sustained Demand Theory</td>
<td>10</td>
</tr>
<tr>
<td>Low-Prevalence Effect</td>
<td>14</td>
</tr>
<tr>
<td>Correctable Responses</td>
<td>17</td>
</tr>
<tr>
<td>Burst of High Signal Probability</td>
<td>21</td>
</tr>
<tr>
<td>Low-Prevalence Effect and Vigilance</td>
<td>23</td>
</tr>
<tr>
<td>Task Manipulations</td>
<td>25</td>
</tr>
<tr>
<td>Knowledge of Results</td>
<td>26</td>
</tr>
<tr>
<td>Instructions</td>
<td>28</td>
</tr>
<tr>
<td>Mental Breaks</td>
<td>29</td>
</tr>
<tr>
<td>Rewards</td>
<td>30</td>
</tr>
<tr>
<td>Motivation</td>
<td>33</td>
</tr>
<tr>
<td>Self-Determination Theory</td>
<td>35</td>
</tr>
<tr>
<td>Potential for Improvement</td>
<td>37</td>
</tr>
<tr>
<td>Gamification</td>
<td>38</td>
</tr>
<tr>
<td>Why Gamification Works</td>
<td>40</td>
</tr>
<tr>
<td>When Gamification Works</td>
<td>42</td>
</tr>
<tr>
<td>Introduction Summary</td>
<td>49</td>
</tr>
</tbody>
</table>

| II. SIMULATED INSPECTION TASK DEVELOPMENT | 51 |
| Background | 51 |
| Experiment Zero: Simulated Inspection Pilot | 55 |
| Simulated Inspection Task Discussion | 59 |
# TABLE OF CONTENTS (continued)

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>III. METHODS AND RESULTS</td>
<td>63</td>
</tr>
<tr>
<td>Experiments 1a-1d: Assessing Individual Gaming Elements</td>
<td>64</td>
</tr>
<tr>
<td>General Methods</td>
<td>64</td>
</tr>
<tr>
<td>Experiment 1a: Badges and Traditional Feedback</td>
<td>66</td>
</tr>
<tr>
<td>Experiment 1b: Points-Based Challenge and Burst of Higher Prevalence</td>
<td>73</td>
</tr>
<tr>
<td>Experiment 1c: Storytelling and Instructions</td>
<td>80</td>
</tr>
<tr>
<td>Experiment 1d: Points-Based Feedback and Traditional Feedback</td>
<td>87</td>
</tr>
<tr>
<td>Experiment 2: Gamified Training</td>
<td>92</td>
</tr>
<tr>
<td>Experimental Design and Power Analysis</td>
<td>94</td>
</tr>
<tr>
<td>Results Summary</td>
<td>105</td>
</tr>
<tr>
<td>IV. GENERAL DISCUSSION</td>
<td>106</td>
</tr>
<tr>
<td>Vigilance &amp; The Low-Prevalence Effect</td>
<td>106</td>
</tr>
<tr>
<td>Gamification</td>
<td>108</td>
</tr>
<tr>
<td>Conclusion</td>
<td>112</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>114</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>129</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1. Summary table of common gaming elements used in gamification</td>
<td>40</td>
</tr>
<tr>
<td>2.1. List of the operational definitions for each of the four types of washer defects</td>
<td>55</td>
</tr>
<tr>
<td>3.1. List of badge descriptions and images</td>
<td>69</td>
</tr>
<tr>
<td>3.2. Points system used during the Points-Based Challenge</td>
<td>76</td>
</tr>
<tr>
<td>3.3. Summary of significant findings across initial series of experiments</td>
<td>91</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1. Four outcomes of a response according to Signal Detection Theory, adapted from Wickens and Hollands (2000)</td>
<td>4</td>
</tr>
<tr>
<td>1.2. The Parasuraman and Davies (1977) vigilance taxonomy</td>
<td>11</td>
</tr>
<tr>
<td>1.3. Summarized results from the original Wolfe et al. (2005) Low-prevalence effect finding</td>
<td>15</td>
</tr>
<tr>
<td>1.4. Decrement slope results from Methot and Huitema (1998) for individual participants across signal probability conditions</td>
<td>20</td>
</tr>
<tr>
<td>1.5. Summary of primary similarities and differences between vigilance tasks and low prevalence visual searches</td>
<td>23</td>
</tr>
<tr>
<td>1.6. Diagram displaying how various manipulations may impact the vigilance decrement or low prevalence effect</td>
<td>25</td>
</tr>
<tr>
<td>1.7. Google Trend for “gamification” as a search term</td>
<td>43</td>
</tr>
<tr>
<td>2.1. Images of acceptable and defective washers used in the experiment</td>
<td>54</td>
</tr>
<tr>
<td>2.2. Screenshot of search stimuli. Feedback was provided during the practice block but not during the experimental session</td>
<td>56</td>
</tr>
<tr>
<td>2.3. Experiment Zero mean hit rates across the seven epochs</td>
<td>58</td>
</tr>
<tr>
<td>2.4. Experiment Zero criterion values across epochs</td>
<td>59</td>
</tr>
<tr>
<td>3.1. Badges used as a gamified version of traditional feedback in Experiment 1a</td>
<td>68</td>
</tr>
<tr>
<td>3.2. Experiment 1a comparing hit rate between Badges and Traditional Feedback conditions</td>
<td>70</td>
</tr>
<tr>
<td>3.3. Experiment 1a mean flow dimension ratings</td>
<td>71</td>
</tr>
<tr>
<td>3.4. Points-Based Challenge used as a gamified version of a burst of high target prevalence trials in Experiment 1b</td>
<td>75</td>
</tr>
<tr>
<td>3.5. Experiment 1b comparing Points-Based Challenge and Traditional High-Prevalence Burst. Criterion values across epochs</td>
<td>77</td>
</tr>
<tr>
<td>3.6. Mean hit rate during two-minute burst between second and third epochs</td>
<td>78</td>
</tr>
<tr>
<td>3.7. Mean reaction time during two-minute burst between second and third epochs</td>
<td>79</td>
</tr>
<tr>
<td>3.8. Storytelling used as a gamified version of instruction framing in Experiment 1c</td>
<td>81</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>3.9. Experiment 1c comparing Storytelling and Traditional Instructions. Mean sensitivity (d') across epochs</td>
<td>84</td>
</tr>
<tr>
<td>3.10. Experiment 1c comparing Storytelling and Traditional Instructions. Mean NASA-TLX Mental Workload Assessment subscale ratings</td>
<td>85</td>
</tr>
<tr>
<td>3.11. Points-based feedback was used as a gamified version of traditional performance feedback in Experiment 1d</td>
<td>88</td>
</tr>
<tr>
<td>3.12. Experiment 2 design</td>
<td>93</td>
</tr>
<tr>
<td>3.13. Experiment 2 mean hit rate during two two-minute burst after second and fifth epochs during training session</td>
<td>98</td>
</tr>
<tr>
<td>3.14. Experiment 2 mean test session hit rate across all three conditions</td>
<td>100</td>
</tr>
<tr>
<td>3.15. Comparing hit rate performance between late-semester participants from Experiment 1c and early-semester participants from Experiment 2</td>
<td>104</td>
</tr>
</tbody>
</table>
CHAPTER 1
LITERATURE REVIEW

The classic humanistic perspective proposes that people instinctively want to grow intellectually, be challenged, and reach their full potential (Maslow, 1954, 1968; Rogers, 1963). From this perspective, humans have an innate desire to engage their interests, be challenged when executing a task, and ultimately perform well and succeed in their endeavors. When opportunity arises, we embrace challenges that allow us to grow as individuals.

Oftentimes, however, we encounter certain tasks within the workplace that do not allow us to prosper, whether it is because the demands of the task are too high or the design of the system is cumbersome and hinders performance (Hancock, 2013; Szalma, 2014; Warm, Parasuraman, & Matthews, 2008). Considering the goal of human factors research is to apply knowledge of human limitations and capabilities towards creating effective, efficient, and satisfactory human-machine systems (Wickens & Hollands, 2000), it is our responsibility to design and develop workplace tasks that allow users to not only succeed, but perhaps enjoy themselves at the same time.

There are plenty of activities that we as humans engage in because of the inherent interest and enjoyment, not because we have to. One of the most popular tasks that possesses this quality is games. People love playing games, not only because they are entertaining and interesting, but also because they can be challenging (Lazzaro, 2004). The characteristics that define games can be inspirational for developing and designing workplace activities (Mollick & Rothbard, 2014). The principle of applying gaming elements to non-game contexts, otherwise known as gamification, has been shown to improve performance and satisfaction in a variety of
environments (Hamari et al., 2014). The current project examines how gamification can be applied to improving detection for rare critical signals in a simulated vigilance search task.

Vigilance has been characterized as sustained attention on a particular task for a prolonged period of time (See et al., 1995). It has been a focus of human factors research since World War II. Norman Mackworth (1948) documented that Naval sonar operators experienced fatigue and suffered reduced detection of critical signals after a half hour of watch, with subsequent gradual performance loss. Mackworth named this decline in performance the “vigilance decrement,” and since that time the phenomenon has been replicated in laboratory experiments and observed in industry and military settings numerous times (Parasuraman, 1986; Wiener, 1987). The decrement is often characterized by drops in one or more of three performance metrics: reduced signal detection, increased miss rates, or increased response time (RT) (See et al., 1995). A decrement coefficient, or decrement slope, is an effective method for observing how overall performance changes over time for each individual. It is a single value representing, on average, how a particular performance metric (e.g., miss rate) changes per unit of time (e.g., Methot & Huitema, 1998).

A similar attention limitation is observed in certain visual search tasks, in which an observer must locate a target among distractors (Klein & MacInnes, 1999; Treisman & Gelade, 1980; Wolfe & Horowitz, 2004). This activity is something humans do on a daily basis, whether it is locating their car keys on the cluttered kitchen counter, a familiar face in a crowd, or a particular name in a list of recent text messages. When studied in the laboratory, visual search performance is typically observed using hundreds of trials with 50% of the trials containing a target and 50% of the trials not containing a target (e.g., Huang & Pashler, 2005). Performance
Signal Detection Theory

Since vigilance tasks require the observer to make a decision about the presence or absence of signals in the world, SDT metrics are often applied. Originally developed by Green and Swets (1966), the goal of the SDT framework is to quantify successful signal detection in an observer (Macmillan & Creelman, 2005). Essentially, this theory is applicable in any situation that requires someone to distinguish important information needing to be identified from unimportant information irrelevant to the task. Environmental information relevant to task success is the signal, while unimportant, distracting information is referred to as noise (Wickens & Hollands, 2000). The observer is responsible for making a decision when deciphering what in
the environment is the signal compared to what is noise. SDT proposes this process can be understood and measured using two components (Green & Swets, 1966): 1) the observer’s ability to discern the signal being monitored from non-signal noise (i.e., sensitivity), and 2) the willingness of the observer to report whether the signal is present or absent (i.e., response bias or criterion). These two elements together provide important information about an observer’s ability to detect a signal within a vigilance task (Wickens & Hollands, 2000). Both sensitivity and criterion are measured by obtaining the four possible outcomes of a signal detection activity (see Figure 1.1).

Sensitivity is typically measured by the $d'$ variable, which is calculated by subtracting the normalized false alarm rate from the normalized hit rate (Macmillan & Creelman, 2005). A high level of target sensitivity would yield a high $d'$ score, indicating the observer is able to successfully and consistently distinguish the target stimulus from noise. Independent of

![Figure 1.1. Four outcomes of a response according to Signal Detection Theory, adapted from Wickens and Hollands (2000).](image-url)
sensitivity, however, is the response criterion of the observer, which reflects their willingness to respond “present” or “absent.” Criterion (c) is also calculated using the false alarm and hit rates of an observer during a signal detection task (Macmillan & Creelman, 2005). The c variable is a measure of operator response-bias and a reflection of the operator preferring to respond “yes, target is present” or “no, target is absent.” If both false alarm and hit rates are high, then the operator would have a low, liberal criterion, indicating that he or she is quite willing to report that the target is present. If both rates are low, then the operator would have a high, conservative criterion, indicating a general unwillingness to respond “target present.”

The appropriate criterion can vary depending on the context. Take, for instance, the Transportation Security Administration agent examining baggage going through security at an airport – a liberal bias is appropriate to avoid a miss, which could result in a destructive threat making it through security. On the other hand, a conservative criterion may be more appropriate for an industrial inspector examining products for minor defects coming off the conveyor line. The inspector may have been instructed to minimize false alarms, which would result in holding up the production and cause unnecessary costs for the manufacturer.

Vigilance

Both sensitivity and criterion play important roles in determining the ability of an observer to successfully detect a critical signal. The following section discusses three vigilance theories that explain the sources of the vigilance decrement and how both components of SDT are affected.
Arousal Theory

One of the earliest theories of the vigilance decrement was the Arousal Theory (Welford, 1968). Typical vigilance tasks are characterized by monotony. During prolonged periods of maintaining attention on a single task, it is considered that the monotony, tediousness, and perceived rare occurrences of critical events during a vigilance task reduce overall brain activity, leading to a reduced sense of arousal and wakefulness. A study from Milošević (1978) provides some empirical neurological evidence to support the relationship between state of arousal and signal detection in a vigilance task. During the vigilance task, participant brain activity was recording using electroencephalography (EEG) imaging. As expected, correct target detection and sensitivity reduced over time. Additionally, participant overall EEG activity reduced over time. The correlation of reduced task performance and reduced neural activity provided evidence to support the relationship between arousal state and vigilance performance.

A study by Dardano (1962) indicated that arousal levels observed via galvanic skin response tend to decline in participants performing a vigilance activity. Galvanic skin response is a commonly used method to measure electrical conductivity of the skin, which is an objective observation of psychophysiological arousal (e.g. Vetrugno, Liguori, Cortelli, & Montagna, 2003). The Dardano (1962) study found that reaction times increased and skin conductance decreased over the course of a 3-hour vigil. The results provide further evidence supporting the role of arousal in vigilance task performance.

In a study investigating the effects of both caffeine (a stimulant) and diphenhydramine (a sedative) on the vigilance decrement, Fine et al. (1994) found that both types of drugs impact performance over time in a vigilance task. Target detection performance was better when
participants were administered moderate doses of caffeine compared to being administered a placebo. Performance was worse, however, when participants were administered the sedative diphenhydramine compared to the placebo. These results indicate that vigilance performance can be modulated by arousal levels via stimulant or sedative drug administration.

The Arousal Theory of the vigilance decrement is also consistent with the Yerkes-Dodson Law. This law proposes that performance on a given task is a function of the level of psychological arousal in the individual (Teigen, 1994; Yerkes & Dodson, 1908). Specifically, optimal performance in a task occurs when arousal state is “moderate.” If an activity is too mentally demanding or stimulating, the state of arousal may become too high and the observer may become anxious and nervous, causing performance to suffer. On the other hand, a boring and mundane activity like a vigilance task will produce low levels of arousal, leading to poor vigilance performance over time. Thus, the principles suggested by the Yerkes-Dodson Law are consistent with reduced performance over time in a vigilance activity.

The empirical evidence from Fine et al. (1994), Milošević (1978), and Dardano (1962), as well as theory from the Yerkes-Dodson law (1908), suggests there is a link between an observer’s arousal state and the ability to maintain adequate performance in a vigilance task. While the previously discussed studies have established the neurological consequence of monotony during a vigilance task, other research has found more subjective outcomes of vigilance. For example, Scerbo (1998) summarized a series of studies designed to observe the impact that vigilance has on perceived boredom. He found that as soon as 10 minutes into a vigilance task, participants may reach their maximum level of boredom. The monotony and dullness of vigilance tasks cause the operator to feel disengaged from the activity, as evidenced
by reduced neural activity and increased sense of boredom. The Arousal Theory postulates that the performance decrement over time is the consequence of this disconnect.

**Expectancy Theory**

The arousal theory of vigilance provided evidence of how state of arousal can play a role in reduced performance over time. In doing so, however, it was unable to account for the critical role of response bias within a vigilance task. The studies previously discussed emphasized the reduced perceptual ability to distinguish signal from noise over time during a vigil, which corresponds to a reduction in sensitivity in SDT. Arousal Theory, however, does not explain the role of response bias in a signal detection task. This is an important element for overall signal detection ability and must be taken into consideration.

Take, for instance, an industry inspector at a bottling factory. This individual must inspect bottle after bottle for any possible defects. The inspector could take some caffeine to possibly improve perceptual sensitivity towards defective bottles, but that will not change how willing he/she is to respond “defect present”. If the bottling factory manager has instructed the inspectors to avoid any “false alarms” of defective bottles to reduce unwanted costs, the inspector will need to take that into consideration. Now a “defect present” response from the inspector is a riskier response and could result in a problematic false alarm for the company. The result of the manager’s instruction will likely produce an criterion shift towards a more conservative approach.

Expectancy Theory emphasizes the role of criterion shifts within vigilance tasks (Baker, 1961). As stated previously, vigilance tasks are characterized by monotony and rare events. This theory suggests that the vigilance decrement is due in no small part to an upward shift of the
response bias toward being more conservative (i.e., less likely to response “target present”).

Returning to the bottling inspector, there might be a defective bottle once an hour, therefore he or she will have a reduced willingness to say “defective bottle present” because of the expectation that defective bottles are rare (Wickens & Hollands, 2000). This is the hallmark of the Expectancy Theory, which emphasizes that the observer’s willingness to response either “yes” or “no” to a target’s presence is shaped by the rarity of events during a vigilance task.

