WELHAF, MATTHEW S. Ph.D. Evaluating the Construct Validity of Sustained Attention
Measures: Performance Indicators, Self-Report Indicators, and Their Covariation. (2022)
Directed by Dr. Michael J. Kane. 289 pp.

The ability to sustain attention is a fundamental cognitive process that is required for many everyday activities. Current measurement approaches focus on either objective behavioral indicators (like reaction time [RT] variability or task accuracy) or subjective self-reports of task-unrelated thoughts (TUTs) as being suitable assessments for sustained attention. However, both types of indicators come with their own unique sources of measurement error, which reduce our accuracy in measuring sustained attention ability and weaken the conclusions we can draw from their findings. In this integrated dissertation, three papers are presented to argue that the covariation between objective and subjective indicators is a more construct-valid way to measure the ability to sustain attention than is either indicator type on its own. The results generally supported this claim, with some caveats. Theoretical implications, remaining concerns, and future directions are discussed to further improve the measurement of sustained attention ability.

EVALUATING THE CONSTRUCT VALIDITY OF SUSTAINED ATTENTION MEASURES: PERFORMANCE INDICATORS, SELF-REPORT INDICATORS, AND THEIR COVARIATION

by

Matthew S. Welhaf

A Dissertation
Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro

2022

Approved by

Dr. Michael J. Kane
Committee Chair

APPROVAL PAGE

This dissertation written by Matthew S. Welhaf has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. Michael J. Kane

Committee Members

Dr. Paul Silvia

Dr. Dayna Touron

Dr. Jasmine DeJesus

September 1, 2022

Date of Acceptance by Committee

September 1, 2022

Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to first thank my advisor, Dr. Mike Kane. His mentorship not only gave me the skills to become a more productive and skeptical scientist and researcher, but also a more thoughtful person. His guidance made it possible for this integrated dissertation and its experiments, and all the other projects completed during my time at UNCG possible. I would also like to thank the members of my committee, Drs. Dayna Touron, Paul Silvia, and Jasmine DeJesus. Your feedback and discussions during the prelim and dissertation process challenged my thinking and made the integrated dissertation process as smooth as possible. I truly appreciate your valuable insight and feedback. I would also like to thank my undergraduate and Masters advisor, Dr. Jonathan Banks, and lab mate Audrey Hood. I would not be where I am today without your guidance and continued support.

I would like to thank my friend and fellow graduate student Liz Gilbert. Liz, while I was originally your first-year liaison, your continued support and encouragement is something I will always be grateful for.

Finally, I would like to thank my partner, Brynn Hudgins. You came on board during the last leg of my time as a graduate student and were there during the highs and lows. Thank you for supporting me in finishing this important part of my academic career. I cannot wait to return the favor as you finish your degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTEGRATIVE INTRODUCTION

People's attention occasionally drifts away from what they are currently engaged in, which can result in negative real-world consequences, such as driving accidents (Broadbent et al., 1982; Reason & Mycielska, 1982; Yanko & Spalek, 2014), as well as professional (Hollenbeck et al., 1995; Reason, 1990) and academic difficulties (Lindquist & McLean, 2011; Steinmayr et al., 2010). Given the important role that sustained attention plays in such fundamental actions, understanding how and why lapses of sustained attention arise is critical for preventing such errors from occurring. However, the study of sustained attention has received little attention compared to other foundational cognitive abilities like executive functions and other aspects of attention (Esterman & Rothlein, 2019; Miyake & Friedman, 2012).

Contemporary research has identified distinct, yet correlated, empirical outcomes that reflect failures of sustained attention. Specifically, reaction time (RT) variability within simple tasks and subjects' reports of task-unrelated thoughts (TUTs) are often used as separate indicators of sustained attention failures and (in)abilities. As I discuss below, however, each of these indicators has their own unique sources of error that muddy the measurement of sustained attention. To remedy this, I argue that the individual-differences covariation of these measures may be a more construct valid approach to measuring sustained attention, an approach that the field has not yet considered. The goal of the research program presented in this dissertation is to investigate and evaluate the construct validity of sustained attention measures.

