American Psychological Association

Journal of Experimental Psychology:
Human Perception and Performance

0096-1523/18/$12.00

2018, Vol. 44, No. 9, 1348–1355
http://dx.doi.org/10.1037/xhp0000538

Can Vigilance Tasks Be Administered Online? A Replication and Discussion

Victoria L. Claypoole
University of Central Florida

Alexis R. Neigel
U.S. Army Research Laboratory, Aberdeen Proving Ground, Aberdeen, Maryland

Nicholas W. Fraulini
University of Central Florida

Gabriella M. Hancock
California State University- Long Beach

James L. Szalma
University of Central Florida

Recently, experimental studies of vigilance have been deployed using online data collection methods. This data collection strategy is not new to the psychological sciences, but it is relatively new to basic research assessing vigilance performance, as studies in this area of research tend to collect data in the laboratory or in the field. The present study partially replicated the results of a newly developed online vigilance task (Thomson, Besner, & Smilek, 2016). A sample of 130 participants completed the semantic vigilance task created by Thomson et al. (2016) in a research laboratory setting. The present results replicated Thomson et al. (2016) when nonparametric and corrected signal detection measures were used. We suggest that some vigilance tasks typically performed in the laboratory could be administered online. However, we encourage researchers to consider the following factors prior to studying vigilance performance online: (a) the type of vigilance task, (b) the length of the vigilance task, and (c) the signal detection indices most appropriate for their research. It is quite possible that some analyses may yield significant results, whereas other signal detection measures may not (i.e., parametric indices vs. nonparametric indices vs. "corrected" indices) and this point is discussed further in our article.

Public Significance Statement

The present research demonstrates the importance of critically thinking about data collection strategies, as well as the usage of the appropriate signal detection indices when assessing vigilance performance. We provide a laboratory-based replication of Thomson et al.'s (2016) online study, and replicate similar performance declines in the laboratory. However, the results yielded only partial support for the original study findings, and importantly, the replication of these results was determined by the signal detection index that was imputed into the statistical analyses. More specifically, the results of Thomson et al. (2016) were replicated using corrected measures of signal detection, but these results did not replicate when traditional measures of signal detection theory were utilized. These results have important implications for the assessment of vigilance performance, particularly when data collection is conducted online, and can shape future research on vigilance.

Keywords: semantic vigilance, signal detection theory, sustained attention, vigilance performance

In a recent publication, Thomson et al. (2016) presented evidence for a decrement in performance over time using a semantic vigilance task administered online. Historically, basic vigilance research has been conducted using laboratory-based or in-the-field data collection methodologies (Davies & Parasuraman, 1982; Warm, 1984), with applied research focusing largely on the latter strategy. Furthermore, basic vigilance research focused on extending theories of signal detection (a common metric used to analyze

This article was published Online First April 30, 2018.

Victoria L. Claypoole, Department of Psychology, University of Central Florida; Alexis R. Neigel, U.S. Army Research Laboratory, Aberdeen Proving Ground, Aberdeen, Maryland; Nicholas W. Fraulini, Department of Psychology, University of Central Florida; Gabriella M. Hancock, Department of Psychology, California State University- Long Beach; James L. Szalma, Department of Psychology, University of Central Florida.

Victoria L. Claypoole and Alexis R. Neigel shared first authorship.

De-identified data can be requested with permission from Alexis R. Neigel at alexis.neigel@gmail.com.

Correspondence concerning this article should be addressed to Victoria L. Claypoole, Department of Psychology, University of Central Florida, 4000 Central Florida Boulevard, Orlando, FL 32816. E-mail: victoria.claypoole@knights.ucf.edu

performance results; cf. Thomson et al., 2016) or theories of information processing (i.e., resource theory, mind-wandering theory, etc.) have rarely utilized online data collection methods. To date, only two studies have conducted basic vigilance research using online task administration (Ralph, Thomson, Seli, Carriere, & Smilek, 2015; Thomson et al., 2016).

Two studies are not sufficient to explain the efficacy of online data collection strategies for vigilance. The lack of online studies in the vigilance literature could exist for several reasons. First, the psychological and sociological sciences are subject to the "file drawer" problem, which plagues this area of research, and the field has been sharply criticized for this effect (for a more detailed discussion of this issue, see Reimers, 2013). Thus, it is possible, and likely probable, that other researchers have attempted to conduct basic research related to vigilance online, but these studies and results remain unpublished due to nonsignificant results (for replication studies conducted online, see PsychFileDrawer.org). Second, a plethora of laboratory studies on individual differences in vigilance performance (Caggiano & Parasuraman, 2004; Craig, 1984; Helton & Russell, 2011, 2013; Humphreys & Revelle, 1984; Lehman, Olson, Aquilino, & Hall, 2006; Matthews, Warm, Shaw, & Finomore, 2014; Sawin & Scerbo, 1995; Shaw et al., 2010; Szalma, 2009; Thackray, Bailey, & Touchstone, 1977), especially differences in task engagement (Matthews, 2016; Matthews, Warm, & Smith, 2017), have demonstrated that individual characteristics influence vigilance performance, and more broadly visual search (Ort, Fahrenfort, & Olivers, 2017). When conducting online experimental research, one must consider the lack of experimental control associated with participation time and location, as well as associated individual differences (i.e., need for autonomy, motivation, etc.). For example, it is impossible to know whether participants are completing the study cozily in bed or during a chaotic afternoon lecture. In that vein, it is consequently impossible to control for any hedonomic factors (Hancock, 2013, 2017) that may influence participants' choice of coping behaviors employed to make the vigilance experiment enjoyable or, at the very least, less monotonous or boring.