In a series of studies from Parasuraman (1979), response bias was shown to shift to a more conservative, cautious approach as the task continued. Participants were asked to identify a critical signal (a temporary increase in decibel level of an intermittent auditory tone) during a 45-minute vigil. Critical auditory signals were presented on average between one to two targets per minute. Throughout the studies, both hit rate and false alarm rate reduced over time, indicating the observer was becoming less willing to respond “target present.” This tendency of an upward criterion shift during a vigilance task has been replicated time and time again, including during visual vigilance tasks (Becker, Warm, Dember, & Hancock, 1995; Loeb & Binford, 1968; Parasuraman & Davies, 1977; Parasuraman, Warm, & Dember, 1987; See et al., 1995; See, Warm, Dember, & Howe, 1997). Expectancy Theory emphasizes that the criterion of the observer adjusts over time so that they expectation of critical signals is rare. Therefore the willingness to response that a target is present becomes diminished.

Additional observations of a conservative response bias during vigilance activities was provided by See et al. (1997). In this study, participants performed a 40-minute vigil in which they were monitoring a line blinking 20 times per minute and responding to critical signals that appeared when the length of the line increased by 20%. The researchers manipulated the signal
probability between participant groups. The signal probability was as low as 5% (one critical
signal per minute) or as high as 75% (15 critical signals per minute). The results indicated a
strong effect of signal probability on response bias. The 5% signal probability condition yielded
a significantly more conservative measure of response bias compared to the 75% signal
probability condition. Since most vigilance tasks feature rare critical signals (Wickens & Hollands,
2000), the conservative response bias is an expected outcome of vigilance activity.

**Sustained Demand Theory**

The two vigilance theories previously discussed both suggest that the decrement may be
explained by the overall monotonous nature and slow-pace of maintaining constant alertness
on a single task, resulting in increased fatigue and reduced expectation of signal presence
(Baker, 1961; Welford, 1968). The Sustained Demand Theory of vigilance contests the viewpoint
that vigilance tasks induce boredom, and instead suggests that such tasks are in fact stressful
for the operator and require substantial mental effort (Warm et al., 2008). The origins of this
theory come from the notion that human attentional resources have a limited capacity (Fisk &
Scerbo, 1987; Kahneman, 1973), and when we engage in certain types of vigilance tasks,
attentional resources become depleted and performance suffers. The important part of this
theory is the *certain types* of vigilance tasks.

Parasuraman and Davies (1977) discussed the inconsistent nature of the vigilance
decrement across a variety of contexts. In their review, the authors noted that while a large
body of research displayed evidence of a vigilance decrement, there was an equitable amount
of conflicting evidence suggesting that the phenomenon was not as ubiquitous as was being
portrayed. Their review and accompanying research experiments provided support for a
vigilance taxonomy, in which the vigilance decrement is observed only under certain conditions (Parasuraman et al., 1987). The goal of this classification scheme was to account for why different task types and characteristics produced different levels and rates of decrement. The taxonomy is made up of the four following categories: 1) task type, 2) event rate, 3) sensory modality, and 4) source complexity. Each category has two sub-types (see Figure 1.2).

**Figure 1.2.** The Parasuraman and Davies (1977) vigilance taxonomy. Classification of vigilance tasks between the four categories. Task type can either be simultaneous (relative judgment where pertinent information about the critical signal is available) or successive (absolute judgment where a critical signal representation is stored in working memory). Source complexity identifies how many sources of information are being monitored (one source for single complexity or more than one source for multiple). Sense modality splits the task into either an auditory or visual task. Event rate describes as how many events (critical and non-critical) occur per unit time.
Task type can either be successive, in which the observer makes an absolute judgment about the critical signal using a representation stored in working memory, or simultaneous, in which the observer makes a comparative judgment using the information provided from the task. For example, Molloy and Parasuraman (1996) employed a semi-realistic vigilance activity that incorporated both successive and simultaneous task types. The successive discrimination was intended to represent an automation failure in an aviation setting. A portion of the search display contained a blinking square, and a critical signal occurred when the size of the blinking square changed, indicating an automation failure. The simultaneous discrimination part of the simulation required the observer to monitor the fuel gauges of the aircraft and regulate the levels to stay in an acceptable range. When the dial on the gauge reached an unsafe level, the operator would respond accordingly. Due to the increased cognitive workload of a successive task, it is much more susceptible to the vigilance decrement compared to the simultaneous task. The simultaneous task using the dial on the fuel gauge provided all pertinent visual information in the display to allow for the observer to make a decision by examining the gauge. The successive task, however, displayed only a blinking square that would change in size, meaning the observer had to maintain the target square size in working memory. This increased mental workload and allowed fewer working memory resources to be available for other tasks.

Event rate is perhaps the most crucial element predicting the vigilance decrement (Teichner, 1974). Event rate is defined as the number of events per unit time (Wickens & Hollands, 2000). The higher the event rate, the greater the vigilance decrement rate. The more events that occur per unit, the higher the mental workload of the observer, giving them less time to process each event. For example, consider an industrial inspector working on the
conveyor line, examining parts as they pass by. The faster the rate of the conveyor line, the greater the likelihood of a vigilance decrement. Research in both laboratories and in the field have found reduced decrement rates for slower event rates, as well as operator-controlled event rates (Parasuraman, 1986).

The remaining two elements within the vigilance taxonomy are sense modality and source complexity. Sense modality can be either auditory or visual and the decrement has been shown to exist in both types of tasks (Parasuraman, 1979; See et al., 1995). Source complexity refers to how many sources of information must be attended to during the task. For instance, an observer monitoring a single pressure gauge attends to a single source of information, while an air traffic controller attends to multiple sources (e.g., ground control, air traffic, weather). Research suggests that multi-source complexity leads to a greater decrement than single-source complexity (Teo & Szalma, 2011).

Additional support for the Sustained Demand Theory of vigilance comes from subjective measures of mental workload from individuals participating in a vigilance task. Measuring subjective mental workload using the NASA-Task Load Index (NASA-TLX), Warm, Dember, and Hancock (1996) observed overall high workload ratings from vigilance tasks, as well as an increase in workload assessment over time. These findings provide additional evidence for the notion that vigilance tasks are resource demanding and require mental effort to perform successfully.

If indeed vigilance tasks are resource demanding, then the vigilance decrement should be further exacerbated when additional mental resources are required for supplementary tasks. A series of studies from Helton and Russell (2011) made such an observation when they paired
a successive visual vigilance task with either an additional spatial working memory or verbal working memory task. Participants performing the additional spatial working memory task were asked to respond when a spatial cue was congruent with a prior spatial probe presented intermittently throughout the vigilance task. During the verbal working memory task, participants responded when presented with a letter from the alphabet that may have been present in a probe set of letters displayed sporadically throughout the task. The results demonstrated that both types of supplementary tasks reduced vigilance performance compared to a non-supplementary activity control condition, indicating increased demand on mental resources led to a greater vigilance decrement.

Evidence from the vigilance taxonomy and mental workload assessments support the Sustained Demand Theory of vigilance. Humans have limited attentional resources and vigilance performance requires mental effort and is stressful. The vigilance taxonomy is an important contribution to the Sustained Demand Theory of vigilance. It provides both researchers and industry personnel with operational guidelines for various vigilance tasks, as well as ways to improve performance for certain types of tasks.

**Low-Prevalence Effect**

Unlike vigilance tasks, criterion (and therefore the relative number of misses) is the only SDT component typically impacted during search tasks featuring rare targets. The following section discusses recent empirical evidence exploring sources of increased miss rates as a consequence of low prevalence searches.

Clinical radiology was the first domain to document the effect of target prevalence on detectability (Kundel, 2000). The medical research findings on prevalence effects were
somewhat inconclusive since most of the research focused on sensitivity and not error rate or
criterion (e.g., Egglin & Feinstein, 1996; Gur et al., 2003; Kundel, 2000). The first systematic
investigation of a Low Prevalence Effect (LP effect) in visual search was demonstrated by Wolfe
et al. (2005). In this study, participants searched for images of tools (e.g., a hammer) as targets
among a set of other non-tool images. All participants searched for targets in three conditions
of varying target prevalence: 50% prevalence, 10% prevalence, and 1% prevalence. Participants
were provided performance feedback at the end of each trial. The primary dependent variable
was miss rate. In the 50% target prevalence condition, miss rates across participants were
around 7%. In the 10% prevalence condition, miss rates rose to 16%. The 1% target prevalence
condition produced miss rates of around 30% (see Figure 1.3). Altering target prevalence from
50% to 1% in the Wolfe et al. (2005) study provided a fourfold increase in miss rates.

![Figure 1.3. Summarized results from the original Wolfe et al. (2005) LP effect finding: increased miss rates as a function of target probability during a visual search task.](image-url)
The dramatic increase in likelihood of a searcher missing a rare target is the hallmark of the LP effect, and has been replicated several times (Biggs, Adamo, & Mitroff, 2014; Fleck & Mitroff, 2007; Godwin, Menneer, Riggs, Cave, & Donnelly, 2015; Rich et al., 2008; Wolfe, Brunelli, Rubinstein, & Horowitz, 2013). The initial LP effect from Wolfe et al. (2005) was observed with target prevalence of 1%, but has been replicated with target prevalence up to 5%, with miss rates ranging from as low as 20% to as high as 50% (Lau & Huang, 2010; Rich et al., 2008; Wolfe et al., 2013). This phenomenon has major human factors implications since socially critical searches often have low target prevalence. For example, radiologist examine mammography images for abnormalities that may represent cancerous tumors. This task becomes especially challenging because breast cancer prevalence in the United States is roughly 0.3% across age groups (DeSantis, Ma, Bryan, & Jemal, 2014), constituting a low prevalence search. Consequently, experienced radiologists are just as susceptible to the LP effect as nonprofessional searchers (Evans, Birdwell, & Wolfe, 2013). Likewise, security checkpoints at airports search for potential harmful weapons during baggage screening. According to the TSA.gov 2014 Year in Review (2015), a total of 2,212 firearms were detected among more than 653 million air travelers, constituting a .00x% prevalence rate. Safe to say these are real world examples of low prevalence search tasks.

The ability to successfully detect a target that appears once out of every 100 opportunities proves to be a difficult task. The following section discusses various research that has explored the boundaries of the LP effect in an effort to identify the source of increased miss rates.
Correctable Responses

A study from Fleck and Mitroff (2007) provided evidence that the LP effect may be linked to motor response errors. The visual search experiment employed in this study was very similar to the one discussed in the initial Wolfe et al. (2005) study. Twenty participants took part in a visual search task looking for an image of a tool among a set of non-tool images. Half of the students were in the control condition that featured high, medium, and low prevalence blocks of 50%, 10%, and 2% target prevalence, respectively. Participants in the experimental condition were allowed the option to correct their response from trial to trial. They were informed that if they wanted to go back to the previous trial, even after executing their initial motor entry by responding “target present” or “target absent,” they could press a different key to return to the previous trial and change their response. This manipulation ameliorated the LP effect so that the miss rates in the low prevalence block for the correctable response condition reduced from 30% to roughly 10%. The authors discuss that this finding indicates that the LP effect originates from motor response error due to RTs steadily decreasing as the experiment progresses. After correctly rejecting a few dozen consecutive trials, participants essentially grow accustomed to responding “target absent.” When allowing participants to correct their response, however, this motor response error is somewhat eliminated, which provides for improved accuracy during low prevalence searches.

The Fleck and Mitroff (2007) study indicates the motor response errors are an important factor in the decreased accuracy observed in low prevalence searches. It suggests that the LP effect may not actually be due to the target prevalence per se, but instead may be explained by a speed-accuracy tradeoff in which an efficient strategy in low prevalence searches is to quickly
respond “target absent,” because that is the likely correct response. While this study provided
evidence for correctable responses solving the LP effect, this manipulation is not often
replicated. For instance Wolfe et al. (2007) and Van Wert, Horowitz, and Wolfe (2009) both
failed to replicate the ameliorated LP effect when employing correctable responses in low
prevalence scenarios. There are two explanations for this inconsistency.

First, evidence suggests that the correctable response strategy is effective only in
certain search contexts. Much like the performance decrement can be affected by task type in
vigilance activities, efficiency in visual search tasks can vary depending on the search stimuli
characteristics. The target stimuli in feature searches, for instance, contain a single
characteristic that distinguishes them from distractors (e.g., a blue square among red squares)
(Treisman & Gelade, 1980). This allows for an efficient search in which the target will likely be
detected very quickly. Search tasks can become more complicated and inefficient when the
target shares the same basic properties as the distractors (e.g., a T among Ls). This type of
search activity requires inefficient processing of the displays since items must be attended to
resolve the spatial configuration of the constituent horizontal and vertical elements (Wolfe et
al., 1990). Rich et al. (2008) examined the LP effect across both feature and serial searches,
while also investigating the efficacy of a correctable response intervention. The results
indicated that while the LP effect is evident in both types of search tasks, the correctable
response manipulation was only effective at improving the LP effect in efficient, feature
searches and not inefficient, spatial configuration searches. The authors concluded that the
source of the LP effect differs depending on the type of search task. Increased miss errors in an
efficient search task are likely caused by increased motor response bias, while increased miss
errors in an inefficient search do not share that source of the error since responses tend to be slower and more variable (Palmer, Torralba, Horowitz & Wolfe, 2011). This interesting result provides a partial explanation as to why the correctable response intervention was not easily replicated. Vigilance research examining enhanced variation in low prevalence tasks may provide an additional explanation.

The second explanation for the inconsistent efficacy of the correctable response strategy comes from Methot and Huitema (1998), who varied target prevalence levels in a vigilance task. In this study, the researchers were interested in the impact of varying signal probability (i.e., target prevalence) on differences in individual performance in a simultaneous vigilance task. In the experiment, participants monitored a replicated pressure gauge for 120 minutes. Critical signals occurred when the arrow indicator entered into a “danger zone” designated on the gauge. Researchers observed signal detection accuracy across three between-groups conditions with eight participants in each group: a high signal probability condition (12% probability of all arrow indicator changes were deemed critical signals), medium signal probability (4%), and low signal probability (1%). The results in this vigilance task align with the LP effect, in which miss rates increased as a function of signal probability (6% miss rate for high, 13% miss rate for medium, and 27% miss rate for low). The primary dependent variable in this study, however, was within-group variance of each participants’ hit rate decrement coefficient. The researchers reported that the coefficient variance increased as signal probability decreased. Specifically, the coefficient variance for the high signal probability condition was 92% lower than the hit rate slope variance for the low signal probability condition (see Figure 1.4). The authors discuss that a low critical signal or target probability
contributes to enhanced individual differences in performance. While the variance data was not provided in Fleck and Mitroff (2007), this inconsistency may provide additional explanation for the unpredictable nature of the correctable response strategy as an effective intervention for alleviating the LP effect.

![Figure 1.4](image)

**Figure 1.4.** Decrement slope results from Methot and Huitema (1998) for individual participants across signal probability conditions. Within-group variance decreased as signal probability increased.

Methot and Huitema (1998) discuss the potential sources of such extreme within-group variation. The authors mention Expectancy Theory as a plausible explanation for their results. As previously discussed, Expectancy Theory emphasizes the role of a criterion shift towards a more conservative approach within vigilance tasks (Baker, 1961). As time passes during a vigilance task and critical signals seldom appear, the expectation for a critical signal decreases, diminishing the likelihood of a “target present” response and increasing the likelihood of a “target absent” response. Methot and Huitema (1998) also emphasize that a decreased signal probability leads to greater uncertainty of the signal presence, which leads to greater variation in detection accuracy across individuals.
Expectancy Theory has provided valuable insight into human information processing during a vigilance task. When critical signals appear rarely over the course of a sustained attention task, the operator becomes biased towards assuming the signal will not appear. Visual search studies on low target prevalence have taken this information processing mechanism into consideration when investigating the potential sources of the LP effect.

**Burst of High Signal Probability**

Perhaps the most effective method for reducing the LP effect in visual search studies has been directly related to the principles proposed by Expectancy Theory. Wolfe et al. (2007) performed a series of studies attempting to improve target detection accuracy in low prevalence searches. One study addressed the conservative criterion shift often associated with the increased miss rates for low prevalence targets. In this experiment, participants examined realistic images of semitransparent x-ray baggage stimuli. Targets appeared as guns or knives within the baggage images. The session was divided into five blocks of 200 trials each, in which targets typically appeared twice every block (1% target prevalence), without performance feedback. The experimental manipulation was introduced as occasional bursts of 40 trials with 50% target prevalence with immediate performance feedback after each trial. Visual search performance was compared to a control condition with no burst of higher target prevalence trials. The results revealed a substantial reduction in miss rates in the high prevalence trials burst condition. By implementing a burst of high prevalence trials, miss rate was reduced from around 45% in the control condition to around 20% in the experimental condition. The authors reasoned that the LP effect is diminished using this method because interspersed high prevalence trials allow individuals to recalibrate their response bias towards a more liberal
criterion. This suggestion is supported by an increase in false alarms reported by the authors. Participants became more willing to respond “target present” when the search task was accompanied by occasional bursts of higher target prevalence trials. This effect has been replicated, even among trained airport security screeners (Wolfe et al., 2013).

While the successful attempt to recalibrate criterion using bursts of high prevalence trials is a recent development in low prevalence visual search studies, this type of intervention has been observed in vigilance for some time. Improving detection of rare signals by injecting artificial signals throughout a vigil has been seen to be an effective intervention in auditory vigilance tasks (Wilkinson, 1964), visual detection vigils (Mukrell, 1975), and prolonged simulated sonar operations (Mackie, Dennis, & Smith, 1994). All three of these studies, along with the high prevalence bursts observed in both the Wolfe et al. (2007) and Wolfe et al. (2013) visual search studies, reported an increase in false alarms as a result of integrating additional target present events. Furthermore, several studies examining the impact of rare critical signals or targets have emphasized that the increase in miss rates is primarily a function of the conservative criterion shift and has not been accompanied with a decrease in sensitivity (Mukrell, 1975; Wolfe et al., 2007). This indicates that an observer’s ability to successfully discriminate signal from noise is not always impacted in low probability environments. Instead, the criterion shift appears to be the source for the LP effect, meaning the explanation provided by Expectancy Theory bridges the gap between vigilance and low prevalence visual search performances.
Low-Prevalence Effect and Vigilance

Clearly the vigilance decrement and the LP effect have similar theoretical traits. Since the current project explores elements shared by both activities, it is important to further understand what they have in common as well as how they differ. Figure 1.5 summarizes the important similarities and differences.