The first empirical paper (Welhaf et al., 2020b) examines the association between performance and self-report measures of sustained attention by testing the robustness of the worst performance rule. The worst performance rule refers to the empirical finding that subjects' cognitive ability is more strongly related to their worst performance, or longest RTs, compared to their best, or even average, performance or RTs. One account of the worst performance rule is that subjects' longest RTs occur, in part, because of momentary lapses of sustained attention (Larson & Alderton, 1990; Unsworth et al., 2010). As noted above, failures of sustained attention can also be measured via self-reported TUTs. In terms of the worst performance rule, then, we would expect to see ability correlations increasing in strength from shortest to longest RTs if the ability measure is closely linked to sustained attention, such as TUT rates. However, if the ability measure were somewhat less closely linked to sustained attention abilities, such as working memory capacity (WMC), we would expect a weaker worst performance rule pattern. Empirical Paper 1 also tested the robustness of our findings regarding sustained attention and the worst performance rule by using a "mini-multiverse" approach to outlier decisions and definitions. Although outlying RTs can be a sign of attention lapses, they can also reflect other processes or behaviors unrelated to sustained attention that impact performance on that given trial (e.g., momentarily forgetting the appropriate key to press or asking the experimenter questions during a task). Decisions on how to handle outlier trials (and subjects) can influence the findings of a study, so to increase the transparency of our analyses, we replicated our correlational models using different subject- and trial-level outlier-exclusion decisions.

The second empirical paper assesses the construct validity of sustained attention measures, and specifically their individual-differences overlap, from a nomothetic span (correlational) approach (Welhaf & Kane, 2022a). It reanalyzes two large-scale, latent variable studies where we could derive objective and subjective measures of sustained attention from multiple tasks (Kane at al., 2016; Unsworth et al., 2021). In addition to modeling RT variability and TUT rates as separate but correlated factors, we also modeled the shared variance in these indicators as the general ability to sustain attention. Each study also measured multiple nomological network constructs (e.g., WMC, motivational state, Big 5 personality traits) which allowed us to test the convergent and discriminant validity of the general sustained attention factor. As in empirical paper 1 (Welhaf et al., 2020b), we also employed a “mini-multiverse” analysis to assess the robustness of our findings across varying definitions and treatments of trial-level and participant-level outliers.

The third, and final, empirical paper further assesses the construct validity of sustained attention measures, but from a construct representation (experimental) approach (Welhaf & Kane, 2022b), by measuring RT variability and TUT rates in tasks of varying sustained attention demands. We asked whether experimentally manipulating the sustained attention demands of the tasks altered not only mean levels of RT variability and TUT rates, but critically their covariation, as well. If the individual-differences overlap in objective and subjective indicators of sustained attention is a more construct valid measure of sustained attention, then their correlation should be sensitive to theoretically derived manipulations. Specifically, when sustained attention demands are maximized, these two indicators should be at least moderately correlated with each other because variation in each measure is significantly caused by sustained attention ability. However, reducing the sustained attention demand of a task should weaken, if not eliminate, the correlation between these indicators because now any remaining between-person variance is primarily caused by nuisance factors unique to either indicator (i.e., *not* sustained attention processes or abilities). Collectively, this line of research seeks to extend our understanding and improve our measurement of the ability to sustain attention, which can in turn help identify individuals who may be especially prone to distraction and potentially reduce the likelihood of human error in everyday life.

The goal of this integrated dissertation is to present work testing whether the individual-differences covariation in objective and subjective measures is a construct valid way to measure the ability to sustain attention. To do this, I first define and dissociate the current view of sustained attention from alternative views. I then provide a brief review of the two construct validation approaches that will be used in the empirical papers presented in this integrated dissertation—"nomothetic span" and "construct representation" (Embretson, 1983)—along with expected outcomes that would support the construct validity of these sustained attention measures, and specifically their individual-differences covariation. I then describe the two main measurement approaches that have been used in the sustained attention literature—"objective" and "subjective" measures—and critically consider some limitations of using either of these approaches in isolation for measuring sustained attention ability. I next present correlational and experimental evidence from the literature that seems to support using the individual-differences covariation in objective and subjective measures to assess sustained attention. Specifically, I review studies that suggest some constructs that should be theoretically related to, and manipulations that should theoretically affect, the ability to sustain attention.