Finally, online data collection methods sacrifice perceptual uniformity across participants. Participants are not completing online vigilance protocols under the same environmental conditions (Helton, Matthews, & Warm, 2009), nor are they utilizing a uniformly lit or uniformly sized screen (Hashimoto, Kumashiro, & Miyake, 2003). Factors such as screen size, noise, and heat are known to effect vigilance performance; and again, these variables are uncontrolled in online basic research related to vigilance. This argument is not to say that online data collection strategies are inappropriate for basic research on vigilance, but rather differences in screen size, device type (e.g., desktop, laptop, mobile phone), and luminosity of each of these devices should be measured and controlled for to the extent possible. At the very least, these environmental characteristics need to be considered by the researchers.

These limitations could explain why so few basic research studies examine vigilance performance online. Thus, studies by Ralph and colleagues (2015), as well as Thomson and associates (2016), are among the first of their kind, though online testing methodologies have been implemented since 1995 (for a history of online experimentation see Reimers, 2013). As so few basic online research studies concerning vigilance performance exist, the present study seeks to replicate the results of Thomson et al. (2016) in a controlled laboratory environment.

Implications for Theories of Vigilance

Vigilance, or the ability to sustain attention over a prolonged period of time, has been an area at the forefront of basic and applied research for nearly 70 years (Davies & Parasuraman, 1982; Mackworth, 1948; Parasuraman, 1985; Warm, 1984). Most notably, vigilance is associated with a decline in performance over time, which is referred to as the vigilance decrement (Davies & Parasuraman, 1982; Jerison, 1970; Mackworth, 1948; See, Howe, Warm, & Dember, 1995; Warm, 1977). This decline is linked to a decline in correct detections (i.e., hit rate), longer response times, and decreased sensitivity over the course of the vigil. This replication is important in understanding how vigilance performance may change across administration modalities. In a broader sense, though, it informs our understanding of the signal detection processes underlying vigilance.

In their original article, Thomson et al. (2016) include a theoretical discussion of signal detection theory, which is informed by their empirical data collected online. Thomson et al. (2016) manipulated the signal-to-noise distribution in their online experiment in order to demonstrate that "shifts in response bias (the observer's "willingness to respond") over time can masquerade as a loss in sensitivity" (p. 70) and "that the metrics employed to measure observer sensitivity in modern vigilance tasks (derived from signal detection theory) are inappropriate and largely uninterpretable" (p. 70). This is an intriguing theoretical stance that challenges years of vigilance research (See et al., 1995). For example, in a comprehensive meta-analysis examining 42 studies over a 12-year period, See et al. (1995) provided evidence for a substantial and consistent sensitivity decrement in vigilance research and across task types. Despite this watershed publication, Thomson et al. (2016) still argue that "the evidence for a sensitivity decrement in modern vigilance tasks is extremely weak" (p. 74) and "actual evidence in favor of a sensitivity decline in vigilance tasks is extremely sparse" (p. 74).

Consequently, the Thomson et al. (2016) perspective on sensitivity has been met with sharp criticism. Fraulini, Hancock, Neigel, Claypoole, and Szalma (2017) described how the experimental paradigm employed by Thomson et al. (2016) creates several potential discrimination decisions for participants (i.e., signal vs. lure, and lure vs. distracter) and three possible measures of sensitivity (i.e., distracter vs. lure, lure vs. signal, and signal vs. distracter). The multiple sensitivities and response criteria can make the interpretation of any one of these measures difficult, as the remaining indices are likely to exert effects on participant decision-making.

The Present Study

Given nontraditional task administration methods and the theoretical implications for signal detection theory described in Thomson et al. (2016), we propose that a laboratory-based replication of Thomson et al. (2016) is necessary to assess the effects of both sensitivity and response bias in vigilance performance. The goals for the present research are to (a) determine whether basic vigilance tasks can be administered similarly online and face-to-face,

and (b) empirically investigate whether performance can be assessed similarly across traditional and novel (i.e., corrected) measures of signal detection theory.

Method

Participants

An a priori power analysis conducted using G*Power 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007) and conventional criteria (i.e., $\alpha = .05$, power = 0.80) indicated the minimum number of participants required for statistical power was equal to 125. In the present study, 130 (78 females; 52 males) undergraduate students were recruited from the psychology research participation system at the University of Central Florida. Participants completed this study for partial course extra credit. The average age of participants was 18.95 years (*Median* = 18.00 years, *SD* = 3.02 years, *Range:* 18–47 years). All participants reported normal or corrected-to-normal vision. Participants were asked to refrain from consuming caffeine prior to their participation in this study.

Measures

Dundee State Stress Questionnaire. The Dundee Stress State Questionnaire (DSSQ; Matthews, 2016; Matthews et al., 2002) was used to measure subjective levels of stress at pre- and posttask. The short version of the DSSQ consists of 20 items that measure three subsidiary factors: distress, task engagement, and worry (Matthews, 2016). Higher scores on each respective subscale indicate elevated distress, worry, or engagement in the task.

NASA-Task Load Index. The NASA-Task Load Index (NASA-TLX; Hart, 2006; Hart & Staveland, 1988) measures perceived workload across six dimensions including: mental demand, physical demand, temporal demand, performance, effort, and frustration. Participants provide a rating (0–100) for each scale, then complete 15 pairwise comparisons. The ratings and pairwise comparisons are then used to calculate a weighted measure of global workload; higher scores indicate greater global workload.