Figure 1.5. Summary of primary similarities and differences between vigilance tasks and low prevalence visual searches.

The theoretical traits shared between vigilance and low prevalence searches are not limited to Expectancy Theory. There is evidence to suggest the mechanisms involved in both Arousal Theory and Sustained Demand Theory are also involved in low prevalence searches. In one of the low prevalence search experiments from Wolfe et al. (2007), participants were asked to rate their subjective levels of alertness and fatigue during each rest break between blocks.
approximately once every 20-30 minutes during the 2-4 hour session). The results revealed that subjective alertness ratings declined significantly over time, indicating participants became more fatigued as the session progressed. As previously mentioned in the Arousal Theory discussion, vigilance activities are typically accompanied with reduced brain activity and increased perceptions of boredom (Milošević, 1978; Scerbo, 1998). Even when enforcing rest breaks throughout a low prevalence search task, participants’ arousal may be diminished. Sustained Demand Theory, which suggests vigilance tasks are stressful and mentally demanding, has subtly presented itself in low prevalence studies. While no low prevalence studies have reported subjective mental workload ratings, Wolfe et al. (2007) mentioned that participants voiced frustration regarding their experience (also a sub-scale of the NASA-TLX, though not used in this study). An activity involving over 2,300 visual search trials in which only two dozen contain a target lends itself to increased frustration and mental demand. Some vigilance research has demonstrated high ratings on the frustration subscale of the NASA-TLX mental workload assessment (e.g., Sawin & Scerbo, 1995). The similarities between Arousal Theory and Sustained Demand Theory suggest that the vigilance and low prevalence search share multiple fundamental aspects of human attention processing.

The current project places these fundamental similarities at the forefront. Of primary interest are the (1) increased miss rates, (2) increased mental demand, and (3) reduced task engagement as potential outcomes of low signal probability environments. As it is, miss rates often become the primary dependent variable in both vigilance and low prevalence research, since too few false alarms are committed to allow for a reliable measurement of sensitivity and criterion (e.g., Fleck & Mitroff, 2007; Kunar, Rich, & Wolfe, 2010; Methot & Huitema, 1998;
Wolfe et al., 2007). Since both measures of SDT provide valuable insights, and because calculated corrections for zero false alarms do exist (Macmillan & Creelman, 2005), they will continue to be discussed and emphasized.

**Task Manipulations**

Numerous manipulations and observations have been made in an effort to explore the boundaries of low signal probability tasks. While the extensive research has examined the limits of the external stimulus characteristics (e.g., event rate, stimulus salience in vigilance tasks, stimulus salience in low prevalence search tasks, and signal probability as seen in See et al., 1995, Teichner, 1974, Biggs et al., 2014, & Loeb and Binford, 1968, respectively; see Figure 1.6), other research has examined manipulations intended to alter the internal mindset of the observer in an attempt to improve performance.

![Figure 1.6](image)

**Figure 1.6.** Diagram displaying how various manipulations may impact the vigilance decrement or low prevalence effect. Performance in low signal probability tasks has shown to be vulnerable by either adjusting stimulus characteristics or altering the observer’s mindset.

Since the vigilance decrement was discovered about 60 years before the LP effect, thus providing a more abundant research archive, most of the literature discussed revolves around
vigilance activities. Each manipulation discussed, however, attempts to address fundamental problems of low signal probability tasks.

**Knowledge of Results**

In a series of studies, Warm, Riechmann, Grasha, and Seibel (1973) demonstrated the power of feedback when it comes to improving reaction time performance during a vigilance task. Participants performed a successive visual vigilance task where they pressed a button whenever the critical signal was presented. The experiment consisted of two primary groups: a knowledge-of-results (KOR) group, who received evaluative feedback displaying their reaction time performance as either improving or worsening throughout the task, and an acknowledgement control group, who received response feedback without any information regarding their performance. As the session progressed, reaction times increased in both groups, which is typical of a vigilance task. The reaction times of the KOR group, however, were consistently faster than the acknowledgement group.

More recent research has examined how different types of KOR can be implemented to improve target detection performance. Szalma, Hancock, Dember, and Warm (2006) performed a series of studies examining how the format of the KOR feedback can be varied. The experiments tested KOR format in four conditions: KOR-hits, KOR-false alarms, KOR-misses, and KOR-composite of all three previous formats. Participants monitored a complex, multidimensional visual display for rare target signals (approximately 6% prevalence) during a 30-minute vigil. During the task, participants received visual feedback consistent with the condition they were in (e.g., KOR-hits participants received only correct detection feedback and KOR-miss participants received feedback only when they failed to detect a critical signal). The
results indicated that all four types of KOR feedback conditions had superior perceptual
sensitivity performance compared to the no-KOR feedback control condition. In particular, the
KOR-hits and KOR-composite conditions produced higher levels of perceptual sensitivity
throughout the task.

Mackie et al. (1994) performed a series of experiments testing the impact of
performance feedback accompanying artificial signal injection in a simulated sonar operations
task with low signal probability. Signals in this task represented potential enemy threats. As
previously discussed, artificial signal injection or interspersing high prevalence target trials in a
low prevalence task can improve detection performance. The experimenters tested the effect
of artificial signal injection with and without KOR. The results indicated that artificial signal
injection improved target hit rates in both conditions, but was improved more with the
inclusion of KOR. The researchers also report that response bias was shifted towards a liberal
criterion when KOR was provided, leading to an increase in false alarms. The researchers
emphasize that false alarms are an acceptable error in a sonar watch environment, since a false
alarm is typically much less disastrous than a miss.

The three previous studies demonstrate the benefits of providing immediate
performance feedback to observers during attention tasks with rare critical signals. By providing
performance feedback immediately after a decision is made, the observer not only has valuable
information regarding their performance throughout the task, but receives confirmation on
whether or not their task strategy is effective. This is why the KOR format is so important. The
KOR-composite and KOR-hit conditions are effective because they are linked to the actual
dimensions of a critical signal, while the KOR-misses have no such association (Dittmar, Warm,
& Dember, 1985; Szalma et al., 2006). The observer learns what to look for and can therefore evaluate their strategy and ultimately improve their performance.

**Instructions**

Participant experience in a vigilance task can be modified by how the instructions are framed. Sawin and Scerbo (1995) examined the effects that two types of instruction have on vigilance performance and workload assessment. Participants were split into two groups. Observers in one group were given the standard vigilance instructions and asked to pay close attention to the critical signals and respond accordingly when they identify the target “flickers.” The second group of participants were told that the vigilance activity was actually a relaxation exercise and that they should leisurely monitor the occasional target “flickers” on the display and respond accordingly. Both groups executed a successive visual discrimination during a 30-minute vigil. The results indicated that the type of instructions did not impact target detection performance, but the “relaxation” instructed group rated their experience as having less perceived mental workload compared to the standard instructions group. Specifically, the relaxation group rated the vigilance task as being significantly less frustrating on the NASA-TLX subscale compared to the standard instruction group as well. As stated in previous sections, vigilance activities are resource demanding and require high mental workload, as suggested by the Sustained Demand Theory of vigilance (Warm et al., 2008). This finding suggests that the subjective level of mental demand and frustration can be adjusted by framing the activity using favorable instructions.

The Sawin and Scerbo (1995) results indicate that perceived mental workload of an operator may be modified without changing any of the physical characteristics of the task. This
is an indication that the apparently stressful and frustrating experience typically induced by a prolonged vigil can be reduced and perhaps avoided by how the experience is framed and presented. While target detection performance between the two groups did not differ, these findings are still valuable. It is an indication that operator motivations may be modified by how the activity is presented and what parts of the task are emphasized. Even with the null performance differences, mental workload perceptions are still a valuable part of vigilance task engagement.

**Mental Breaks**

Ariga and Lleras (2011) examined an experimental manipulation intended to address the demanding nature of vigilance tasks. The researchers hypothesized that, because sustained attention on a particular task results in a depletion of limited attentional resources, allowing observers brief mental breaks would cause attentional resources to “recover” and target detection performance to improve. The experimenters tested this hypothesis using two groups of observers, a switch-group and no-switch-group, both performing two tasks throughout their participation. The primary vigilance task was a successive visual discrimination during a 40-minute vigil. Before the start of the vigilance exercise, participants in both groups were presented with a series of digits that they were asked to memorize and told they would be presented with a recognition set of digits later on. The secondary digit memory task was presented either intermittently during the vigil for the switch-group, or presented at the end of the vigil for the no-switch-group. The results of the study showed that target detection performance among the switch-group was greater than the no-switch-group.
Additional evidence that brief mental breaks can improve vigilance performance was observed by researchers investigating the consequences of complete rest breaks during a vigilance activity. In a study from Ross, Russell, and Helton (2014), participants performed a 40-minute visually successive vigil. The experiment had two between-subjects groups: a continuous control group that did not rest during the task, and a rest-break group that was given two 1-minute breaks after 20 minutes and 30 minutes into the vigil. Perceptual sensitivity over the course of the vigil was recorded for both groups and was the primary performance metric of interest. The results indicated that the rest-break group had a shallower perceptual decrement than the control group over time.

The findings from the two previously discussed studies are supported by the Sustained Demand Theory of vigilance. Sustained Demand Theory emphasizes the role of attentional resources during a vigilance task. Vigilance requires mental effort for a prolonged period of time. Therefore allowing participants to take brief mental breaks from a vigilance task, as in Ariga and Lleras (2011), or allowing them to rest for a minute or two as in Ross et al. (2014), allows the limited attentional resources to “replenish” in some sense when returning to the vigilance task.

**Rewards**

Sipowicz, Ware, and Baker (1962) were one of the first to explore the performance enhancements when vigilance operators are rewarded (or punished) with monetary incentives. Their experiment investigated the performance differences among four groups of participants. The first group was provided with an initial monetary payment and lost money for each signal they missed (reward group). The second group received knowledge-of-results for each missed
signal (KOR group). The third group received both of these manipulations (reward + KOR group). The fourth group was the control that did not receive any of these manipulations. All four groups participated in an 3-hour vigilance task and the researchers observed their miss rates. The reward + KOR group had the lowest miss rates, followed by the reward group, KOR group, then the control group. There was no significant difference in miss rate performance between the reward group and KOR group. These findings indicate that monetary incentives, when tied to performance and paired with KOR, can enhance target detection during a vigilance activity.

A more recent study from Tomporowski and Tinsley (1996) examined how monetary incentives may impact vigilance performance among both younger and older participants. In their first experiment, the researchers split participants into three groups. The first group was made up of older adults (mean age = 63) who were paid a flat rate for their participation. The second group was made up of younger adults (mean age = 21) who were also paid a flat rate for their participation. The third group was also younger adults but they were not paid for their participation. All three groups completed a successive visual discrimination task during a 60 minute vigilance exercise. The hit rates remained stable throughout the vigil for both paid older and younger participants. The unpaid younger group, however, displayed both lower overall target detection and a vigilance decrement as time progressed. The results indicate that monetary incentives, even when not tied to performance, can be an effective motivator for improved performance during a vigilance task. In their second experiment, the researchers again tested a group of both paid older adults and unpaid younger adults, but this time they removed any mention of monetary incentive to older adults until after the participants had completed the activity. The target detection performance among older adults remained stable.
and high throughout the task, while the performance of the unpaid younger adults consistently
deteriorated over time. The results of this experiment indicate that the benefits of monetary
incentives provided in a vigilance task may not be as clear after all.

The results from Tomporowski and Tinsley (1996) demonstrate that monetary reward
may be an effective manipulation to enhance operator mindset during a vigilance task. When
examining these studies further, however, there are two important components that must be
addressed. First, the results of the second experiment with paid older adults actually indicates
that monetary payment may not make a difference in vigilance task performance. Not only
were those participants unaware that they were being paid until after they finished the activity,
the authors describe how a number of them were interested in participating because they
wanted to see how they would perform on tasks assessing their attention and memory skills. A
number of these participants also declined monetary payment for their participation. Second,
while the unpaid younger adults were not rewarded with money for their participation, they
were incentivized to participate. The authors describe how the younger adults were recruited
from the university undergraduate population, and they were required to participate in
department research activities or suffer the consequences of a coursework penalty. While they
did not receive monetary payment for their participation, they were still being incentivized to
participate.

The results from both Sipowicz et al. (1962) and Tomporowski and Tinsley (1996)
demonstrate that there likely exists a benefit to vigilance performance when monetary rewards
are provided for participation, particularly when incentives are tied to performance. It seems
that incentives provide a motivational boost and encourage the operator to do well at the task.
Motivation

It has been established that sustained attention can be modified and even improved with manipulations intended to adjust the operator mindset (Ariga & Lleras, 2011; Helton & Russell, 2015; Ross et al., 2014; Tomporowski & Tinsley, 1996). Some of these manipulations have utilized external rewards and punishments as a mechanism to adjust operator mindset and improve performance. While the results suggest using monetary incentives can be an effective measure to improve performance, other research has examined some of the detriments of using external rewards and punishments (e.g., monetary incentives) as a motivator. For instance, research from psychologist and economist Dan Ariely has found there can be unintended performance consequences when using money as an incentive for task participation (Ariely, Gneezy, Loewenstein, & Mazar, 2009; Heyman & Ariely, 2004). When using external rewards and punishments as a motivator, it can be interpreted as an attempt to control and manipulate behavior (Pink, 2011; Ryan & Deci, 2000). This tendency can actually reduce task performance and has been seen to be prevalent in particularly boring and uninteresting tasks.

Heyman and Ariely (2004) were one of the first to empirically demonstrate the unintended consequences of monetary incentives. The researchers had three groups of participants perform a “mind-numbing task devoid of [inherent interest]” by dragging a ball across a screen for several minutes (p. 790). The first group of participants did not receive any monetary payment. The second group received a small monetary payment. The third group received a moderate monetary payment. The results indicated that performance was highest when participants were either paid a moderate sum or no money at all. When participants
received a small payment, performance was the lowest. The important takeaway from this study is that payment for performing a boring task that itself has no inherent interest can actually be detrimental to performance if the payment is small.

This finding is especially important for vigilance research, because there is strong evidence to suggest that vigilance tasks are boring and devoid of inherent interest. In a study from Upadhyay and Singh (2013), participants’ level of intrinsic motivation (i.e., the motivation to perform a task because the task itself is interesting and enjoyable) was assessed before and after a 40-minute successive visual discrimination vigilance exercise. The results indicated that reported levels of intrinsic motivation in the individuals were significantly lower after they participated in the vigilance activity. This is a strong indication that the sustained attention requirement of vigilance tasks has adverse effects on an individuals’ intrinsic motivation.

Since tasks requiring sustained attention for rare targets could be viewed as incredibly boring and devoid of any intrinsic motivation, they are perhaps quite susceptible to the unintended consequences of performance and motivation deficits from using an extrinsic incentive like monetary reward. While some of the research previously discussed does suggest that monetary incentives can improve performance, recent research in behavioral economics has established that it is important to pay an adequate amount to achieve the optimal level of performance and avoid performance deficits (Gneezy, Meier, & Rey-Biel, 2011; Gneezy & Rustichini, 2000). Regardless, what may be more powerful than providing monetary incentives for superior performance would be applying certain manipulations to a vigilance task that modify operator mindset in a more powerful way. For instance, manipulating the task and
mindset so the vigilance activity becomes more interesting and enjoyable because it now has
intrinsic value.

**Self-Determination Theory**

An influential theory of intrinsic motivation comes from social psychologists Edward Deci and Richard Ryan. Self-Determination Theory is a theory of human behavior and psychology that proposes human beings seek growth in their endeavors, want to be challenged in their tasks, and have an inherent self-motivation to succeed (Deci & Ryan, 1985). The theory is comprised of three primary elements: autonomy, competence, and relatedness (Ryan & Deci, 2000).

Autonomy is the idea that individuals desire an internal control of their own behavior as opposed to a feeling of being externally controlled. Individuals want freedom for their actions and choices. This is why monetary incentives can have unintended consequences in social settings like volunteering to participate in a research experiment (Heyman & Ariely, 2004). The individuals no longer perceive themselves to have autonomy and instead feel that they are participating in a task to earn a reward, and a paltry reward at that.

Competence is the sense of mastery an individual obtains when engaging in a challenging task in which they have the desire and ability to improve their performance. Some research examining the psychological state of *flow* (Csikszentmihalyi, 1990) reflects the power of competence as a component of self-determination. Flow occurs when an individual becomes so involved and engaged in an activity that he or she loses a sense of time and space; it becomes a perfect balance of the individual’s skill level and the challenge of the task (Csikszentmihalyi, 1990).
Lastly, relatedness is the idea that an individual wants to feel meaningfully connected to other people rather than feeling isolated and unimportant. It can be a powerful motivator when one feels that they have a sense of purpose, and that they are contributing to the greater good of humanity (Pink, 2011).

The elements of Self-Determination Theory have been thoroughly researched and explored over the past few decades. Little effort has been made, however, to apply these principles to vigilance activities. The following section highlights the only two existing studies that have examined the possible benefits of facilitating intrinsic motivation in a vigilance task.

Dember, Galinsky, and Warm (1992) were the first researchers to overtly apply intrinsic motivation principles to improve performance in a vigilance task. Specifically, the researchers were interested in the role of autonomy in vigilance task performance. In their study, participants performed a 50-minute successive vigil of lines blinking on a display. The critical signal was defined as a blinking line with a shorter display duration. There were two primary groups in the study: participants in the autonomy condition were allowed to choose their own difficulty level (easy or hard), while participants in the control condition were assigned to one of the two difficulty levels. The difficulty of the vigilance task was defined by the display duration of the critical signal. The critical signal had a much longer display duration difference from neutral events in the easy difficulty compared to the hard difficulty condition. The results of the study showed that the autonomy condition participants had a significantly shallower decrement over time compared to the control group, regardless of difficulty level. This is an indication of the benefits of intrinsic motivation, and specifically the role of autonomy, within a vigilance task. Furthermore, the manipulation in this study was independent of the task itself and only applied
to the operator mindset. There was no reward, punishment, or even a break from the task that improved performance in the autonomy group. Instead, the only difference was the presence of freedom to engage in the task based upon their own level of interest.