Defining Sustained Attention As Attention Consistency

Before determining the best ways to assess sustained attention, it is important to define the construct and differentiate it from associated constructs, such as vigilance, vigilant attention, and processing speed (for reviews, see Esterman & Rothlein, 2019; Fortenbaugh et al., 2017). I will, therefore: (a) briefly review, and dissociate, other abilities related to sustained attention, (b) define sustained attention for the current study, *focusing on short-term consistency rather than long-duration changes*, and (c) discuss why sustained-attention ability is necessary.

Historically, research on sustained attention has relied on long-duration vigilance tasks, in which subjects watch or listen for rare target events to occur, with vigils typically ranging from many tens of minutes to a few hours. Here, the main variable of interest is the "vigilance decrement," or the drop over time in target detection or detection speed (Mackworth, 1950; Lim & Dinges, 2008; Parasauraman, 1986; Parasauraman & Davies, 1977). Recent work using shorter laboratory tasks has found vigilance decrements at smaller time scales (over the course of a few minutes; Dinges & Powell, 1985; Esterman et al., 2013, 2014; Unsworth & Robison, 2020). Although these long- and short-duration changes in sustained attention are likely related, the current review and studies will focus on short-term fluctuations (i.e., over the course of a few seconds or minutes), which are more likely to apply broadly to everyday settings and to the modest-duration tasks that are common to cognitive psychology research.

For example, Esterman et al. (2013) found that subjects shifted frequently between two attentional states across the duration of a task. Using RT data from a continuous performance task, Esterman et al. (2013) identified periods where subjects were "in the zone" or "out of the zone," based on short-term deviations from their average performance (i.e., being "in the zone" was defined as periods in which RTs were close to mean levels, whereas being "out of the zone" was defined as periods of time where RTs were highly deviant from mean levels). Each period lasted only from a few seconds to a minute (see also Esterman, et al., 2014; Rosenberg et al., 2013, 2015; Weismann et al., 2006).

Likewise, the contents of conscious thought are dynamic (see James's [1890] "stream of thought"), drifting from being activity-focused to being off-task. Subjects report being off-task roughly 30–50% of the time during simple laboratory tasks of modest duration, as well as during daily-life activities (e.g., Kane et al., 2016, 2017; Killingsworth & Gilbert, 2010; Unsworth &

McMillan, 2013; Unsworth et al., 2021), but with substantial variation around these mean estimates. Some subjects reporting very little mind wandering during a task (or day) and others reporting being off-task an overwhelming percentage of the time.

Subjects also show trait-like propensities for TUT-*content* fluctuation during periods of rest (Kucyi, 2018) and within demanding laboratory tasks (Welhaf et al., 2020a; Zanesco, 2020). For example, previous research has found that subjects switch between TUT-content categories (e.g., endorsing worry at one probe and then fanciful daydreams at the next probe) at different rates while performing simple attention tasks (Welhaf et al., 2020a). This variability in TUT-content stability was also related to individual differences in some cognitive ability and personality characteristics. Thus, subjects not only fluctuate between focusing on their current goal and thoughts unrelated to their current task, but they also appear to fluctuate in *what* they are mind wandering about. Understanding how these moment-to-moment fluctuations in attention occur can help us better understand how and why lapses arise, and the downstream consequences that such sustained attention failures produce.

For the present program of research, I define sustained attention as *the purposeful act of maintaining optimal task focus to successfully, and consistently, perform goal-relevant actions*. Note that this definition emphasizes two critical aspects not captured by all views of sustained attention. Specifically, some definitions of sustained attention fail to emphasize the consistency (vs. inconsistency) of attention and the optimization of processing for action. Robertson et al. (1997), for example, conceptualized sustained attention as "…the ability to self-sustain mindful, conscious processing of stimuli whose repetitive, non-arousing qualities, would otherwise lead to habituation and distraction by other stimuli" (p. 747; see also descriptions of "vigilant attention;" Langner & Eickhoff, 2013; Lim & Dinges, 2008; Robertson & O'Connell, 2010). This view emphasizes the role of stimulus characteristics in engaging attention but does not consider the role of sustained attention in preparation for action or response. Likewise, traditional vigilance views of sustained attention are concerned with relatively long-range performance trajectories over time and often overlook the moment-to-moment (in)consistency of attention. Sustained attention (from a vigilance perspective) may thus be defined as, “the ability of organisms to maintain their focus of attention and to remain alert to stimuli over prolonged periods of time” (Warm et al., 2008, p. 433).