Stimuli

The list of stimuli used in the Thomson et al. (2016) task was requested and obtained with permission from David R. Thomson via e-mail communication. The stimuli were programmed into SuperLab stimulus presentation software, which was used to randomize the presentation of the stimuli on a desktop computer.

The stimulus set was identical to that used in Thomson et al. (2016) and was presented in the same manner as the original experiment (i.e., black background with white text and a fixation cross presented during interstimulus intervals). The standard task consisted of 10 critical signals (i.e., four-legged creatures such as "cougar," "llama," "squirrel," etc.) and 90 neutral trials (i.e., nonsignals such as "apple," "phone," "wire," etc.). The lure task consisted of 10 critical signals, 10 lures (i.e., non-four-legged creatures; such as "canary," "lobster," "walrus," etc.), and 80 neutral trials. This resulted in 100 trials per block across five periods on watch in each condition. Each word appeared for 200 milliseconds, followed by a fixation cross that remained visible for 1100 milliseconds, which resulted in a total response window of 1300 milliseconds. As indicated by Thomson et al. (2016), a critical signal (i.e., four-legged creature) was never immediately preceded or followed by another critical signal.

Apparatus and Laboratory Setup

Participants completed the study in a quiet laboratory room with the facilitation of a research assistant. Stimuli were presented using SuperLab software (Version 4.5) on a Dell Optiplex 745 desktop computer. Participants were seated approximately 50.8 cm from the computer screen in a uniformly lit, private cubicle. Up to two participants could complete the experiment simultaneously and independently, though it was most common to have only one participant complete the study at a given time. When two participants were in the laboratory, each participant was instructed to complete the task individually. Participants were told not to adjust the settings of the monitor (e.g., brightness, contrast, height, angle, etc.). Importantly, the research assistant was not present in the room during the vigil, but did return to the laboratory upon completion of the task to administer the posttask self-report measures.

Procedure

Upon arrival, participants were randomly assigned to either the standard task ($N = 67$) or lure task ($N = 63$). Participants were instructed to power down any electronic devices (e.g., cellphones, tablets, laptops, etc.) and surrender any timepieces (watch or otherwise) to the researcher prior to beginning the experiment.

First, participants reviewed the electronic informed consent form and then completed the pretask measures (i.e., pretask DSSQ). At this time, participants were instructed to ask any questions pertaining to the research experiment. If participants had no questions, then the experiment proceeded. The research assistant then prepared either the lure task or standard task on the computer and asked participants to review the instructions prior to beginning the task. The research assistant indicated that participants should alert them when the task had ended by stepping outside of the research laboratory. The researcher then left the laboratory at this time and waited in another room until alerted by participants approximately 12 min after beginning the task, or tasks in the case of multiple participants.

The instructional set used in the present study was identical to the task instructions provided in Thomson et al. (2016, p. 76). After the vigilance task concluded, the researcher returned to the room and opened the posttask measures (i.e., posttask DSSQ, NASA-TLX) for participants to complete. The NASA-TLX and posttask DSSQ were counterbalanced to control for order effects. Participants then completed demographics upon the conclusion of the study to avoid any priming or bias effects.

Stress and Workload Results and Discussion

We extend the Thomson et al. (2016) research by including subjective measures of participant stress (see Table 1). Stress is important in the assessment of vigilance performance and typically accompanies a decline in performance over time (Dillard et al., 2013; Matthews et al., 2002; Matthews, Szalma, Panganiban, Neubauer, & Warm, 2013; Matthews, Warm, Reinerman-Jones,

Table 1
Perceived Stress and Workload Across Task Types

Measure	Standard ($N = 67$)	Lure ($N = 63$)	t statistic	p value	Cohen's d
Pre-task engagement	19.29 (4.49)	19.39 (4.24)	−.13	.90	.02
Pre-distress	4.66 (3.36)	5.03 (3.25)	−.65	.52	.11
Pre-worry	16.52 (5.34)	17.52 (4.81)	−1.21	.27	.20
Post-task engagement	17.67 (5.14)	18.32 (5.01)	−.73	.47	.13
Post-distress	6.24 (5.07)	8.56 (5.22)	−2.57	.01	.45
Post-worry	14.63 (5.84)	15.41 (5.65)	−.78	.44	.14
Global workload	51.13 (17.72)	54.94 (15.14)	−1.31	.19	.23
Mental demand	53.54 (30.49)	59.43 (26.70)	−1.17	.24	.21
Physical demand	14.79 (18.69)	15.38 (17.94)	−.18	.86	.03
Temporal demand	57.00 (27.96)	61.65 (27.21)	−.96	.34	.17
Performance	52.42 (33.85)	55.90 (25.29)	−.67	.51	.12
Frustration	29.04 (27.53)	39.49 (30.35)	−2.06	.04	.36
Effort	50.19 (29.18)	58.89 (25.16)	−1.82	.07	.32

Note. Scores reported in parentheses indicate the standard deviation of the mean. NASA-TLX scores are reported raw and untransformed. These results did not reach significance using a Bonferroni-corrected alpha criterion to adjust for multiple comparisons.

Langheim, & Saxby, 2010; Matthews et al., 2010; Matthews et al., 2014; Szalma & Teo, 2012). Global workload was measured in Thomson et al. (2016), but not explicitly reported; we report participant ratings of workload and the factors influencing workload in Table 1.