A study from Szalma and Hancock (2006) replicated the findings from Dember et al. (1992), also using a visual successive task. In this study, difficulty of the vigilance task was defined by the event rate. A low event rate was used for the easy difficulty level and a high event rate was used for the high difficulty level. Participants in the autonomy condition showed a greater hit rate throughout the task, while participants in the no choice condition exhibited a significantly greater number of misses throughout the task.

While intrinsic motivation principles seem to have benefits on sustained attention performance, there has yet to be a thorough exploration into how other areas of intrinsic motivation, if facilitated properly, might improve performance.

**Potential for Improvement**

At this moment, it might appear that the vigilance decrement is susceptible to improvement from various manipulations. From knowledge of results to task instructions, it seems that the limitation of prolonged sustained attention might be improved with the proper precautions, and that system operators can be put in a position to succeed with appropriate rewards or occasional breaks during task engagement. While these manipulations might improve vigilance performance, it is important to examine the bigger picture with regard to designing for the limitations of human sustained attention.

Hancock (2013) wrote that the vigilance decrement is “primarily iatrogenic in origin;” it is a technology-induced side effect of creating tasks that require prolonged, sustained attention.
He proposed that performance decline in vigilance tasks is the unintended consequence of poor task and display design. He suggests that, moving forward, human factors professionals make a better effort to design tasks that account for the natural boundaries of human cognition, specifically the limitations of sustained attention. While Hancock makes a compelling case, one could argue he stops short in regard to the burden on human factors professionals and their responsibility for successful task design.

While the goal of any human factors specialist is to create systems and activities that accommodate the limitations and capabilities of the human operator, I believe it is also our responsibility to create systems that are also satisfying and enjoyable. The successful task design will allow the human user to accomplish their goals while also creating a more enjoyable experience in which the three elements of intrinsic motivation (autonomy, competence, and relatedness) are allowed to thrive. A particularly compelling type of human recreation to examine for inspiration towards promoting intrinsic motivation is that of games. People play games because they encourage all three elements of intrinsic motivation (Przybylski, Rigby, & Ryan, 2010), and these principles could be applied to vigilance tasks so they not only improve performance, but also become more enjoyable.

**Gamification**

The video game industry has become one of the most powerful industries in the world. In 2013, consumers spent over $21 billion on games (Entertainment Software Industry, 2014), which is more than all other types of at-home media entertainment combined (i.e., film, television, and radio; Statistica, 2014). By 2018, the gaming software market is expected to reach a $100 billion value as an industry (Brightman, 2014). Furthermore, in the United States,
users spend on average two hours a day playing mobile games (NPD Group, 2014). This evidence suggests video games possess certain characteristics that make them engaging and enjoyable that may be applied to other tasks that take a prolonged period of time, such as vigilance tasks. The emerging process of enhancing activity experience by introducing gaming elements has become known as gamification.

Gamification is the application of gaming elements (e.g., achievement rewards, levels, leaderboards; see Table 1.1 for a summary) to non-gaming contexts (Deterding et al., 2011; Hamari et al., 2014). The primary goal of gamification is to enhance the task experience while perhaps improving task performance (Deterding et al., 2011; Hamari et al., 2014). While the term “gamification” is relatively new, the practice of enhancing an activity, whether intrinsically boring or exciting, using game-like features has been around for several decades (Michael & Chen, 2005). “Serious games,” for instance, were first developed during the Cold War as an activity to create a learning platform utilizing gaming characteristics to educate military personnel on war strategies (Abt, 1970). While the primary goals of gamification and serious games are to enhance the learning experience using game-like features, serious games utilize an overall gaming scenario while gamification applies specific features to non-game scenarios (Deterding et al., 2011; Ma, Oikonomou, & Jain, 2011; Michael & Chen, 2005).
Table 1.1. Summary table of common gaming elements used in gamification (Deterding et al., 2011; Hamari et al., 2014; Kapp, 2012; McGuire & Jenkins, 2008; Zichermann & Cunningham, 2011).

<table>
<thead>
<tr>
<th>Gaming Element</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Point System</strong></td>
<td>Scoring structure intended to elicit a specific behavioral outcome; flexible method that can be utilized in a variety of ways (e.g., participants can “cash in” points for rewards or points can be used as an objective measurement of participant performance)</td>
</tr>
<tr>
<td><strong>Leaderboard</strong></td>
<td>Visible ranking of participant performance; encourages a competitive social component to participation</td>
</tr>
<tr>
<td><strong>Badges</strong></td>
<td>Visual representation of a reward for task completion or reaching a certain level of achievement</td>
</tr>
<tr>
<td><strong>Levels</strong></td>
<td>Discrete and sequential changes to either the overall game scenario or increases in task difficulty; demonstrates progression through the activity</td>
</tr>
<tr>
<td><strong>Challenges</strong></td>
<td>Additional user tasks encountered at various points of the activity; intended to provide additional interest and add meaning</td>
</tr>
<tr>
<td><strong>Storytelling</strong></td>
<td>Creating a theme for the activity; gives a sense of purpose and meaning to the experience, providing important context and direction</td>
</tr>
</tbody>
</table>

This section is split into three parts. The first part discusses how gamification is supported by theories of motivation (specifically Self-Determination Theory). The next part identifies numerous applications of gamification, both in industry and academics. And the final part explores three specific gaming elements that might benefit performance in a vigilance search activity.

**Why Gamification Works**

Before applying gaming elements to an activity, it is important to first understand what it is about games that promotes intrinsic motivation. By establishing what gaming elements might enhance certain psychological outcomes, the application of gaming elements may then
be more effective. Przybylski et al. (2010) provide a motivational model of video game engagement that informs what aspects of gaming contribute to the motivation to play them. This theory explains the drive people have to play video games through the lens of Self-Determination Theory. Through a series of empirical studies, the authors have identified how certain aspects of video games promote the specific elements of Self-Determination Theory. Games that allow exploration encourage a sense of autonomy and freedom, which enhances the video game experience. Common gaming elements like levels and challenges provide the player with increased competence by improving his or her skill, creating a more immersive experience, and giving clear indications of progress. Finally, relatedness may be enhanced by playing video games that allow users to connect with other players in a shared experience or simulating a meaningful experience. Online gaming, in particular, has encouraged this type of activity, with players being allowed to work on a team with other players to accomplish shared goals. This summary of the motivational model of video game engagement gives important context towards understanding how gamification, when designed effectively, enhances activity experience.

Richter, Raban, and Rafaeli (2015) provide additional theoretical support for gamification and the potential towards enriching motivation. By describing several theories of motivation, Richter et al. (2015) argue that well-designed gamification practices can potentially modify overall psychological well-being within a given activity. The authors emphasize that while intrinsic motivation is enhanced, users may also experience increased self esteem, self efficacy, and approach their full potential. These powerful outcomes might be achieved when applying gamification in a well-designed approach by aligning certain gaming elements with
specific psychological outcomes. Richter et al. (2015) continue to discuss how gamification enhances the humanistic desire to be challenged, improve skills, and succeed. When certain gaming elements are meticulously applied early in the design process, they can yield positive motivation and performance outcomes on the end users. The following section discusses some research that supports the argument from Richter et al. (2015) for implementing gaming elements in an effort to enhance the activity experience.

**When Gamification Works**

Over the past few years, gamification has become a prevalent application in industry and the focus of research in academics. A Google Trends report during the time this project is being conducted reveals it had roughly zero searches until early 2011, when the term “gamification” became prevalent relative to the total Google search volume (see Figure 1.7). As an industry, gamification is expected to approach a value of $3 billion by 2016 (Peterson, 2012). Gamification principles have been applied to a variety of domains over the past several years, including health, education, and corporate operations. Even consumer products like the FitBit utilize gaming elements like badges and leaderboards to enhance personal fitness activity (FitBit, 2015). This engaging interaction has helped FitBit become the industry leader among wearable fitness products (Kelly, 2014).
Figure 1.7. Google Trend for “gamification” as a search term. The y-axis represents total searches for a term relative to total Google searches. Interest in the topic has increased since 2010.

Outside of consumer products, corporations have begun to explore the benefits of gamification within their operations. For instance, Canadian airline company WestJet Airlines implemented gamification within their Enterprise Resource Planning system (Kanaracus, 2014). Employees now have the opportunity to earn points and badges for completing tasks (e.g., submitting expense reports) within the system. While the results of the implementation have not been revealed, WestJet Airlines provides an example of how companies are implementing gamification.

True Office, a technology company, implemented gamification within their Information Technology compliance training course (Baxter, Holderness, & Wood, 2014). The corporate training course was intended to educate employees on best practices for company security activities, like updating login information and not sharing passwords. The company added
elements like a storyline and progressive milestones so that employees felt more engaged during the training. The company was interested in observing the benefits of gamification in this context, so they compared post-training knowledge results with a control training group that did not receive gamification. The results indicated that gamifying the compliance training module led to improved knowledge and increased subjective levels of enjoyment for the training activity (Baxter et al., 2014).

While applying specific gaming elements in non-game contexts is a recent development, game-based learning in education has been examined for at least the last couple of decades (Mishra & Foster, 2007). Research on game-based learning in education has examined the benefits and drawbacks of modifying course exercises to include game elements in an effort to increase motivation and improve learning and retention (Mishra & Foster, 2007). The following two summaries relate to gamification in education.

Online courses in particular have long suffered from reduced student motivation and engagement (Waschull, 2005). A study from Domínguez et al. (2013) investigated gamification applied to an online communication sciences course. The primary objective of the course was to teach undergraduate students how to use various computer programs (e.g., Microsoft Excel, Word, and PowerPoint). Students enrolled in the course were split into two groups: one gamified and one control. Students in the gamified group had the opportunities to earn trophies, complete challenges, and view a leaderboard representing student rankings of scores on course exercises. The gaming elements were only applied to optional course exercises that were also available to students in the control group. The results of this study indicated students in the gamified group outperformed the control group students on certain tasks, such as
practical application of the lessons and completing simple tasks. The control group students, however, outperformed the gamified students on written examinations that assess deeper understanding of the material. A similar outcome occurred in a study from Daubenfeld and Zenker (2014).

In the Daubenfeld and Zenker (2014) study, gamification was applied to an undergraduate physical chemistry course. The researchers assessed students’ objective performance scores and subjective attitudes towards the course. Both metrics were compared to the metrics of the same course from a previous term that used the same core lesson plan but did not contain any gaming elements. Throughout the gamified course, students progressed through the course curriculum via the “learning pathway through the island of phases” as part of the overall narrative theme the course instructors introduced at the beginning of the term. Quizzes and exams were gamified and students completed “labyrinths” (i.e., word problems) at the end of each “level” (i.e., chapter). The results of this study found that students in the gamified chemistry course had higher levels of course satisfaction and reported being more motivated in the course. In terms of objective performance metrics, the gamified course also had a lower failure rate and improved memory scores compared to the control course. The results also indicate, however, that the gamified group did not have higher-level thinking scores than the control. The authors elaborated that the students in the gamified course had difficulty transferring knowledge towards solving more difficult problems. The results of this study, along with the previous Dominguez study, support the idea that gamification can be beneficial in the right circumstances, but can inhibit deeper understanding of material. This outcome is supported by criticisms of game-based learning.
Mishra and Foster (2007) performed an extensive literature review on game-based learning practices. After reviewing the literature, they claim that a potential weakness of game-based learning is that it leads to “superficial” learning that does not produce improved problem solving abilities and deeper understanding of the material. The authors suggest that game-based learning practices should maintain focus on the material to be learned and not shift focus towards the otherwise irrelevant gaming elements. Of equal importance, they claim, is understanding the audience that will be interacting with the gamified activity. Both of these suggestions provide important principles towards developing best practices for implementing gamification in an activity. Failure to do so, the authors claim, results in the unintended consequences observed in both the Domínguez et al. (2013) and Daubenfeld and Zenker (2014) examples.

In the Domínguez et al. (2013) study, one of the gaming elements used was a leaderboard where students could observe where their course performance ranks among their peers. In the post-course feedback from students, the authors described how numerous students found the leaderboard to be discouraging as the semester progressed. Students also complained it was not a fair representation of the knowledge level among the class. This outcome demonstrates the importance of understanding the audience when implementing gaming elements into an activity. The researchers anticipated that a leaderboard would be a source of motivation for the students (and for many, it was), but did not anticipate the unintended consequences of discouragement and frustration among other students.

In the Daubenfeld and Zenker (2014) study, one of the things the authors concluded about why students did not demonstrate a deeper understanding of the course material was
because their gamification approach used too much of a fantasy narrative and not enough real-life examples. For instance, the objective of the course was for the students to make their way through the “island of phases” towards the “Castle of the Math Mage.” The authors suggested that this approach is beneficial when students are learning the basics and fundamentals, but more practical approaches are more appropriate for advanced learning.

Both of these studies made subtle mistakes when designing for gamification in a learning environment. The learning outcomes were affected when the researchers unintentionally ignored either the demanding course content or individual differences among the students enrolled. The outcomes of these studies, however, still provide valuable information in regard to designing for gamification. Additional evidence suggests that the outcomes can be favorable when gamification is prudently applied.

A study from Denny (2013) implemented a single gaming element in an e-learning environment. The researchers examined whether badge-based achievements as a gaming element would increase student interaction in an e-learning environment. The course of over 1,000 students were randomly assigned to two groups: gamified or control. Students in the gamified group had the opportunity to earn badges for certain contributions to course discussions and uploading study questions for the rest of the class. Badges were visible and detailed on a separate webpage and appeared similar to sports medals (e.g., a silver medal) with a graphic representative of the task that was completed (e.g., a notebook for sharing notes). The researchers were primarily interested in assessing how online activity differed between the two groups. The results indicated that students in the gamified group answered more questions and spent an overall greater number of minutes on the course page than the
control group. The results of this study demonstrate that careful application of gaming elements can have positive outcomes.

The examples discussed thus far have focused on gamification applications to personal fitness, corporate operations, and education domains. But if the goal of the current project is to observe potential benefits of gamification in a vigilance environment, none of these domains quite fit the monotonous nature of vigilance activity. In a study from Flatla, Gutwin, Nacke, Bateman, and Mandryk (2011), however, researchers applied gamification to the mundane process of Human-Computer Interaction (HCI) calibration tasks (i.e., adjusting the computer settings to accommodate for human limitations). The authors discuss how HCI calibration tasks are typically tedious and lengthy, and users often forgo them despite their importance. For instance, a typical color calibration task would display multiple colored circles on the display and require the user to respond “same” or “different” to whether or not the colors are uniform. By carefully considering various types of calibration tasks and matching them with specific game elements, the researchers created gamified calibration tasks. The gamified color calibration task became similar to the game Space Invaders, where users controlled a spaceship and shot lasers at any “different” colored circles that were displayed. A story/theme gaming element was also utilized, in which users were instructed to defend against attacking, evil invaders. The researchers were interested in both the level of enjoyment by the users as well as the quality of the calibration when compared to the non-gamified version of the tasks. The results revealed that participants found the gamified calibration activity significantly more enjoyable than the non-gamified version, as well as no difference in the quality of the calibration. Flatla et al. (2011) provide two important pieces of information. First, gamification
can improve enjoyment during a dull and tedious task. While this is not surprising, it is important empirical evidence nonetheless. Second, they also demonstrated the importance of careful and deliberate consideration when designing tasks using gamification. By analyzing what gaming elements would best compliment the various types of HCI calibration tasks, they maximized the benefits of their intervention.

The results from the Flatla et al. (2011) study demonstrate that gamification can improve task engagement in a tedious color calibration task. Calibration quality, however, did not improve under such conditions. This is a recurring theme in gamification. Creating a more engaging and interesting activity by applying gaming elements can undoubtedly enhance the overall task experience, but it does not guarantee that performance will improve. This should not undermine the benefits of gamification, but instead encourage task designers to be cautious when implementing gaming elements in non-gaming contexts.

Emphasizing the importance of careful gamification consideration, Kapp (2012) argues that gamification is not just applying points and rewards to an activity. What makes true games engaging and interesting goes beyond basic game mechanics. Core gaming elements such as storytelling, skill-based challenges, and achievement recognition promote engagement. When an activity is designed with these elements in mind, they can enhance the experience.

**Introduction Summary**

The current state of vigilance and low-prevalence search reveals a need to re-evaluate task design (Hancock, 2013; Wolfe et al., 2007). Human factors researchers are especially capable of providing such improvement recommendations. The current project explores how
gamification may be a viable and potentially effective intervention to improve performance and user experience.

The next chapter will discuss the details of the simulated inspection task used to test sustained attention during low-prevalence visual search. A simulated inspection task was created as a way to study the overlapping attentional mechanisms of vigilance and visual search for rare targets.
CHAPTER 2
SIMULATED INSPECTION TASK DEVELOPMENT

Background

The following chapter presents details on the simulated inspection task that was designed for the current study. A special task was created for this study so performance in vigilance and low-prevalence visual search could both be observed. Additionally, it was important to create a task that resembled a realistic low-prevalence vigil while using equipment within the confines of academic research settings.

Criticisms of vigilance research in academic laboratories date back to at least Smith and Lucaccini (1969). Detractors of laboratory vigilance research contend that the tasks and stimuli used in such studies are not representative of real-world vigilance activities (e.g. Upadhyay & Singh, 2013) and that efforts must be made to bridge the gap between basic and applied vigilance research (Adams, 1987). Researchers have addressed this concern in recent years. For instance, Szalma, Schmidt, Teo, and Hancock (2014) created a semi-realistic monitoring activity by applying vigilance elements to a first-person perspective video game task. The current project attempted to contribute to the semi-realistic laboratory vigilance research effort by creating a simulated inspection task.

Industrial Inspection

Industrial inspection has long been identified as an activity requiring vigilance for detecting low probability targets (Parasuraman et al., 1987). For example, Fox and Haslegrave (1969) were some of the first vigilance researchers to document vigilance performance and LP
effect outcomes among inspectors examining defects on small metal screws. Vigilance and the LP effect still play a role in industrial inspections today.

According to the Bureau of Labor Statistics (2015), just under 500,000 workers in the United States “inspect, test, sort, sample, and weigh... raw, processed, machined, fabricated, or assembled parts or materials for defects and deviations from specifications.” Their employment results in over $17 billion of estimated annual expense for the manufacturing industry. Clearly industrial inspection is a relevant area to consider when developing a simulated vigilance and low target prevalence task.

Creating a simulated industrial inspection task is considered a valid effort for a semi-realistic vigilance task and low prevalence search for a variety of reasons. First, previous research has documented that prevalence rates for defective items in industrial inspection settings range from as high as 40% to as low as less than 1% (Parasuraman et al., 1987). This range allows for the target prevalence levels to potentially elicit a LP effect. Second, industrial inspection settings have previously been characterized as vigilance tasks, featuring elements like high event rates and rare critical signals (Craig, 1984; Pesante-Santana & Woldstad, 2001; Swets, 1977). Lastly, creating an industrial inspection simulation allows for the experiments to methodically manipulate task parameters like event rate and the design of visual defects. This final point is critical for assessing reliable vigilance and SDT metrics (Pesante-Santana & Woldstad, 2001). We decided to create a simulated inspection task for small, round, metal washers.
Washer Defects Pilot

Our motivations for the washer defects created for the study were for them to be (1) realistic defects that could occur during washer manufacturing and (2) clear to an observer that a defect is present. A pilot study was conducted to confirm successful detection of washer defects. The operational definitions of the defects and d' scores for each defect are provided in Table 2.1. Five volunteers (3 female) participated in a 2-interval forced choice (2-IFC) task in which they decided which of two washers displayed in succession contained a defect. Volunteers were informed they would be comparing pictures of small metal washers and provided a brief verbal description of the defects, but did not view any sample images of the defects before the experiment began. The washers were displayed, one at a time, for 300 milliseconds (ms) with a 250 ms inter-stimulus interval. Each of the four defect types was presented 27 times, providing 108 total trials. Presentation order of defective vs. acceptable washer presentation was randomized. Sensitivity (d') was calculated using percentages of hits and false alarms (any incorrect trial) for each volunteer (Macmillan & Creelman, 2005). All four defects resulted in d' scores of at least 1.72, indicating that each defect was reliably detectable.
Figure 2.1. Images of acceptable and defective washers used in the experiment.
Table 2.1. List of the operational definitions for each of the four types of washer defects. All four defect types had average sensitivity (d') scores of above one, indicating they were detectable (Macmillan & Creelman, 2005).

<table>
<thead>
<tr>
<th>Defect Types</th>
<th>Operational Definitions</th>
<th>Sample Acceptable Washers</th>
<th>Same Defective Washers</th>
<th>2-IFC d' (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inner Indentation</td>
<td>1.5% pixels removed from inner ring</td>
<td></td>
<td></td>
<td>1.93 (1.05)</td>
</tr>
<tr>
<td>Outer Indentation</td>
<td>1.5% pixels removed from outer ring</td>
<td></td>
<td></td>
<td>1.72 (0.50)</td>
</tr>
<tr>
<td>Irregular Inner Oval</td>
<td>85% reduced width or height</td>
<td></td>
<td></td>
<td>1.98 (0.84)</td>
</tr>
<tr>
<td>Irregular Outer Oval</td>
<td>90% reduced width or height</td>
<td></td>
<td></td>
<td>2.98 (0.57)</td>
</tr>
</tbody>
</table>

**Experiment Zero: Simulated Inspection Pilot**

The primary objective of pilot experiment zero was to obtain baseline vigilance and low prevalence data in a simulated inspection activity. The results obtained from experiment zero were also useful for designing the various gaming elements to be featured in this specific task.

**Participants.** A group of 15 undergraduates (eight female) with an average age of 21.33 years (SD = 1.19) from Wichita State University were recruited using the SONA system and earned course credit for participation. All participants provided informed consent and were debriefed upon completion.

**Materials.** Five identical 2.4-GHz Mac Mini computers running MATLAB 2010A with the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) were used for the experiment. Participants sat approximately 45 cm from a 17-in. diagonal Dell 1704FPV screen at 1,280 X 1,024 pixel resolution and 60-Hz screen refresh rate. Images of nine different round metal washers were
used as the visual search stimuli. Individual washer stimuli measured at 3.44° visual angle diameter and were each displayed within square regions measuring 4.29° x 4.29° visual angle. Including a black border between the square regions, the entire 3x3 search array of washers was 14.53° visual angle. See Figure 2.2 for reference.

![Figure 2.2](image.png)

Figure 2.2. Screenshot of search stimuli. Feedback was provided during the practice block but not during the experimental session.

**Procedure.** Before the experimental task began, participants were informed that they would be inspecting images of nine small, round, metal washers in a 3x3 grid arrangement. They were instructed to search for a single defective washer that would only occasionally appear in the set. They viewed sample images of all four types of defective washers alongside images of acceptable washers (Figure 2.1).

On each trial, a washer search array would translate across the display from the right to the center. A black rectangle stretching across the horizontal extent of the screen was displayed as a background to give the idea of a conveyor line moving the items from right to left (as in Figure 2). After arriving in the center of the screen, the search array was stationary for 1.5
seconds. If during that time, the participant detected a defective washer, they pressed the SPACEBAR key to send the array up, off the conveyor line and out of view. Otherwise the search array transitioned off screen to the left as the next array arrived. Transition time of arrays moving between search trials was 0.25 seconds. With a 1.5 second duration for each trial, this created an event rate of 34 events/min. Participants completed a three minute practice block with immediate trial feedback for correct target hits, misses, and false alarms via a message saying “HIT,” “MISS,” or “FALSE ALARM.” The defective washer was highlighted in green during hit feedback and in red during miss feedback. Target prevalence during the practice block was 25%.

Following the practice block, participants entered into the experimental session. Each experimental session was divided into seven continuous nine-minute epochs and lasted approximately 60 minutes total. Note that participants did not receive a break during the experiment and were unaware of the grouping of trials into epochs. Targets appeared nine times every epoch, creating a 2.9% target prevalence rate with 63 total target-present trials across all 2,142 trials. Target location in the search array and defect type all appeared with equal likelihood.

**Results.** Three primary dependent measures were analyzed: hit rates across epochs (see Figure 2.3), overall decrement coefficient, and overall miss rate. Average hit rates across the 63-minute session ranged from as high as 46.67% ($SD = 17.92\%$) during the first epoch to as low as 34.07% (18.05%) during the third epoch. A one-way Analysis of Variance (ANOVA) revealed no significant change in hit rate across epochs ($p > .10$).
A decrement coefficient for hit rate was calculated for each participant and represents the average percentage hit rate change per epoch. The average decrement coefficient across participants was -0.50\% (3.83\%). The minimum coefficient observed was -5.56\% and the max was 4.76\%.

The hit rate per epoch results and decrement coefficient values both indicate performance did not decrease over time. Most participants remained somewhat stable in their performance across the session, revealing the absence of a vigilance decrement. Eight of the 15 participants did demonstrate a hit rate decrement with negative decrement coefficients, but overall this task did not lead to declining performance over time. We will return to this point later.

The overall miss rate across participants was 57.85\% (12.32\%). Additionally, criterion (c) values were obtained for each epoch for each participant and were calculated as $-0.5 \times [ Z(\text{hit}) +$
Z(false alarm) ] (Macmillan & Creelman, 2005). A one-way ANOVA revealed that average c significantly increased during the first three epochs, \( F(2,28) = 5.08, p < .05, \eta^2_p = .27 \), starting at around 0.73 (0.61) and increasing to as great as 1.13 (0.37) during the third epoch (see Figure 2.4). The upward criterion shift and substantial miss rates align with the previously discussed findings of a LP effect.

Figure 2.4. Experiment Zero criterion values across epochs. Error bars are +/- one standard error of the mean.

Simulated Inspection Task Discussion

Despite high miss rates in the pilot study, we did not observe a vigilance decrement since participant performance did not significantly deteriorate over time. There are a couple of explanations for this finding. The first explanation comes from critics of laboratory vigilance research who contend the vigilance decrement is a somewhat artificial finding in basic science studies. Previously mentioned was Adams’ (1987) contention that too much vigilance research focuses on the basic perceptual boundaries of sustained attention and not enough on
applications to realistic contexts. A major component of this argument is that a vigilance decrement is rarely observed in real world situations, including in industrial inspection settings (see Mackie, 1984 for a review). This presents the notion that this task did not produce a decrement perhaps because it did not represent the stimuli typically used in laboratory settings. Related to this point is research also using realistic scenarios in academic vigilance studies.

Recent studies have found a similar non-decrement outcome when attempting to create a dynamic and realistic vigilance task. Szalma et al. (2014) created a dynamic vigilance task that integrated first-person video-game characteristics to simulate walking through a village and searching for rare explosive devices. Like our simulated inspection task, the pace of simulation was controlled by the program and intended to be monotonous and tedious. Participants searched the streets and sidewalks of the village for any potential explosive devices. Szalma et al. (2014) failed to observe a performance decrement in the task and hypothesized that their activity did not represent the “perceptually impoverished context[s]” that are typically used to observe vigilance decrements (p. 1327). The results and rationale from Szalma et al. (2014), along with the relatively animated nature of the current simulated inspection activity, may help explain the lack of a vigilance decrement in the current task.

Furthermore, multiple iterations of the simulated inspection task were tested in an attempt to observe a decrement. The initial washer inspection task was 27 minutes long with a one-minute practice block. A performance decrement was not observed and most participants’ detection performance improved over time. This was an indication participants were learning how to detect the defects. To avoid a learning effect, the next iteration featured larger and more obvious defects, as well as a three-minute practice block. Performance across participants
in this version no longer improved over time and was stable across the duration, but there was still no decrement. The 63-minute version previously discussed was the next iteration tested.

Lastly, an additional version of the simulated inspection task was tested that featured images of various travel objects (toothbrush, magazine, phone charger, etc.) instead of washers. The targets appeared as occasional prohibited travel items (e.g., gun, knife, or bomb). This version was designed to be more demanding for the observer and perhaps more likely to produce a decrement since target items were clearly perceptible yet each item would need to be inspected one-by-one to find the target. The results mirrored the washer inspection experiments, in which no decrement was observed throughout various iterations. Additionally the overall miss rate for the travel version of the task was lower than the washer (27% compared to 58%, respectively). Therefore the washer version was considered superior for examining the LP effect.

Though no decrement was observed, this outcome was used as a logistical and statistical advantage for the primary experiments. Since this task did not produce a decrement even after an hour yet still produced a significant upward criterion shift, the task time was shortened to three epochs during a 27-minute vigil. This duration is supported by real-world inspection tasks, in which it is common for industrial inspectors to take a break after 20-30 minutes of inspection (Parasuraman et al., 1987). Individual vigils lasting less than 30 minutes each allowed for two vigils to be conducted in one experimental session. Now the design of the dissertation experiments could potentially remain within-subjects, which would strengthen their statistical power.
In summary, the results of the simulated inspection pilot experiment provided valuable information towards studying gamification of low-prevalence vigils. First, substantial miss rates of approximately 55% and significant upward criterion shift over time are indications of the LP effect. This provides an effective methodology towards studying any effects of gamification on the LP effect. Second, the multiple pilot sessions provided valuable information towards working out any unforeseen issues that may have appeared during the primary experiments and ultimately improved the final design of the inspection task. The final version used in the primary experiments feature well-defined task instructions, clearly detectable washer defects as target stimuli, a three-minute practice block, a successive 27-minute inspection session with 2.9% target prevalence. The final version is a rigorously tested method of observing the potential effects of gamification on performance and user experience in a low-prevalence vigil. The next chapter will discuss the motivations, hypotheses, methods, and results of gamification applied to the simulated inspection task.
CHAPTER 3

METHODS AND RESULTS

The following chapter presents the background for applying gamification to a low-prevalence vigil, followed by the hypotheses, methods, and results for the dissertation experiments. The study consisted of five total experiments: four initial experiments compared single gaming elements with existing task manipulations (Experiments 1a-1d) and a training experiment evaluated the effectiveness of gamified, traditional, and control training sessions (Experiment 2). Experiment 1a assessed badges as a gamified form of performance feedback. Experiment 1b assessed a points-based challenge as a burst of high target prevalence trials. Experiment 1c assessed storytelling as a gamified way to frame the purpose of the task. Experiment 1d assessed points-based feedback as a gamified form of performance feedback. The results of the initial series of experiments were used to inform which of the most effective individual gaming elements would be implemented in Experiment 2. Experiment 2 took part in two phases: a training phase and a test phase. The training phase consisted of a gamified training, traditional training, and a control (no feedback) training. The test phase was a control task with no feedback.

Careful consideration must be made when applying gamification. Understanding the demand characteristics of a task is imperative for designing game elements within said task. As presented in the previous chapter, the simulated inspection task created for the current project results in the low-prevalence (LP) effect: substantial miss rates for rare targets. According to Expectancy Theory, the source of this outcome is from an upward criterion shift \((c)\). Additionally, empirical evidence in vigilance research and anecdotal evidence in LP effect research has
identified increased mental workload and frustration during such activities. This information is key towards designing game elements for improving the signal detection accuracy and subjective assessments of task experience.

**Experiments 1a-1d: Assessing Individual Gaming Elements**

The first four experiments (Experiments 1a-1d) each assess a specific gaming element intervention in preparation for the Experiment 2 training task. The general methods for the first four experiments will be discussed, followed by the specific methods, results, and brief discussion for each experiment. The overall method for Experiments 1a, 1b, 1c, and 1d, including participant recruitment, materials, and planned analyses, remained constant.

**General Methods**

**Experimental Design and Power Analysis.** Experiments 1a, 1b, 1c, and 1d were independent two-way repeated measures designs, with the independent variables being epoch (three levels) and gaming intervention (two levels). The dependent variables collected at each of these levels were detection performance measures, including hit/miss rate and SDT metrics (criterion and sensitivity).

G*Power 3.1 software was used to conduct an a priori power analysis (Faul, Erdfelder, Lang, & Buchner, 2007). Using the ANOVA: Repeated measures, within factors formula with six measurements (three epochs X two gaming conditions), a moderate partial eta squared ($\eta^2_p$) effect size value (0.2), alpha level of .05, power level of 0.8, and conservative correlations of zero between measures, the calculated total sample size was $N = 10$. A larger sample size of $N = 20$ was collected for each experiment to allow for potential post-hoc analyses. The gaming intervention was intended to be a within-groups factor, but two of the four initial series of
experiments revealed potential carryover effects between conditions after the first 10 participants completed the activity. Therefore these two experiments (1c and 1d) were redesigned with gaming intervention as a between-groups factor and data was collected from a minimum of 40 new participants (20 participants in each condition).

**Participants.** All participants were recruited from Wichita State University using the SONA system. Each participant provided informed consent and earned course credit for their participation in accord with the Institutional Review Board guidelines.

**Materials.** The materials and apparatus for all experiments were the same as the equipment used in the simulated inspection pilot study. Subjective task experience measures were obtained for each the gaming intervention conditions. They included the NASA-TLX (Appendix A) as a measure of subjective mental workload (Hart & Staveland, 1988) and the Flow Short-Scale (Appendix B) as a measure of task engagement (Jackson, Martin, & Eklund, 2008).

**Procedure.** The overall task procedure was the same as described for the simulated inspection pilot study. All four experiments consisted of two versions of the inspection task: a traditional manipulation version and a gamified version. The specific task instructions were the same for both versions (“Press the SPACEBAR key if you detect a defective washer in the set”). Each version consisted of three continuous, nine-minute epochs in which targets appeared once in all nine locations per epoch (2.9% prevalence). All four defect types were equally likely to appear across the session. For experiments 1a and 1b, the order of gamified version or traditional version was counterbalanced across participants. Gaming condition was a between-groups factor for experiments 1c and 1d. At the end of each inspection task, participants
completed the NASA-TLX and Flow Short-Scale questionnaires. All questionnaires were presented on the computer and each rating was entered via a mouse click.

Statistical Analyses. The primary statistical test used for all detection performance data was a 2x3 repeated measures Analysis of Variance (ANOVA), with two levels of gaming intervention (present or absent) and three levels of epochs. The Bonferroni approach was used to control family-wise error rate across all pairwise comparisons. In any case of Mauchly’s sphericity test showing that the sphericity assumption was not met, \( p < .05 \), the Greenhouse-Geisser correction was used. Additionally, participants who had an overall sensitivity (d’) of less than one were excluded from the analyses.

The nonparametric Wilcoxon Signed-Ranks Test was used to evaluate differences in subjective assessments of mental workload and task engagement. This allowed for comparisons of the six specific NASA-TLX subscale ratings (ranging from 1 to 21) and nine subscale levels of flow (ranging from 1 to 6) between gaming conditions. The cutoff for significance was reduced to \( p = .008 \) for the 6 NASA-TLX subscale comparisons and \( p = .006 \) for the nine flow short scale subscale comparisons to control for Type I error.

Experiment 1a: Badges and Traditional Feedback

As established, knowledge-of-results and immediate performance feedback are effective interventions towards improving overall target detection performance (Szalma et al., 2006). With this in mind, badges as a type of gaming element could be a modified version of performance feedback that promote intrinsic motivation and enhance overall task experience.