Thomson et al. (2016) noted "a trend toward the lure condition being rated as more demanding, rushed, and frustrating than the standard condition," but these self-reports were not significantly different (p. 76). In the present replication, our results indicated that participants perceived the lure task to be substantially more frustrating and effortful than the standard task, though participants in the lure task reported only slightly more mental demand and temporal demand than participant in the standard task (as seen in Table 1). Global workload was similar between task types.

As for measures of stress, participants in the lure task reported more engagement with the task at posttest, which could indicate that the lure task is slightly more engaging or challenging, but this is a claim that would need to be explored further in future research. Like Thomson et al. (2016), none of these results reached statistical significance.

Performance Results and Discussion

Separate mixed measures ANOVAs were performed using period on watch as the within-subjects factor and type of task as the between-subjects factor to identify significant differences across several measures of performance: proportion of correct detections, number of false alarms, and response time. Like Thomson et al. (2016), we use a Huynh-Feldt epsilon statistic to correct for violations of sphericity. In some instances, the Levene's Test of the Equality of Variances was also violated, thus these results should be interpreted with caution (note that the presence or absence of such a violation was not reported in the original Thomson et al. [2016] publication). Table 2 includes an overview of performance across task types and administration methodologies.

A majority of the performance results were consistent between experiments; many of the trends reported in Thomson et al. (2016) were reproduced in the present laboratory study. Importantly, a traditional vigilance decrement was observed over time overall, as well as an increase in response time over period on watch, which is indicative of the vigilance decrement (Warm, Parasuraman, & Matthews, 2008). However, our results do not demonstrate a clear, linear decline, as indicated by Thomson et al. (2016), rather the standard task committed slightly fewer correct detections over time and the lure task committed significantly fewer correct detections over time with sporadic performance increases in Periods 4 and 5 on watch (i.e., initial decrease then increase over time after Period 3 of watch).

Interestingly, in both the Thomson et al. (2016) study and the present experiment, participants in the standard task reported significantly more correct detections than participants in the lure task. This finding is important, as the lure task is meant to increase only false alarm responses to facilitate tests of declines in sensitivity. However, the lures also affect correct detection performance, which implies that perhaps shifts in response bias and sensitivity are both influencing vigilance performance.

Signal Detection Results and Discussion

Separate mixed measures ANOVAs were performed using period on watch as the within-subjects factor and type of task as the between factor to identify significant differences in sensitivity and response bias, as well as corrected sensitivity and response bias (these results are reported in Table 3). Sensitivity was calculated using the proportion of correct detections and proportion of false alarms using A', d', and corrected A'. Response bias was calculated using the proportion of correct detections and proportion of false alarms using B_D", c, and corrected B_D''.

Thomson and colleagues (2016) propose a new statistical methodology for analyzing vigilance performance when false alarms are subject to a floor effect. These new signal detection indices are referred to as "corrected sensitivity" and "corrected response bias," and are calculated when parametric tests of signal detection are biased by low false alarm rates. In the present study, we followed of Thomson and colleagues' (2016) method for calculating these

Table 2
Trends in Vigilance Performance Across Task Type and Administration Method

Measure	Laboratory (N = 130)	Online (Thomson et al., 2016) (N = 133)
Correct detections	• Significantly more correct detections were indicated over time in the standard task compared with the lure task. • A sharp decline in performance over time in the lure task was observed (with some sporadic performance increases in certain periods), but in the standard task there was a small, but steady performance decrement.	• Significantly more correct detections were indicated in the standard task, than in the lure task. • A steady vigilance decrement was observed in both the lure and standard tasks.
Total false alarms	• Significantly more false alarms were committed in the lure task than in the standard task. • A significant decline in the number of distracter false alarms committed over time was observed across task types.	• More false alarms were committed in the lure task compared with the standard task. • False alarms significantly declined over time across task types.
Lure false alarms	• A significant decline in the number of lure false alarms committed over time was observed in the lure task.	• A significant main effect of watch period over time indicated that lure false alarms declined over time in the lure task.
Response time	• Significantly longer response times were observed in the lure task compared with the standard task. • Response times steadily increased overall as a function of time on task across conditions.	• Significantly longer response times for participants were observed in the lure task. • Response time increased overall across tasks types.

Note. The following significant results are reported for the present laboratory experiment. There was a significant main effect of task type on proportion of correct detections, $F(1, 128) = 10.84$, $p = .001$, $\eta_p^2 = .08$, and period on watch, $F(4, 125) = 14.90$, $p < .001$, $\eta_p^2 = .10$, $\varepsilon = .73$. There was a significant main effect of task type on the number of total false alarms, $F(1, 128) = 100.33$, $p < .001$, $\eta_p^2 = .44$, a significant main effect of period on watch on total false alarms, $F(4, 125) = 48.62$, $p < .001$, $\eta_p^2 = .28$, $\varepsilon = .72$, and a significant interaction between period on watch and task type on total false alarms committed, $F(4, 125) = 28.91$, $p < .001$, $\eta_p^2 = .18$, $\varepsilon = .72$. There was a significant main effect of period on watch on the number of lure false alarms committed over time, $F(4, 59) = 53.63$, $p < .001$, $\eta_p^2 = .46$, $\varepsilon = .86$. There was a significant main effect of task type on mean response time, $F(1, 128) = 66.71$, $p < .001$, $\eta_p^2 = .34$, and a significant main effect of period on watch, $F(4, 125) = 37.95$, $p < .001$, $\eta_p^2 = .23$, $\varepsilon = .75$.

corrected measures. The corrected number of false alarms was calculated by adjusting the proportion of lure false alarms entered into traditional signal detection analyses. The aforementioned proportion of lure false alarms was calculated by dividing the number of lure false alarms committed by each participant by 10 lure stimuli.

To obtain corrected statistics, separate one-way repeated-measures ANOVAs were performed for sensitivity and response bias over time.

Table 3
Trends in Vigilance Performance Across Task Type, Administration Method, and Each of the Relevant Signal Detection Indices

Measure	Laboratory (N = 130)	Online (Thomson et al., 2016) (N = 113)
Sensitivity	• The utilization of the A' and d' indices indicated that participants in the lure task demonstrated less perceptual sensitivity toward the task stimuli than participants in the standard task. • The A' and d' indices indicated an increase in perceptual sensitivity overall over time across the task types. • The corrected A' index indicated a significant increase in perceptual sensitivity over time.	• The A' index indicated a sharp decline in sensitivity over time across the task types. • A decline in sensitivity over time was not observed using the d' index for lure false alarms only. • A numerical increase in sensitivity over time was observed using corrected A' (but see Figure 8A; p. 79, Thomson et al., 2016).
Response bias	• Participants in the standard task demonstrated more conservatism over time overall than participants in the lure task when both the B_D'' and c indices were used to compute response bias. • The corrected B_D'' statistic indicated a significant increase response bias over time overall in the lure task.	• The B_D'' index indicated an increase in conservatism over time across the lure and standard tasks. • The c index was not calculated or utilized in the original study to examine shifts in response bias. • The corrected B_D'' statistic indicated a significant increase response bias over time overall in the lure task.

Note. There was a significant main effect of task type on sensitivity as indexed by A', $F(1, 128) = 12.15$, $p = .001$, $\eta_p^2 = .09$, and a significant main effect of period on watch for this index, $F(4, 125) = 9.15$, $p < .001$, $\eta_p^2 = .07$, $\varepsilon = .72$. There was a significant main effect of task type on sensitivity using the d' index, $F(1, 128) = 57.94$, $p < .001$, $\eta_p^2 = .31$, there was also a main effect of period on watch using d', $F(4, 125) = 8.79$, $p < .001$, $\eta_p^2 = .06$, $\varepsilon = .88$, and a significant watch by task type interaction on sensitivity using d', $F(4, 125) = 2.61$, $p = .042$, $\eta_p^2 = .02$, $\epsilon = .88$. There was a significant main effect of task type on response bias as indexed by B_D'', $F(1, 128) = 76.68$, $p < .001$, $\eta_p^2 = .38$, main effect of period on watch for this statistic, $F(4, 125) = 45.70$, $p < .001$, $\eta_p^2 = .26$, $\varepsilon = .88$, and interaction between period on watch and task type using this statistic, $F(4, 125) = 24.76$, $p < .001$, $\eta_p^2 = .16$, $\varepsilon = .88$. There was a significant main effect of task type on response bias using the c index, $F(1, 128) = 8.25$, $p = .005$, $\eta_p^2 = .06$, as well as a main effect of period on watch using the c index to compute response bias, $F(4, 125) = 34.44$, $p < .001$, $\eta_p^2 = .21$, $\varepsilon = .79$, and a significant interaction between period on watch and task type using this statistic, $F(4, 125) = 11.75$, $p < .001$, $\eta_p^2 = .08$, $\varepsilon = .79$.

We used the corrected A' statistic and the corrected B_D'' statistic described in Thomson et al. (2016) to analyze lure *only* false alarm performance (i.e., ignoring nonsignal events).

Sensitivity

Both A' and d' (the latter is calculated using parametric test assumptions; Green & Swets, 1966; Macmillan & Creelman, 2005) indicated that participants completing the lure task displayed lower levels of perceptual sensitivity than those in the standard task, but perceptual sensitivity increased overall as a function of time on task for both groups. In the present study, a significant increase overall over time was indicated by both the A' (see Figure 1) and d' indices (see Figure 2) of sensitivity. Our results also indicated a significant period on watch by task type interaction using d', $F(4, 125) = 2.61$, $p = .042$, $\eta_p^2 = .02$, $\varepsilon = .881$, but this interaction was not significant when A' was used to compute sensitivity. This is perhaps the greatest disparity between our research and the original Thomson et al. (2016) study. Thomson et al. (2016) demonstrated a clear decline in sensitivity over time when using A' to calculate sensitivity, but interestingly sensitivity remained relatively stable when using d' to calculate sensitivity (for lure false alarms only, p. 78).

The results of a one-way ANOVA using corrected A', which assesses sensitivity to lure false alarms *only*, indicated a significant main effect of period on watch, $F(4, 248) = 16.374$, $p < .001$, $\eta_p^2 = .209$, $\varepsilon = .814$. This analysis indicated a significant increase in perceptual sensitivity over time. In Thomson et al. (2016), a nonsignificant main effect of period was reported using the corrected A' index, though the authors note that a numerical increase in sensitivity over time was observed when using corrected A' (p. 78), but their original figure (p. 79) indicates otherwise.

Response Bias

Both the B_D'' and c (that latter index is calculated using parametric test assumptions; Green & Swets, 1966; Macmillan & Creelman, 2005) indices of response bias indicated that participants in the standard task were generally more conservative in their responses than participants in the lure task (albeit, participants in the lure task were more conservative in their responses in Period 3 when the c index was utilized in the present study). Furthermore, there were significant main effects of period on watch when the B_D'' and c indices were used to calculate response bias, which demonstrated that responses overall trended toward greater conservatism over time across conditions. Figure 3 depicts response bias over time across conditions when the B_D'' index is used, and Figure 4 shows response bias over time across conditions when the c index is utilized.