As a gaming element, badges are visual representations of a rewards for task completion or reaching certain levels of achievement (Antin & Churchill, 2011). They are
commonly used on social media sites like Wikipedia or StackOverflow to encourage and reward valuable contributions to the site content (Antin & Churchill, 2011). When designed effectively, gamified rewards like badges represent task success and achievement, which can improve learning, performance efficiency, and self-efficacy within the task (Kapp, 2012; Richter et al., 2015).

In a low-prevalence vigilance task, badges might be able to improve overall target detection performance. Since the tendency in low prevalence detection tasks is for hit rates to be substantially low, badges awarded for target hits and designed to encourage a liberal response criterion might serve as a more engaging version of feedback. The purpose of Experiment 1a was to observe the potential benefits of badges as a form of gamified traditional feedback (see Figure 3.1).

**Hypothesis 1:** Overall miss rate for the badges condition will be significantly lower than the overall miss rate for the traditional feedback condition.

**Hypothesis 2:** Task engagement will be rated significantly higher for the badges condition compared to the traditional feedback condition.
Participants. Twenty participants (10 female, $M = 23.80$ years old, $SD = 8.68$) performed both versions of the simulated inspection task. Half of the participants performed the gamified version first and the other half performed the traditional version first.

Method. The badge condition featured additional visible representations of task performance (see Table 3.1). Badges were designed to encourage a liberal criterion bias so that participants would be more inclined to commit false alarm errors than miss errors. After the practice block and before the experimental session, all three levels of all three badges were described and displayed on the screen. Participants were informed that they could earn the badges based on their performance. At the start of the experimental task, the names of all three badges were displayed at the bottom of the screen, but the badge graphic itself was locked. Participants earned each level of badge in order (i.e., bronze, then silver, then gold). Precision and bullseye
Badges were not awarded until three targets have been presented so that they were not administered after the first trial or if the first target is detected. Because these two badges were earned based on overall performance, they were awarded near the completion of each epoch. For instance, if a participant had a false-alarm rate of less than 2% throughout the first epoch, the bronze-level precision badge was awarded (approximately eight minutes into the session). If he or she maintained an overall false-alarm rate of less than 2% throughout the first and second epochs, the silver level precision badge was awarded (approximately 16 minutes into the session). If he or she maintained an overall false-alarm rate of less than 2% throughout the majority of the session, then the gold level precision badge was awarded near the end of the session. The Precision and Bullseye badges featured progress bars, indicating when the next badge would be awarded. The hit streak badge featured a visible counter so that the participant was aware of how many targets in a row they had detected and how many were left until the next badge level would be earned. At the completion of the task, participants were shown a graphic informing them of all three badges that they earned.

Table 3.1. List of badge descriptions and images. Once a participant earns a badge, they will retain it for the duration of the session.

<table>
<thead>
<tr>
<th>Badge Type</th>
<th>Performance Measure</th>
<th>Design Purpose</th>
<th>Badge Level and Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bullseye</td>
<td>Overall hit rate &gt;55%</td>
<td>Encourage “target present” response</td>
<td><img src="image1.jpg" alt="Locked" /> <img src="image2.jpg" alt="Bronze" /> <img src="image3.jpg" alt="Silver" /> <img src="image4.jpg" alt="Gold" /></td>
</tr>
<tr>
<td>Hit Streak</td>
<td>4 consecutive defects detected</td>
<td>Encourage consistency and promote a skill-based challenge</td>
<td><img src="image5.jpg" alt="Locked" /> <img src="image6.jpg" alt="Bronze" /> <img src="image7.jpg" alt="Silver" /> <img src="image8.jpg" alt="Gold" /></td>
</tr>
<tr>
<td>Precision</td>
<td>Overall false alarm rate &lt;2%</td>
<td>Discourage abundance of “target present” responses</td>
<td><img src="image9.jpg" alt="Locked" /> <img src="image10.jpg" alt="Bronze" /> <img src="image11.jpg" alt="Silver" /> <img src="image12.jpg" alt="Gold" /></td>
</tr>
</tbody>
</table>
Results

Detection Performance. A series of 2x3 (Gaming Condition x Epoch) repeated measures ANOVAs were conducted for the three primary detection performance measures of sensitivity, criterion, and hit rate. There was no significant main effect nor interacting effects of task order as a between-subjects covariate. The analyses detected no significant main effect of either gaming condition or epoch for any of the three detection performance measures, as well as no significant interactions for any of the detection performance measures (all ps > .10, see Figure 3.2). Additionally, a paired samples t-test was conducted to determine the outcome of the hypothesis that the badges condition would result in a significantly lower overall miss rate compared to the traditional feedback condition (H1). The results do not support H1 due to no significant difference in overall miss rate between the two conditions (p > .05).

Figure 3.2. Experiment 1a comparing hit rate between Badges and Traditional Feedback conditions. Error bars are +/- one standard error of the mean.
**Subjective Experience.** A series of Wilcoxon Signed-Rank tests were conducted for participant ratings on the NASA-TLX and the Flow Short Scale to assess the hypothesis of increased task engagement rates in the badges condition (H2). The analyses detected no significant differences on any NASA-TLX subscales (all ps > .10). Two Flow Short-Scale subscale comparisons returned marginal significant differences: concentration, $z = -2.23, p < .03$ and autotelic (i.e. rewarding) experience, $z = -2.41, p < .01$ (see Figure 3.3), indicating support for H2. There were no significant differences on the remaining flow state dimensions (all ps > .05).

![Flow State](image)

**Figure 3.3.** Experiment 1a mean flow dimension ratings. Error bars are +/- one standard error of the mean.

**Discussion**

The purpose of Experiment 1a was to observe the potential benefits of implementing badges as a gamified form of feedback. The badges were hypothesized to (H1) alleviate the
substantial miss rates for rare targets by encouraging liberal response criterion and (H2) promote positive task engagement. The results of Experiment 1a do not support H1 due to no difference in detection performance between badges and traditional feedback. The results do, however, support H2 due to participants rating two flow dimensions more favorable in the badge version of the inspection task than traditional feedback.

When badges were made available, participants reported (1) feeling more rewarded during the simulated inspection task and (2) having greater concentration during the task. It is no surprise badges yielded a more rewarding experience given they are a gaming element known to provide a visual representation of task achievement, which can improve self-efficacy (Kapp, 2012). The difference in concentration ratings, however, are a little more interesting.

One explanation for why badges made participants feel they concentrated more is the additional information they provided may have allowed them to focus more on the intended task. Previous research revealed individuals may experience task-unrelated thoughts during a sustained attention task, where attention is directed internally and away from the primary task at hand (e.g. Scerbo, 1998). When badges were presented on the display, participants may have been more focused on information they provided regarding the current task than on task-unrelated thoughts.

Despite the positive task engagement provided by the badges, the demands of the low-prevalence inspection task may have been too great for them to encourage improved detection performance. A search task featuring rare targets provides rare opportunities for task success. It therefore became discouraging to adopt the liberal response criterion even though the badges were designed specifically to encourage it. Furthermore, badges were a tedious gaming
element to design, development, and implement, requiring several pilot phases to finalize. Consequently, it appears badges that as a gamified form of feedback during a simulated inspection task are not a particularly effective intervention and do not encourage improved performance for rare targets.

**Experiment 1b: Points-Based Challenge and Burst of Higher Prevalence**

Low signal probability environments typically produce a conservative criterion, in which the operator is more likely to respond “target absent.” As established, criterion can be recalibrated towards a more liberal approach when a burst of high target prevalence trials are interspersed during the activity (Wolfe et al., 2007). With this intervention in mind, a burst of high target prevalence trials framed as a points-based challenge or “lightning round” could be a modified version of a burst that could also provide the benefit of recalibrating criterion while enhancing overall task experience.

Points are among the most widely-used gaming elements in gamification and often considered an essential part of gamified systems (Hamari et al., 2014; Zichermann & Cunningham, 2011). Point systems can be designed in a variety of ways (e.g., users can accumulate points so they can spend them on rewards) and are intended to elicit a specific behavioral outcome (Kapp, 2012). Miranda and Palmer (2014), for instance, observed increased intrinsic motivation when applying a point system with a high score and bonus points to encourage fast and accurate performance in a visual search task.

In game design, challenges are additional tasks encountered at various points of the activity intended to reinforce task confidence (Duggan & Shoup, 2013; Zichermann & Cunningham, 2011). Like badges, completing challenges represents task success and proficiency.
This sense of accomplishment might enhance task competence and intrinsic motivation, in which task confidence is reinforced (Kapp, 2012; Richter et al., 2015).

A burst of high target prevalence trials presented as a points-based challenge could create a more engaging experience compared to a non-gamified burst of trials. A point system that encourages correct target detections could compel users to respond “target present,” possibly shifting criterion towards a more liberal approach. The presence of high target prevalence burst of trials is already known to reduce miss rates via a criterion shift, so a points-based challenge that promotes skill improvement and task competence may further enhance the experience. The purpose of Experiment 1b was to examine the potential benefits of a points-based challenge or “lightning round” as a modified version of a burst of high target prevalence trials.

**Hypothesis 3:** Criterion will be significantly lower (more liberal) in the points-based challenge condition compared to the high prevalence burst condition.

**Hypothesis 4:** Task engagement will be rated significantly higher for the points-based challenge condition compared to the high prevalence burst condition.
Participants. Twenty participants (eight female, $M = 21.95$ years old, $SD = 5.81$) performed both versions of the simulated inspection task. Half of the participants performed the gamified version first and the other half performed the traditional version first.

Method. The burst of high prevalence trials was introduced after the second epoch in both conditions. This allowed for detection performance to be observed before and after the burst was injected. The burst consisted of a two-minute session (68 trials) with 50% target prevalence (34 targets). Before the burst started, participants were informed via a message on the screen that the next set of trials was different and featured performance feedback. For the gamified condition, the participants were informed that the upcoming set of trials was a “lightning round” in which they had the opportunity to accrue points based on their
performance (see Table 3.2 for point system details). Both the points earned or lost after each trial and total points accumulated were visible on the screen during the burst. Once the two-minute burst ended, a message appeared informing the participant that the round ended and congratulated them on earning however many points they earned. For the traditional burst condition, the participants were informed that the upcoming set of trials was a “practice round” that featured immediate performance feedback. Once completed, a message informed them the round has ended and they were instructed to continue with regular inspections.

Table 3.2. Points system used during the Points-Based Challenge version of the high prevalence burst. Overall point system was designed to encourage a liberal criterion and subsequently reduce target miss rate. Points were administered on each trial during the “Lightening Round” but not during the other parts of the task.

<table>
<thead>
<tr>
<th>Trial Outcome</th>
<th>Number of points</th>
<th>Design Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit</td>
<td>+10</td>
<td>Encourage “target present” response</td>
</tr>
<tr>
<td>False Alarm</td>
<td>-1</td>
<td>Discourage abundance of “target present” responses</td>
</tr>
<tr>
<td>Miss</td>
<td>-25</td>
<td>Substantial punishment for misses to discourage bias towards “target absent” response</td>
</tr>
<tr>
<td>Correct Rejection</td>
<td>+1</td>
<td>Small point value to discourage bias towards “target absent” response</td>
</tr>
</tbody>
</table>

Results

Detection Performance. A series of 2x3 (Gaming Condition x Epoch) repeated measures ANOVAs were conducted for the three primary detection performance measures of sensitivity,
criterion, and hit rate. There were no significant main effects or interaction for sensitivity (all ps > .05). For response criterion, the analyses detected a significant main effect of epoch, \(F(2,36) = 7.61, p < .01, \eta_p^2 = .30\), with follow-up pairwise comparisons indicating participants had a significantly lower response criterion (i.e. more liberal response bias) during the third epoch following the two-minute intervention (\(p < .01\), see Figure 3.5). There was no main effect of gaming condition nor significant interaction (both ps > .05), which does not support the hypothesis that the points-based challenge would have a significantly lower criterion (H3). A similar result pattern was observed for hit rate performance, with the analyses detecting a significant main effect of epoch, \(F(2,36) = 7.30, p < .01, \eta_p^2 = .29\), and follow-up pairwise comparisons indicating participants had a significantly greater hit rate during the third epoch following the two-minute intervention (\(p < .01\)). There was no main effect of gaming condition nor significant interaction (both ps > .05).

**Figure 3.5.** Experiment 1b comparing Points-Based Challenge and Traditional High-Prevalence Burst. Criterion values across epochs. ANOVA revealed significant main effect of epoch
indicating downward criterion shift following the two-minute high-prevalence burst after the second epoch regardless of gaming condition. Error bars are +/- one standard error of the mean.

**Detection Performance During Two-minute Burst.** Further analyses were conducted using the performance data observed during the two-minute burst. Two paired samples t-tests revealed participants had a higher target hit rate during the points-based challenge (\(M = 87.21\%, \ SD = 6.77\%\)) compared to the traditional two-minute burst (\(M = 82.35\%, \ SD = 9.20\%\)), \(t(19) = 2.41, p < .05, \eta_p^2 = .23\) (see Figure 3.6), as well as faster reaction time during the points-based challenge (\(M = 785.32\ ms, \ SD = 95.25\ ms\)) compared to the traditional two-minute burst (\(M = 838.78\ ms, \ SD = 71.36\ ms\)), \(t(19) = 2.36, p < .05, \eta_p^2 = .23\) (see Figure 3.7). These results suggest the points-based challenge encouraged overall more proficient inspection performance compared to traditional performance feedback during the two-minute burst of high prevalence.

![Figure 3.6](image.png)

**Figure 3.6.** Mean hit rate during two-minute burst between second and third epochs. Error bars are +/- one standard error of the mean.
Subjective Experience. A series of Wilcoxon Signed-Rank tests were conducted for participant ratings on the NASA-TLX and Flow Short-Scale to assess the hypothesis of increased task engagement rates in the points-based challenge condition (H4). The analyses detected no significant differences on any NASA-TLX or Flow Short-Scale subscales (all $p$s > .10).

Discussion

Experiment 1b replicated previous findings showing a burst of high-prevalence trials will shift criterion towards a more liberal response, encouraging an increased target detection rate during subsequent low-prevalence trials (Wolfe et al., 2007). This intervention was further supported with the large effect size that was observed. The point system used in the gamified burst was hypothesized to (H3) lower criterion more than the high-prevalence burst condition and (H4) promote positive task engagement. While the results of Experiment 1b do not support
either of the hypotheses, the point system was effective during the two-minute inspection activity with 50% target prevalence.

This success revealed the effectiveness of a point system as a gamification technique within a simulated inspection task, further substantiated but the large effect sizes observed for both the hit rate and RT results. It suggests that perhaps a point system applied during a prolonged search for rare targets may lead to improved performance compared to traditional feedback. Furthermore, if the effectiveness observed during the two-minute burst can be applied during 3% prevalence, then it becomes a viable gaming element to use during the training phase of Experiment 2. Experiment 1d was conducted to assess this possibility.

**Experiment 1c: Storytelling and Instructions**

Subjective assessments of mental workload following participation in a vigilance activity can be altered by task instructions. As discussed, vigilance activities framed as “relaxation exercises” were considered less mentally demanding that those framed as “challenging” (Sawin & Scerbo, 1995). Therefore task instructions play a key role in how a user will approach a vigilance task. In gamification, task instructions can take the form of the story or theme of the task.

Game designers Rollings and Adams (2003) claim that the narrative aspect of a game might be considered the most fundamental element in game design. The authors also emphasize that the story driving a game does not need to be intricate or elaborate, and that oftentimes a short, simple backstory will increase engagement. For instance, the short narrative provided in the Flatla et al. (2011) study on gamifying a color calibration task discussed in the first chapter contributed to the users’ increased enjoyment in the activity. A story for an activity
gives the user a sense of purpose and meaning for the experience, provides important context and direction, and may increase intrinsic motivation (Kapp, 2012; Richter et al., 2015).

The purpose of Experiment 1c was to observe the potential benefits of a story or theme as a gamified version of task instructions. Giving a user with a sense of purpose and meaning for their participation provides them with a sense of direction and context, which may improve performance and/or enhance task experience.

**Hypothesis 5:** Mental workload ratings will be significantly lower in the storytelling condition compared to the instructions condition.

**Hypothesis 6:** Task engagement will be rated significantly higher for the storytelling condition compared to the instructions condition.

![Figure 3.8. Storytelling used as a gamified version of instruction framing in Experiment 1c.](image)
Participants. This experiment became a between-subjects design after potential order effects were observed in the initial 10 participants. Data were collected from 41 new participants (30 female, $M = 20.59$ years old, $SD = 3.71$). Twenty one participants were assigned to the storytelling condition and 20 assigned to the traditional instructions condition.

Method. The practice and experimental sessions of the inspection task were similar to how they were conducted in the pilot study (i.e., immediate performance feedback after each trial during practice but not during the experimental session). Before practice and during the instructions, both groups were informed of the basic nature of the task. An additional set of instructions were provided in the storytelling condition. The additional script for the storytelling condition was as follows:

*For this task, imagine you work for the National Aeronautics and Space Administration (NASA). Specifically, you work in the aeronautical parts and manufacturing division. Your job consists of several different responsibilities, one of which includes inspecting manufactured parts for potential defects.*

*Recently your manager informed you that NASA administrators have specifically asked for you to inspect a certain set of parts intended to be used for NASA’s highly anticipated “Journey to Mars” mission. These parts are of utmost importance since the mission will endure a long duration and extensive use. A defective item could potentially lead to disastrous consequences for the mission and its crew. Successful item inspection, however, will likely lead to the first human mission to Mars. Good luck!*
No additional information was provided during the instructions condition. Participants assessed their subjective mental workload and flow state at the completion of each condition.