A one-way repeated-measures ANOVA was performed for response bias over time for the lure task only. The corrected B_D'' statistic, which assesses lure false alarms *only*, indicated a significant main effect of period on watch, $F(4, 248) = 21.092$, $p < .001$, $\eta_p^2 = .254$, $\varepsilon = .971$. This result indicates a conservative shift in response bias over time toward lure stimuli for participants in the lure task. A similar main effect of period on watch was obtained when B_D'' was used in Thomson et al. (2016, p. 79).

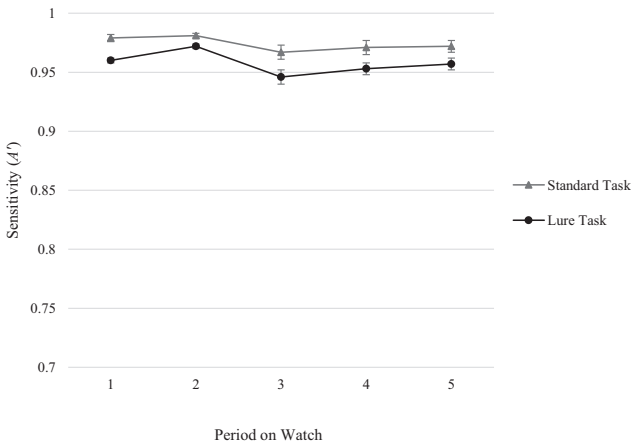

Figure 1. Sensitivity (A') as a function of time for the lure task and standard task ($N = 130$). Error bars reflect the standard error of the mean.

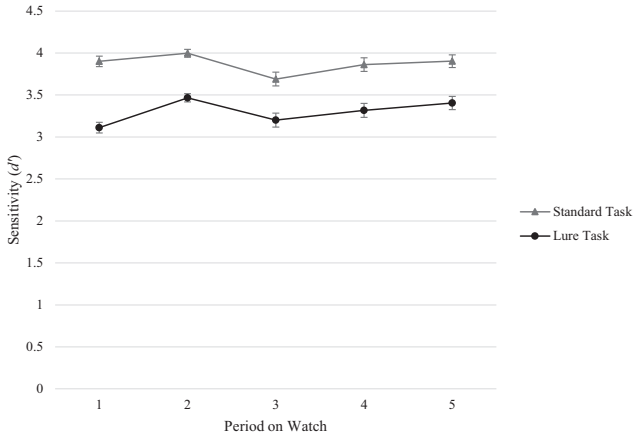

Figure 2. Sensitivity (d') as a function of time for the lure task and standard task ($N = 130$). Error bars reflect the standard error of the mean.

General Discussion

The purpose of the present study was twofold: (a) demonstrate a replication of the online vigilance task used in Thomson et al.

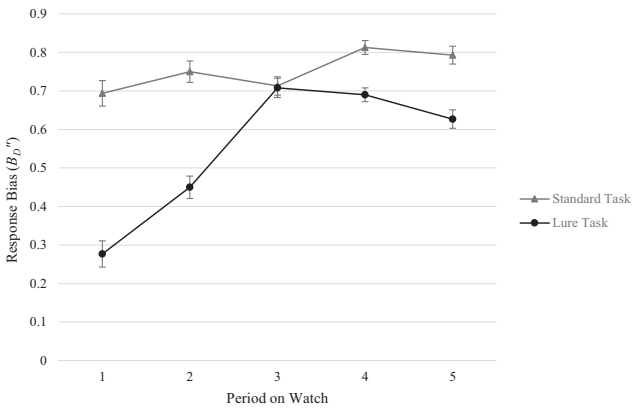

Figure 3. Response bias (B_D'') as a function of time for the lure task and standard task ($N = 130$). Error bars reflect the standard error of the mean.

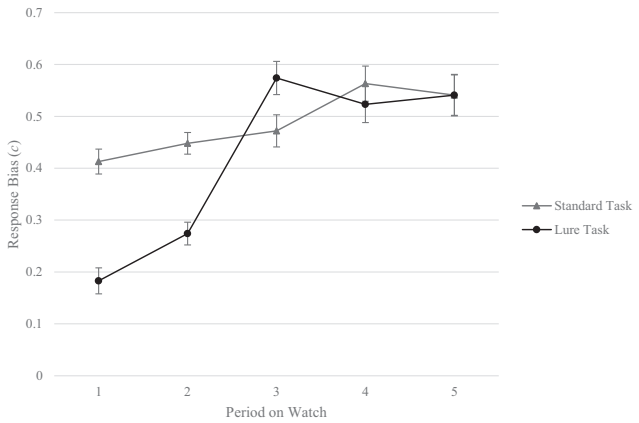

Figure 4. Response bias (*c*) as a function of time for the lure task and standard task ($N = 130$). Error bars reflect the standard error of the mean.

(2016) in the laboratory, and (b) explore the results that did or did not replicate based on the type of signal detection indices used to determine performance, sensitivity, or response bias. We also expand on the previous work by Thomson et al. (2016) by including self-report measures of perceived stress and workload for this semantic vigilance task, which are measures that typically accompany vigilance research.