Results

Detection Performance. A series of 2x3 (Gaming Condition x Epoch) repeated measures ANOVAs were conducted for the three primary detection performance measures of sensitivity, criterion, and hit rate. For sensitivity, there were no significant main effects (all ps > .05) but a significant interaction between gaming condition and epoch, $F(1.72, 67.02) = 3.60, p < .05, \eta_p^2 = .08$ (see Figure 3.9). There were no main effects or interactions for either response criterion or hit rate (all ps > .05).

Additional analyses were conducted to better understand the significant interaction of gaming condition and epoch on sensitivity. To control for Type I error across main effects, alpha was reduced to .025. The first analysis examined differences in sensitivity between groups at each epoch and detected no significant differences between groups at any epoch, ($p > .05$). The next analysis compared differences in sensitivity scores across epochs for story and instruction conditions separately.

There were no significant differences in sensitivity scores across epochs in the story condition, $F(2,40) = 0.03, p > .05, \eta_p^2 = .00$, but there were significant differences for the instruction condition, $F(1.54, 29.34) = 11.10, p < .01, \eta_p^2 = .37$. Follow-up pairwise comparisons were conducted to evaluate the differences among the mean sensitivity scores at each epoch in the instruction condition. Alpha was reduced to .008 to control for Type I error. The mean sensitivity score during the first epoch was significantly lower than the mean scores for both...
subsequent epochs. The results of the interaction and subsequent follow-up tests suggest participants in the storytelling condition demonstrated more stable target sensitivity throughout inspection session compared to the instructions condition.

**Figure 3.9.** Experiment 1c comparing Storytelling and Traditional Instructions. Mean sensitivity (d') across epochs for the storytelling and traditional instructions conditions. ANOVA revealed a significant interaction between gaming condition and epoch, indicating differences in sensitivity over time between the two conditions. Error bars are +/- one standard error of the mean.

*Subjective Experience.* A series of Mann-Whitney *U* tests were conducted for participant ratings on the NASA-TLX to assess the hypothesis that the storytelling condition would provide lower workload ratings (H5), as well as the Flow Short-Scale to assess the hypothesis of increased task engagement rates in the storytelling condition (H6). For the NASA-TLX ratings, the analyses detected a significant difference on the mental workload dimension, \( z = -2.68, p < .008 \), and a moderate difference on the temporal workload dimension, \( z = -2.42, p < .02 \). Participants in the storytelling condition provided significantly higher ratings on both.
dimensions compared to the traditional instructions condition (see Figure 3.10). There were no other significant differences on the remaining NASA-TLX dimensions or Flow Short-Scale dimensions (all $p$s > .05).

![Figure 3.10. Experiment 1c comparing Storytelling and Traditional Instructions. Mean NASA-TLX Mental Workload Assessment subscale ratings. Error bars are +/- one standard error of the mean.](image)

**Discussion**

Experiment 1c investigated how using storytelling to add meaning and purpose to the simulated inspection task compares to traditional task instructions. The NASA narrative was hypothesized to (H5) reduce subjective mental workload ratings and (H6) promote positive task engagement. While the results do not support either of the hypotheses, the story did produce two significant effects. First, the storytelling manipulation produced more stable target sensitivity performance over time compared to traditional instructions. Second, the story
produced higher levels of subjective mental workload compared to traditional instructions. Previous studies examining outcomes of different instructions for a sustained attention task provide helpful insight for interpreting these results.

Previous vigilance research has found detection performance is superior when an inspection task is framed as challenging as opposed to monotonous (Lucaccini, Freedy, & Lyman, 1968). Additionally, more stable detection performance is observed when participants feel their performance is being explicitly evaluated or tested in contrast to performing the task without such mindset (Huntermark & Witte, 1978). These factors potentially played a role in the storytelling condition. Stable target sensitivity over time could be in part due to participants approaching the task as being more of a challenge with the feasibility of meaningful consequences. The moderate effect size observed for the interaction further supports this interpretation.

Embracing the narrative of an industrial inspector employee for NASA produced an unanticipated outcome of increased subjective mental workload. There are couple ways to interpret this outcome. First, previous research examining serious games have found that higher workload ratings are more reflective of how simulated tasks would be conducted in the real-world (Iuppa & Borst, 2014). Conveniently, the simulated inspection task used for the study provides a realistic attention task. Add this realism to a plausible storyline, and now participant mindset further resembles a real-world inspection task. Second, aspects of the Prospect Theory of decision making provide further information towards meaningful interpretation.

Kahneman and Tversky (1979) suggest that, when presented with a decision, human beings have a tendency to classify outcomes as either losses or gains, with losses being typically
more impactful than gains. When applied to the current study, participants in the storytelling condition may have emphasized the loss outcome (disastrous consequences for the space mission crew) more than the gain outcome (successfully sending humans to Mars). This concern may have lead to the task being considered more demanding.

**Experiment 1d: Points-Based Feedback and Traditional Feedback**

The purpose of Experiment 1d was to investigate the potential benefits of using points-based feedback during a 27-minute low target-prevalence inspection task (see Figure 3.11). The benefits observed in Experiment 1b showed improved hit rate and reaction time during the two-minute burst, but it was unknown if this intervention would be effective during a longer inspection session. Additionally, one of the benefits of using a point system as a gamification technique is its flexibility. Points as a gaming element are known to be widely used for this reason, as well as the observed benefits to performance (Hamari et al., 2014; Zichermann & Cunningham, 2011). The results of this experiment would be used to inform how the points-based feedback would be implemented into Experiment 2. If they were as effective throughout the entire session as observed in Experiment 1b, then a thorough points-based feedback system would be implemented in the gamified training of Experiment 2. If not, then the points-based challenge during a high-prevalence burst would be implemented, just as conducted in Experiment 1b.
Figure 3.11. Points-based feedback was used as a gamified version of traditional performance feedback in Experiment 1d.

**Hypothesis 7:** Hit-rate performance will be significantly greater in the points-based feedback condition compared to the traditional feedback condition.

**Hypothesis 8:** Task engagement will be rated significantly higher for the points-based feedback condition compared to the traditional feedback condition.

**Participants.** This experiment became a between-subjects design after potential order effects were observed from the initial 10 participants. Data were collected from 40 new participants (32 female, $M = 19.98$ years old, $SD = 3.49$). Half of the participants were assigned to the points-based feedback condition and the other half assigned to the traditional feedback condition.
**Method.** Participants in the points-based feedback condition were provided with the exact point system used in Experiment 1b but for the entire duration of the session. Participants in the traditional feedback condition were provided with the immediate performance feedback previously described. There was no high-prevalence burst intervention in either condition. All participants completed the NASA-TLX and Flow Short-Scale following the inspection activity.

**Results**

*Detection Performance.* A series of 2x3 (Gaming Condition x Epoch) repeated measures ANOVAs were conducted for the three primary detection performance measures of sensitivity, criterion, and, to assess the hypothesis that performance would be greater in the points-based feedback condition (H7), hit rate. The analyses detected no significant main effect of either gaming condition or epoch for any of the three detection performance measures, as well as no significant interactions for any of the detection performance measures (all ps > .10), indicating H7 was not supported.

*Subjective Experience.* A series of Mann-Whitney U tests were conducted for participant ratings on the NASA-TLX and Flow Short-Scale to assess the hypothesis of increased task engagement rates in the badges condition (H8). The analyses detected no significant differences on any NASA-TLX or Flow Short-Scale subscales (all ps > .05), indicating H8 was not supported.

**Discussion**

Experiment 1d compared the effectiveness of points-based feedback with traditional feedback. The point system used throughout the task was hypothesized to (H7) improve hit-
rate performance compared to traditional feedback and (H8) promote positive task engagement. The results of Experiment 1d do not support either of the hypotheses. Point-based feedback was no more effective than traditional feedback during a low target prevalence inspection task. The point system was most effective during the two-minute inspection activity with 50% target prevalence as observed in Experiment 1b. Research has shown that a point system as a form of gamified feedback can enhance performance (Kapp, 2012; Richter et al., 2015). The question remains why it was effective during a 50% target prevalence search and not 3% prevalence. One interpretation comes from the broader research area on feedback. Feedback Intervention Theory emphasizes the relationship between task demands and feedback effectiveness (Kluger & DeNisi, 1996). Because a 3% prevalence inspection is more demanding than a 50% prevalence inspection, the efficacy of the point system was reduced.
Experiments 1a-1d: Results Summary

Table 3.3. Summary of significant findings across initial series of experiments. Black dash (—) indicates no significant difference between gamified intervention and traditional intervention, checkmark (✓) indicates significant difference.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Gaming Element</th>
<th>Traditional Element</th>
<th>Target Sensitivity ($d'$)</th>
<th>Response Criterion ($c$)</th>
<th>Target Hit Rate</th>
<th>Reaction Time</th>
<th>Mental Workload</th>
<th>Task Engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Badges</td>
<td>Feedback</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
</tr>
<tr>
<td>B</td>
<td>Points-Based Challenge</td>
<td>High-Prevalence Burst</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>✓</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>C</td>
<td>Storytelling</td>
<td>Instructions</td>
<td>✓</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>✓</td>
<td>—</td>
</tr>
<tr>
<td>D</td>
<td>Points-Based Feedback</td>
<td>Feedback</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
Experiment 2: Gamified Training

The purpose of the four initial series of experiments was to compare an individual gamified intervention with an equivalent traditional intervention. The results were also used to inform which combination of gaming elements would be applied to the training session of Experiment 2. The two gaming elements that produced performance benefits (Table 3.3) were the Points-Based Challenge from Experiment 1b and Storytelling from Experiment 1c. Therefore these two gaming elements were implemented in the gamified training condition for Experiment 2. The traditional training condition featured the traditional intervention counterparts. The control training condition had no interventions, gamified, traditional, or otherwise. The test phase resembled an industrial inspection scenario in the real world and therefore did not include any task interventions (see Figure 3.12). The purpose of Experiment 2 was to investigate the potential benefits gamified training may have on subsequent simulated inspection task performance.

Training in vigilance has long been shown to be an effective method of improving detection performance (Szalma, Miller, Hitchcock, Warm, & Dember, 1999; Teo, Schmidt, Szalma, Hancock, & Hancock, 2014; Warm et al., 2009; Wiener, 1963, 1968; Wiener & Attwood, 1968). Specifically, when operators train for a vigilance task with immediate performance feedback, their performance on a subsequent vigil without any feedback is superior compared to a training group without any feedback during training (Warm et al., 2009; Wiener, 1963).

In real-world low prevalence visual searches, however, training does not seem to be as effective. Wolfe et al. (2013) observed that newly trained airport security screeners were just as vulnerable to the LP effect as the general population. This indicates that professional training
for detecting threatening objects among x-ray baggage images does not protect against increased miss rates for rare targets. On the bright side, the security personnel in the Wolfe et al. (2013) study did demonstrate improved detection performance from a burst of high prevalence trials. While formal training does not seem to alleviate the LP effect, an industrial inspector may benefit from occasional gamified practice searches of high-prevalence trials.

These findings allow for the possibility of gamification techniques being integrated in an industrial inspection environment. If gaming elements are an effective training technique, companies may want to explore how to integrate them into a training or practice program so that operators and inspectors can participate in an enhanced training experience. The current projects looked to implement gamification techniques in a semi-realistic vigilance search task for low prevalence targets.

**Figure 3.12.** Experiment 2 design. The purpose of this experiment was to investigate potential benefits gamified training may have on subsequent performance. Gamified, traditional, and control training conditions were between-groups conditions.
Hypothesis 9a: In the training phase, the control training group will have the worst performance.

Hypothesis 9b: In the training phase, the control training group will have the highest mental workload ratings.

Hypothesis 9c: In the training phase, the control training group will have the lowest task engagement ratings.

Hypothesis 10: In the training phase, the gamified group will report higher levels of task engagement compared to both the traditional and control groups.

Hypothesis 11: Both gamified and traditional training groups will outperform the control training group in the test phase.

Hypothesis 12: The gamified training group will outperform the traditional training group in the test phase.

Experimental Design and Power Analysis

Experiment 2 was a three-factor mixed design, with the within-subjects variables being Training Session (two levels: training day and test day) and Epoch (six levels during each day); the between-subjects variable were Gaming Intervention (three levels: gamified training, traditional training, and control training). The dependent variables collected at each of these levels were the same as the previous experiments.

As before, G*Power 3.1 software was used to conduct an a priori power analysis. Using the ANOVA: Repeated measures, within-between factors formula with three groups (gaming intervention), 12 measurements (6 epochs X 2 days), a moderate $\eta_p^2$ (0.2), alpha level of .05,
power level of 0.8, and conservative correlations of 0 between measures, the calculated total sample size was \(N = 15\). A larger sample size of \(N = 46\) was used to allow for potential post-hoc analyses.

**Planned Analyses.** The primary statistical tests for both the detection performance data and subjective task experience measures was the same as the previous experiments. The same corrections and exclusions were also enforced.

**Participants.** Data were collected from 46 total participants (30 female, \(M = 20.72\) years old, \(SD = 5.83\)). Two participants, one in the gamified training condition and one in the traditional training condition, were removed from analyses for having training session hit rate and/or target sensitivity performance more than two standard deviations below the mean. Statistical tests were conducted using data collected from the remaining 44 participants (29 female, \(M = 20.30\) years old, \(SD = 4.97\)). Participants were randomly assigned to one of three training conditions: gamified (\(N = 14, M = 19.14\) years old, \(SD = 1.35\)), traditional (\(N = 16, M = 20.06, SD = 3.11\)), and control (\(N = 14, M = 21.71, SD = 8.07\)). Chi-square analyses and ANOVA revealed no differences in gender ratio or age between the three conditions (all \(p > .10\)).

**Method.** Participants were randomly assigned to one of three conditions. The gamified training condition featured the two effective gaming elements examined in the first set of experiments (points-based challenge and storytelling) during the training session. The traditional training condition featured the two traditional manipulations that aligned with the gamified manipulations (high prevalence burst and instructions) during the training session. The control condition did not feature any task manipulations during the training session.
During the training phase, participants completed the simulated inspection task for 54 consecutive minutes (six continuous nine-minute epochs). Event rate and target prevalence remained constant from the previous experiments. Overall each manipulation resembled how it was implemented in the previous experiments except for minor changes to the burst intervention. High prevalence bursts, both gamified and traditional, occurred twice: once after the second epoch and again after the fifth epoch.

Participants completed the mental workload and flow state assessments at the end of the session. A researcher coordinated with the participant on when to return at some point during the following six to eight days for the test phase.

The test phase of the experiment was the same for all conditions. It featured basic task instructions, a practice block with immediate performance feedback, and a six-epoch experimental session with no feedback. This was an attempt to resemble a realistic industrial inspection scenario in which neither feedback nor knowledge of results are provided to the inspector as they are performing the task. Participants completed the subjective workload and flow state assessments at the end of the session.

Results

*Training Session Detection Performance.* A series of 3x6 (Training Condition x Epoch) mixed ANOVAs were conducted for the three primary detection performance measures of sensitivity, criterion, and hit rate. Performance metrics for both the gamified and traditional conditions from the two high-prevalence bursts were excluded from analyses. For sensitivity, the analyses detected a significant main effect of epoch, $F(5,205) = 3.46, p < .01, \eta_p^2 = .08$, with follow-up pairwise comparisons indicating target sensitivity modestly improved during the sixth
and final epoch in comparison to the second epoch \((p = .09)\). The modest effect size also indicates that the epoch main effect is likely driven by the burst of high prevalence presented after the 2nd and 5th epochs in the gamified and traditional training conditions. There was no main effect of training condition nor significant interaction (both \(ps > .10\)), indicating no support for the hypothesis that the control training group would have the worst performance \((H9a)\). Nor were there any significant outcomes for criterion or hit rate (all \(ps > .05)\).

**Gamified vs. Traditional Detection Performance During Two-minute Bursts.** Further analyses were conducted using the performance data observed during the two high-prevalence bursts. A 2x2 (Training Condition x Burst) ANOVA was conducted for both hit rate and RT, given both of these performance metrics were observed to benefit from points-based feedback in Experiment 1b. While there was no significant difference observed for RT \((p > .05)\), the analyses revealed a significant main effect of condition on hit rate performance, \(F(1,28) = 5.04, p < .05\), \(\eta_p^2 = .15\) (see Figure 3.13). Participants receiving points-based feedback during the bursts had an overall higher hit rate \((M = 84.35\%, SD = 7.03\%)\) compared to traditional feedback \((M = 76.84\%, SD = 12.34\%)\). These results replicate the findings from Experiment 1b that points-based feedback during a higher prevalence inspection task can improve detection performance compared to traditional feedback.
Figure 3.13. Experiment 2 mean hit rate during two two-minute burst after second and fifth epochs during training session. ANOVA revealed significant main effect for condition, indicating significantly higher hit rate during gamified bursts than traditional bursts. Error bars are +/- one standard error of the mean.

Training Session Subjective Experience. A series of Kruskal-Wallis tests were conducted for participant ratings on the NASA-TLX to assess the hypothesis of greatest mental workload ratings in the control condition (H9b) and Flow Short-Scale to assess the hypothesis of lowest task engagement ratings in the control condition (H9c) and greater task engagement ratings for the gamified condition (H10). The analyses detected a significant difference on the NASA-TLX temporal workload subscale, $\chi^2(2, N = 44), = 6.63, p = .04$, but no significant differences on the Flow Short-Scale (all $p > .10$), indicating no support for H9c or H10. Follow-up tests were conducted to evaluate pairwise differences among the three groups on the temporal subscale. The cutoff for significance was reduced to $p = .008$ to control for Type I error. The results indicated a marginally significant difference between gamified training condition ($M = 13.21, SD$
= 3.27) and traditional training condition ($M = 16.56, SD = 3.37$), $p = .01$, indicating participants in the traditional training condition perceived the inspection task as being more temporally demanding than participants in the gamified training condition. Despite this outcome, the results do not support H9b (greatest mental workload ratings for the control condition).