Importantly, several of the laboratory results presented here were consistent with the results reported by Thomson et al. (2016). For example, vigilance performance (i.e., correct detections, false alarms, and response time) was nearly identical to Thomson et al. (2016). However, a steady vigilance decrement was observed in Thomson et al. (2016) across the conditions, whereas correct detection performance in the laboratory-based study tended to be more sporadic. There was a sharp decline in correct detection performance in the lure task, but this decrement demonstrated a moderate increase after the third watch period.

There were key differences between the results of the present laboratory study and online study when signal detection analyses were performed. This outcome is likely attributable to both the data collection methodology used and to the spread of signal detection indices that we have reported here. For example, we demonstrated that sensitivity significantly increased over time across conditions, and for the corrected sensitivity statistics, but a nonsignificant increase in sensitivity over time was reported by Thomson et al. (2016, p. 76). Similar trends were observed when examining response bias and each of the indices across task types: response bias tended to shift toward more conservatism over time in the lure and standard conditions.

The corrected signal detection metrics included in Thomson et al. (2016) may indeed be useful in that they isolate lures as a unique decision-making category. Unfortunately, this could introduce multiple decision-making criteria for participants randomly assigned to the lure task. These criteria include (a) comparisons between distracters (i.e., all neutral events) and lures and (b) comparisons between lures and signals (i.e., all target stimuli), inherently reshaping the underlying decision-making space for signal detection analyses (for a lengthy commentary on the efficacy of corrected signal detection indices, see Fraulini et al., 2017). The corrected statistics may also be pertinent to understanding comparisons between other possible measurements of sensitivity and response bias (nonsignal vs. lures, and overall nonsignals vs. signals; Macmillan & Creelman, 2005). We caution, though, against using these corrected statistics to inform the vigilance literature broadly.

Final Thoughts and Future Directions

As the insertion of lure stimuli into vigilance tasks alters the perceptual sensitivity and conservatism in response bias associated with vigilance performance, it will be important in future work to more closely scrutinize what drives the decision-making criterion strategies employed by participants in tasks utilizing lure stimuli. More specifically, future research will need to examine how individual differences may influence these shifts in sensitivity or response bias over the course of the vigil.

Furthermore, research past and present using lure stimuli (i.e., Thomson et al., 2016) demonstrates how performance and decision-making shift in the presence of these stimuli, but little is known about the perceptual features of the lure itself (i.e., the specific lure criteria stored in long-term memory). Thus, we beg the question: what makes a lure a lure, and not a distracter stimulus? Is it the perceptual similarity of the lure to the target stimulus? We encourage future research to consider this query, particularly in the study of alternative stimuli sets.

References

Caggiano, D. M., & Parasuraman, R. (2004). The role of memory representation in the vigilance decrement. *Psychonomic Bulletin & Review, 11,* 932–937. http://dx.doi.org/10.3758/BF03196724

Craig, A. (1984). Human engineering: The control of vigilance. In J. S. Warm (Ed.), *Sustained attention in human performance* (pp. 247–290). Chichester, UK: Wiley.

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. London, UK: Academic Press.

Dillard, M. B., Warm, J. S., Funke, G. J., Vidulich, M. A., Nelson, W. T., Eggemeier, T. F., & Funke, M. E. (2013). Vigilance: Hard work even if time flies. *Proceedings of the Human Factors and Ergonomics Society, 57,* 1114–1118. http://dx.doi.org/10.1177/1541931213571249

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Fraulini, N. W., Hancock, G. M., Neigel, A. R., Claypoole, V. L., & Szalma, J. L. (2017). A critical examination of the research and theoretical underpinnings discussed in Thomson, Besner, and Smilek (2016). *Psychological Review, 124,* 525–531. http://dx.doi.org/10.1037/rev0000066

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Hancock, P. A. (2013). In search of vigilance: The problem of iatrogenically created psychological phenomena. *American Psychologist, 68,* 97–109. http://dx.doi.org/10.1037/a0030214

Hancock, P. A. (2017). On the nature of vigilance. *Human Factors, 59,* 35–43. http://dx.doi.org/10.1177/0018720816655240

Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX): 20 years later. *Proceedings of the Human Factors and Ergonomics Society, 50,* 904–908. http://dx.doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52,* 139–183. http://dx.doi.org/10.1016/S0166-4115(08)62386-9

Hashimoto, M., Kumashiro, M., & Miyake, S. (2003). Effects of screen size and task difficulty on vigilance performance of older adults. *Journal of University of Occupational and Environmental Health, 25,* 375–386. http://dx.doi.org/10.7888/juoeh.25.375

Helton, W. S., Matthews, G., & Warm, J. S. (2009). Stress state mediation between environmental variables and performance: The case of noise and vigilance. *Acta Psychologica, 130,* 204–213. http://dx.doi.org/10.1016/j.actpsy.2008.12.006

Helton, W. S., & Russell, P. N. (2011). Working memory load and the vigilance decrement. *Experimental Brain Research, 212,* 429–437. http://dx.doi.org/10.1007/s00221-011-2749-1

Helton, W. S., & Russell, P. N. (2013). Visuospatial and verbal working memory load: Effects on visuospatial vigilance. *Experimental Brain Research, 224,* 429–436. http://dx.doi.org/10.1007/s00221-012-3322-2

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review, 91,* 153–184. http://dx.doi.org/10.1037/0033-295X.91.2.153

Jerison, H. J. (1970). Vigilance, discrimination and attention. In D. I. Mostofsky (Ed.), *Attention: Contemporary theory and analysis* (pp. 127–147). New York, NY: Appleton.

Lehman, E. B., Olson, V. A., Aquilino, S. A., & Hall, L. C. (2006). Auditory and visual continuous performance tests: Relationships with age, gender, cognitive functioning, and classroom behavior. *Journal of Psychoeducational Assessment, 24,* 36–51. http://dx.doi.org/10.1177/0734282905285238

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *The Quarterly Journal of Experimental Psychology, 1,* 6–21. http://dx.doi.org/10.1080/17470214808416738

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory* (2nd ed.). Mahwah, NJ: London Erlbaum Associates.

Matthews, G. (2016). Multidimensional profiling of task stress states for human factors: A brief review. *Human Factors, 58,* 801–813. http://dx.doi.org/10.1177/0018720816653688

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., . . . Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion, 2,* 315–340. http://dx.doi.org/10.1037/1528-3542.2.4.315

Matthews, G., Szalma, J. L., Panganiban, A. R., Neubauer, C., & Warm, J. S. (2013). Profiling task stress with the Dundee Stress State Questionnaire. In L. Cavalcanti & S. Azevedo (Eds.), *Psychology of stress* (pp. 49–91). Hauppauge, NY: Nova Science Publishers, Inc.

Matthews, G., Warm, J. S., Reinerman-Jones, L. E., Langheim, L. K., & Saxby, D. J. (2010). Task engagement, attention, and executive control. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition: Attention, memory, and executive control* (pp. 205–230). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4419-1210-7_13

Matthews, G., Warm, J. S., Reinerman-Jones, L. E., Langheim, L. K., Washburn, D. A., & Tripp, L. (2010). Task engagement, cerebral blood flow velocity, and diagnostic monitoring for sustained attention. *Journal of Experimental Psychology: Applied, 16,* 187–203. http://dx.doi.org/10.1037/a0019572

Matthews, G., Warm, J. S., Shaw, T. H., & Finomore, V. S. (2014). Predicting battlefield vigilance: A multivariate approach to assessment of attentional resources. *Ergonomics, 57,* 856–875. http://dx.doi.org/10.1080/00140139.2014.899630

Matthews, G., Warm, J. S., & Smith, A. P. (2017). Task engagement and attentional resources. *Human Factors, 59,* 44–61. http://dx.doi.org/10.1177/0018720816673782

Ort, E., Fahrenfort, J. J., & Olivers, C. N. L. (2017). Lack of free choice reveals the cost of having to search for more than one object. *Psychological Science, 28,* 1137–1147. http://dx.doi.org/10.1177/0956797617705667

Parasuraman, R. (1985). Sustained attention: A multifactorial approach. In M. Posner (Ed.), *Attention and performance: XI*. Hillsdale, NJ: Erlbaum.

Ralph, B. C., Thomson, D. R., Seli, P., Carriere, J. S., & Smilek, D. (2015). Media multitasking and behavioral measures of sustained attention. *Attention, Perception & Psychophysics, 77,* 390–401. http://dx.doi.org/10.3758/s13414-014-0771-7

Reimers, S. (2013). Developments in information technology and their implications for psychological research: Disruptive or diffusive change? *Learning at City Journal, 3,* 45–53.

Sawin, D. A., & Scerbo, M. W. (1995). Effects of instruction type and boredom proneness in vigilance: Implications for boredom and workload. *Human Factors, 37,* 752–765. http://dx.doi.org/10.1518/001872095778995616

See, J. E., Howe, S. R., Warm, J. S., & Dember, W. N. (1995). Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin, 117,* 230–249. http://dx.doi.org/10.1037/0033-2909.117.2.230

Shaw, T. H., Matthews, G., Warm, J. S., Finomore, V. S., Silverman, L., & Costa, P. T., Jr. (2010). Individual differences in vigilance: Personality, ability and states of stress. *Journal of Research in Personality, 44,* 297–308. http://dx.doi.org/10.1016/j.jrp.2010.02.007

Szalma, J. L. (2009). Individual differences in human-technology interaction: Incorporating variation in human characteristics into human factors and ergonomics research design. *Theoretical Issues in Ergonomics Science, 10,* 381–397. http://dx.doi.org/10.1080/14639220902893613

Szalma, J. L., & Teo, G. W. L. (2012). Spatial and temporal task characteristics as stress: A test of the dynamic adaptability theory of stress, workload, and performance. *Acta Psychologica, 139,* 471–485. http://dx.doi.org/10.1016/j.actpsy.2011.12.009

Thackray, R. I., Bailey, J. P., & Touchstone, R. M. (1977). Physiological, subjective, and performance correlates of reported boredom and monotony while performing a simulated radar control task. In R. R. Mackie (Ed.), *Vigilance: Theory, operational performance, and physiological correlates*. New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4684-2529-1_12

Thomson, D. R., Besner, D., & Smilek, D. (2016). A critical examination of the evidence for sensitivity loss in modern vigilance tasks. *Psychological Review, 123,* 70–83. http://dx.doi.org/10.1037/rev0000021

Warm, J. S. (1977). Psychological processes in sustained attention. In R. R. Mackie (Ed.), *Vigilance: Theory, operational performance and physiological correlates* (pp. 623–644). New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4684-2529-1_30

Warm, J. S. (1984). *Sustained attention in human performance*. Chichester, UK: Wiley.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors, 50,* 433–441. http://dx.doi.org/10.1518/001872008X312152

Received May 24, 2017
Revision received January 12, 2018
Accepted January 17, 2018 ■