*Testing Session Detection Performance.* A series of 3x6 (Training Condition x Epoch) mixed ANOVAs were conducted for the three primary detection performance measures of sensitivity, criterion, and hit rate to assess the hypotheses that the control condition would have the worst performance (H11) and the gamified group would have the best performance (H12). There was no significant main effect nor interacting effects of number of hours between training and testing sessions as a between-groups covariate. The analyses detected no significant main effect of either training condition nor epoch for any of the three detection performance measures, as well as no significant interactions for any of the detection performance measures (all $ps > .10$, see Figure 3.14), indicating no support for either H11 or H12.
Figure 3.14. Experiment 2 mean test session hit rate across all three conditions. Error bars are +/- one standard error of the mean.

Testing Session Subjective Experience. A series of Kruskal-Wallis tests were conducted for participant ratings on the NASA-TLX and Flow Short-Scale. The analyses detected no significant differences on either measure (all $p > .05$).

Transfer of Training Detection Performance. A series of 2x3x6 (Session x Training Condition x Epoch) mixed ANOVAs were conducted for the three primary detection performance measures of sensitivity, criterion, and hit rate to assess overall performance across sessions. The analyses detected no significant main effect of either session, training condition, nor epoch for any of the three detection performance measures, as well as no significant interactions for any of the detection performance measures (all $ps > .05$).

Discussion

The purpose of Experiment 2 was to investigate the differences between gamified, traditional, and control inspection training. It was hypothesized that during training (H9a) the
control group would have the worst detection performance, (H9b) highest workload ratings, and (H9c) lowest task engagement ratings, while (H10) the gamified group would have the highest levels of positive task engagement. The results of Experiment 2 training phase do not support any of these hypotheses.

It was further hypothesized that during the test phase (H11) the control group would have the lowest detection performance and (H12) the gamified group would outperform the traditional group. The test phase results do not support either of these hypotheses. Previous research has found that training with comprehensive feedback does not transfer to greater task success compared to limited feedback (Wiener, 1963). This helps to interpret the non-significant difference between the two experimental conditions, but does not help explain the null results between all three conditions. Most literature examining training transfer in a sustained attention task reports improved performance when transferring from a training session to a testing session (e.g., Warm et al., 2009). The low-prevalence (LP) effect in visual search literature, however, helps provide a different perspective.

The LP effect is a stubborn source of increased miss rates that is difficult to alleviate (Wolfe et al., 2007). One way to think about the null outcome is that the 2.9% target prevalence rate used in the current study yields such a strong LP effect that training or practicing, regardless of any training manipulation, provides little to no performance improvement. Previous research examining the LP effect among professional radiologists, whose job encompasses searches for rare critical signals, has found they are just as susceptible to increased miss rates (Evans et al., 2013). Additionally, neither novice nor expert radiologists improved search for rare targets over multiple testing sessions (Nakashima, Kobayashi, Maeda,
Therefore the lack of detection improvement in the current study may be due to the stubbornness of the LP effect that could not be overcome with one hour of training.

**Point System.** The point system created for the current study appeared in three experiments. In Experiment 1d, it did not result in any performance differences compared to traditional feedback during a 27-minute session with 3% target prevalence. In Experiments 1b and 2, however, it encouraged improved target detection during the two-minute 50% prevalence bursts.

The point system may have provided performance enhancements during the high prevalence bursts for a couple of reasons. First, points-based feedback has been demonstrated to enhance performance and improve task proficiency (Kapp, 2012; Richter et al., 2015). The point system was designed to encourage a liberal response criterion and therefore an increased target hit rate. This outcome was achieved without sacrificing overall task proficiency because there were no differences in false-alarm rates or criterion between the two feedback versions.

Second, previous research examining the effectiveness of various types of feedback in a vigilance task found that providing observers with comprehensive performance feedback (i.e., overall hit rate displayed at various intervals during the activity), in addition to immediate trial feedback, can enhance detection performance (Szalma, Teo, Hancock, & Murphy, 2011). The current point system allowed participants to gain an understanding of their overall performance, as well as how they were performing trial-to-trial. This level of feedback may have provided more meaningful information they could use towards developing an improved inspection strategy. Despite this benefit, the point system was ineffective during the 3% target
prevalence. Previously discussed was the interpretation provided by Feedback Intervention Theory (Kluger & DeNisi, 1996), in which the point system was likely less effective because of the higher demands of a 3% prevalence search compared to a 50% target prevalence search.

**Storytelling.** The results of Experiment 1c indicated participants provided with the NASA narrative showed stable target sensitivity during the course of the inspection task, while participants in the traditional instructions showed a change in target sensitivity over time. Unfortunately, this difference was not replicated in Experiment 2. Target sensitivity performance during the training session was no different over time between the three conditions. Research examining semester effects provides insight into why the manipulation did not replicate.

Semesters at Wichita State University last 16 weeks. The data for Experiment 1c was collected during the last two weeks of the semester, while data for Experiment 2 was collected during the first five weeks of the following semester. Empirical evidence has revealed that, in university research using undergraduate psychology students, both performance and motivation in a sustained attention task are lower during later weeks of the semester (Nicholls, Loveless, Thomas, Loetscher, & Churches, 2015). Therefore it is possible Experiment 1c participants were poorer performers going into the activity compared to Experiment 2 participants. This notion was confirmed with supplemental analyses comparing hit rate performance between the 2 control conditions of both experiments. A 2x3 (Semester Time X Epoch) ANOVA detected a modest main effect of semester time, $F(1,32) = 3.50$, $p = .07$, $\eta^2_p = .10$ (see Figure 3.15). This result, along with evidence presented in Nicholls et al. (2015), helps clarify the failure to replicate Experiment 1c results. If indeed the participants in 1c were less
able to perform the search task and/or less motivated, then the NASA narrative was still effective at improving performance among late semester poor performers. Likewise, early semester participants from Experiment 2 were better performers, so the NASA narrative was ineffective at improving preexisting good performance.

![Hit Rate](image)

**Figure 3.15.** Comparing hit rate performance between late-semester participants from Experiment 1c and early-semester participants from Experiment 2. All participants performed the control version of the inspection task featuring neither gamified nor traditional interventions. Error bars are +/- one standard error of the mean.

The late semester 1c participants in the storytelling condition also rated the task as being more demanding on two workload subscales compared to participants in the traditional instructions condition. This result was not replicated in Experiment 2 either. The Experiment 1c results previously discussed mention how Prospect Theory of decision making may suggest participants receiving the NASA narrative emphasized disastrous consequences (losses) more than a successful mission (gains), thus producing greater workload. Findings from additional
semester effects research support this explanation. Grimm, Markman, and Maddox (2012) report, when completing math problems, late-semester participants are risk-averse (i.e., they are more likely to minimize losses as opposed to maximize gains), while early-semester participants are more likely to be risk-seekers. Indeed if late semester participants in Experiment 1c emphasized avoiding losses, thus avoiding disastrous consequences for the space mission, the increased workload ratings likely reflect such mindset.

Results Summary

The current project investigated the potential benefits of applying gamification techniques to a simulated inspection task across five experiments. The initial four series of experiments compared single gaming elements to existing task manipulations known to impact performance. Experiment 1a found that badges provided a more rewarding experience but did not improve detection performance compared to traditional feedback. Experiment 1b found that a point system improved performance during a two-minute 50% target prevalence session but did not impact subsequent performance any differently than traditional feedback. Experiment 1c found that participants maintained stable target sensitivity throughout the course of the inspection when provided with a story to frame the activity. Experiment 1d found no benefits of using a point system as gamified feedback. Regrettably, Experiment 2 found no differences between gamified, traditional, or control training on the inspection ask. The following chapter presents meaningful discussion on the current findings, including the theoretical implications on vigilance, low prevalence visual search, and gamification.
CHAPTER 4
GENERAL DISCUSSION

The current project is built upon three theoretical influences: vigilance, low prevalence visual search, and gamification. Five total experiments were conducted to assess the potential benefits of applying gamification techniques to performance during a low-prevalence vigil. The initial four experiments assessed individual gaming elements: badges, points-based feedback during a burst of high target prevalence, storytelling, and points-based feedback throughout the inspection session. Experiment 2 assessed how gamified training compares to traditional and control training. The following section discusses both theoretical and practical implications of the current findings.

Vigilance & The Low-Prevalence Effect

The simulated inspection task created for the current investigation was an attempt to study vigilance inspection in an academic research environment. As previously discussed, vigilance research has often been criticized for using tasks with low external validity to study vigilance decrements (e.g., Hancock, 2013). Tasks featuring unnatural stimuli (e.g., a blinking rectangle that subtly alters in size to indicate a critical signal) oftentimes do not represent how sustained attention activities are conducted in the real world. The current project attempted to address these concerns when creating the simulated inspection task. The activity incorporated several important factors that define vigilance tasks, such as high event rate, low signal probability, and machine pace. Together, these elements created a monotonous task. Despite efforts to observe a performance decrement, however, data from 12 total experiments, across seven pilot experiments and five dissertation experiments, suggest no drop in performance
over time in this task. Findings and discussions from other realistic vigilance research provide helpful understanding.

Previously discussed were findings from Szalma et al. (2014) in which no vigilance decrement was observed during a task featuring first-person video-game characteristics. The dynamic nature of the Szalma et al. (2014) task, as well as the animated nature of the current task, may explain the lack of performance reduction over time. Observers may be able to maintain a higher level of alertness throughout the course of the inspection because arousal does not reduce over time. While the current simulated inspection task never produced a vigilance decrement, the observations still provide valuable lessons towards studying sustained attention. Human factors researchers are often faced with the challenge of studying psychological tendencies within the confines of academic settings. Efforts should continue to be made towards designing laboratory tasks that resemble realistic activities.

The low-prevalence (LP) effect observed in the present study proved to be persistent. The most effective intervention across all studies was the burst of high-prevalence trials (also see Wolfe et al., 2007). An interesting observation to this intervention was that performance during the burst did not have any impact on performance during the subsequent LP trials. The gamified burst featuring points produced greater detection performance during the high-prevalence trials, but the improved performance made no difference to LP detection performance. However, both types of burst interventions did generate improved detection during the subsequent LP epochs. This finding suggests that performance during a high-prevalence burst is unimportant towards improving performance during a LP search. The criterion shift that occurs during the burst is perhaps more sensitive to overall target probability
rather than actual detection performance. This strengthens the value of a high-prevalence burst intervention because the benefits may be achieved regardless of detection performance.

**Gamification**

The goal of gamification in the current project was to improve performance in a simulated inspection task through enhanced task engagement and increased self-efficacy. While the current results indicate this goal was marginally obtained, they should not be used to suggest gamification efforts are generally unsuccessful. Instead, they should emphasize (1) substantial miss rates during low-prevalence search tasks are exceedingly difficult to overcome, and (2) effective gamification application requires even more astuteness than we previously considered. The gaming elements studied in the current project were designed and applied with the demands of the specific inspection task in mind. The following section discusses the important takeaways regarding the unfavorable cost-benefit outcome of badges, limited effectiveness of the point system, and individual differences revealed by the storytelling element.

**Badges.** The results of Experiment 1a revealed that badges as a gamified form of feedback during a simulated inspection task did not encourage improved performance for rare targets. Furthermore, discussed in the previous chapter was the rigorous development of the badges as applied to the current task. These details provide important information for task designers considering implementing badges.

The characteristics of the simulated inspection task created a demanding environment for designing badges as a form of gamified feedback. It is likely that the monotonous nature of the task does not allow for badges to be an effective feedback mechanism. Tasks requiring
immediate performance feedback do not appear to benefit from badges. The existing empirical evidence of the success of badges has been observed in less monotonous and more socially engaging activities such as the social media application FourSquare or contributing to Wikipedia entries (Antin & Churchill, 2011). With this in mind, perhaps badges may be more beneficial in a search task when awarded during a rest break instead of during the repetitive trial-to-trial experience. Further research is needed to assess how different badge designs and payoff schedules may benefit detection performance.

**Point System.** The point system used in the current study was an effective gaming element providing improved target detection performance during 50% prevalence. Additionally, unlike the badges, the point system was a straightforward gaming intervention to implement. The point system payoff matrix was designed to encourage “target present” responses, thus providing improved target detection in 50% prevalence. An advantage of using a point system is the flexibility of being able to restructure the payoff matrix to encourage a different response bias. If task designers wanted to encourage “target absent” responses, they could design the payoff matrix to reward correct rejections greater than hits. Despite these advantages, however, the point system used in the current study showed limited effectiveness. While it did improve detection performance during a 50% prevalence burst, it was ineffective at improving performance during 3% target prevalence.

The point system was also the only gaming element tested at 50% target prevalence. Therefore it is currently unknown if the other gaming elements assessed in the initial series of experiments, badges and storytelling, may be as reliably effective in 50% prevalence conditions. Perhaps badges would have an impact on detection performance during a search for highly
prevalent targets. Further research is needed to understand the relationship between gamification and different levels of target prevalence within a simulated inspection task.

If task designers are thinking about implementing gamification techniques, they should strongly consider using a point system. Besides being the most widely-used gaming element in gamification, the flexible nature of point system design allows for a straightforward application (Hamari et al., 2014; Zichermann & Cunningham, 2011). Additionally, the current point system produced a strong effect size indicating the substantial impact on performance in comparison to traditional feedback. Points during high target prevalence was a reliably significant and meaningful intervention for improving performance. The outcomes of the current project suggest a point system would be a potentially successful gamification technique when applied appropriately.

**Storytelling.** Of the individual gaming elements, the most effective outcome during a low-prevalence vigil came when the population was primarily poor performers and likely less motivated (late-semester participants in Experiment 1c). It could be that gamification is most effective for these types of individuals. Research findings by game designer Jane McGonigal supports this viewpoint. She has argued that the psychological benefits of game play, such as improved self-efficacy, are experienced most by discouraged or demoralized individuals (McGonigal, 2011). To demonstrate this argument, McGonigal and her colleagues created a web- and mobile-based application known as SuperBetter using game design principles that allow people to track personal wellness goals (Roepke et al., 2015). For instance, if an individual sets a goal of completing a 5K race, they can using SuperBetter to set and track the goal. They will encounter skill-based challenges or quests tracking their fitness progress towards achieving
their goal. McGonigal and her colleagues have observed individuals using SuperBetter, particularly those suffering from depression, can improve their sense of personal well-being (Roepke et al., 2015). Therefore gamification may be most effective for unmotivated individuals seeking a more intrinsically rewarding experience. This result was observed among late-semester unmotivated poor performers in Experiment 1c who showed improvement when provided with a sense of purpose and meaning for their role in the task. While these results are suggestive, further research is required to determine whether gamification is indeed most effective for this special population.

The valuable implication learned from the current project is that providing a narrative for participants performing monotonous search tasks does provide a favorable cost-benefit outcome. Compared to the other two gaming elements assessed (badges and a point system), the storytelling element was the least effortful to implement. A single additional message presented during the instructions framed the task to be a meaningful activity. Furthermore, this simple addition provided performance benefits to unmotivated poor performers in Experiment 1c and did not hinder detection performance among good performers in Experiment 2. The storytelling element appears to be an effective and efficient gamification technique for potentially improving performance in a low-prevalence vigil.

Lastly, the storytelling manipulation in Experiment 1c produced a moderate effect size. This is important to note, particularly in comparison to points as the other effective gaming element. Points produced a large effect size when presented in inspections of high target prevalence. The moderate effect size of storytelling provided to poor performers in Experiment
1c and the large effect size of points provided during high prevalence provide valuable information towards revealing the boundaries of gamification in simulated inspection tasks.

**Conclusion**

To summarize, the current project was the first systematic investigation of gamification in a simulated inspection task with low prevalence targets requiring sustained vigilance. The initial series of experiments compared single gaming elements with existing interventions known to impact performance or experience. Badges as a gamified version of performance feedback did not change performance but did provide participants with a more rewarding experience. A points-based challenge presented in a burst of high target prevalence trials had no lasting effect in subsequent low target prevalence trials, but did improve performance during the burst. Inspection performance over time was more stable when the task was framed as a more purposeful activity in contrast to regular task instructions, particularly among poor performers. The results of Experiment 2 suggest gamified simulated inspection training does not transfer to improved performance compared to traditional or control training. In general, detecting rare targets is a difficult task resistant to improvement.

The current project presents mild benefits of gamification in an inspection task. Nevertheless, it is still worthwhile for human factors researchers to embrace the challenge of improving demanding tasks. Human beings should be put in the best position for success, especially when performing difficult tasks. The fundamental goal of human factors is to improve performance and experience by applying the knowledge of human limitations and tendencies to task design (Wickens & Hollands, 2000). While achieving this goal can be challenging and
demanding tasks are particularly difficult to improve, we will continue to explore new ways to improve the human experience.
REFERENCES


Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If you don’t find it often, you often don’t find it: Why some cancers are missed in breast cancer screening. *PLoS One, 8*(5), e64366.


APPENDICES
APPENDIX A: NASA-Task Load Index (NASA-TLX)

**Mental Demand**
How mentally demanding was the task?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Very Low

**Physical Demand**
How physically demanding was the task?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Very Low

**Temporal Demand**
How hurried or rushed was the pace of the task?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Very Low

**Performance**
How successful were you in accomplishing what you were asked to do?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Perfect

**Effort**
How hard did you have to work to accomplish your level of performance?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Very Low

**Frustration**
How insecure, discouraged, irritated, stressed, and annoyed were you?

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Very Low
### APPENDIX B: Flow State Short Scale (FSS)

<table>
<thead>
<tr>
<th>Item</th>
<th>Strongly Disagree</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I felt I was competent enough to meet the high demands of the situation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I did things spontaneously and automatically without having to think.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I had a strong sense of what I want to do.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I had a good idea while I was performing about how well I was doing.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was completely focused on the task at hand.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I had a feeling of total control over what I was doing.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The way time passed seemed to be different from normal.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The experience was extremely rewarding.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I was not worried about what others may have been thinking of me or my performance.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>