This vigilance perspective typically measures sustained attention through accuracy or RT changes across the duration of the task by examining block-by-block changes. The current perspective of sustained attention ability instead focuses on the variation in trial-to-trial consistency in response or conscious focus as being more representative of the ability to sustain attention. Traditional vigilance tasks like the Mackworth Clock Test (Mackworth, 1948) also present a very different goal than do tasks typically used to assess attention consistency. The goal in these vigilance tasks is to respond to single targets that appear at random and unpredictable (infrequent) times, while tasks frequently used to measure attention consistency require responses on (most) every trial. Thus, even though the term “sustained attention” has been used to describe different phenomena, there are fundamental differences in the goals and definitions between sustained *vigilant* attention and sustained *moment-to-moment* attention. This current view of sustained attention is more aligned with the notion that “attention consistency” is a critical component of the human attention system (Unsworth & Miller, 2021).

Why is the ability to actively sustain attention, from moment-to-moment, necessary? Research suggests that human’s default cognitive state is one of exploration and openness to distraction. For example, Klinger (1971; 2009) argued that personal goals become and remain accessible as “current concerns” until they are fulfilled or abandoned. These concerns compete for attention while we are engaged in ongoing activities, with environmental cues triggering concern-related thoughts into awareness. This sensitivity of attention to current concerns is supported by several studies that have subtly primed self- or concern-related information. For example, in a dichotic listening paradigm, where streams of information are presented simultaneously through different channels, attention is often drawn to the channel where self- or concern-related information is presented (Bargh, 1982; Gollwitzer & Bargh, 1996; Klinger, 1978). Further, evidence from studies using daily thought-sampling techniques shows that people spend large portions of their daily lives mind-wandering about their concerns (Kane et al., 2007; Klinger & Cox, 1987).

Relatedly, a prominent theory of mind wandering, the *Control Failures × Current Concerns* view (McVay & Kane, 2010) argues that people experience mind wandering (in part) because these off-task, concern-related thoughts are continuously and automatically cued by stimuli in their immediate environment. These off-task thoughts may be inhibited or otherwise regulated, however, when executive control processes are adequately deployed to prevent access of such thoughts to conscious awareness. Subsequent research has supported this notion by examining the effects that experimentally cuing subjects’ personal concerns has on TUTs (e.g., Kopp et al., 2015; McVay & Kane, 2013; Vannuci et al., 2017). For example, McVay and Kane (2013) found that TUT reports were more frequent following triplets of personally relevant, goal-related words (compared to non-relevant triplets) inserted into a go/no-go task. As will be discussed in subsequent sections, however, these cuing manipulations have been used primarily for their effects on TUTs, with only minimal interest in objective behavioral changes that should also reflect sustained attention failures.

Collectively, these findings suggest that attention is easily, and naturally, drawn to personally relevant stimuli and information that is not necessarily relevant to the current activities we are engaged in. Thus, an active sustained attention ability is needed to counteract this default orientation of attention to self- or concern-related information and focus on, and successfully perform, the primary task in front of us when it is critical to do so.[1]

Two General Approaches to Construct Validation

Because psychological constructs are unobservable, we must rely on our measures to assess how individuals differ on such hypothetical traits, abilities, or tendencies. How well our measures capture variation in target constructs is a question of construct validity (Borsboom et al., 2004; Cronbach & Meehl, 1955; Embretson, 1983; Strauss & Smith, 2009). The following section briefly reviews two schools of thought on construct validity and construct validation: the "nomothetic span" (correlational) approach and the "construct representation" (experimental) approach (Embretson, 1983). A goal of the current dissertation study is to use both approaches, in a complementary fashion, to assess the construct validity of sustained attention measures.

[1] Although beyond the scope of the current review, certain brain systems also appear to support the necessity of an active sustained attention ability. The locus coeruleus-norepinephrine (LC-NE) system fluctuates between modes of exploitation (i.e., optimal performance via task focus) and exploration (i.e., openness to task disengagement; Aston-Jones & Cohen, 2005; Cohen et al., 2004; Usher et al., 1999). LC-NE activity follows an inverse-U pattern. On the low end of the curve, tonic LC firing is weak and associated with hypoarousal and distractibility. At the peak of the curve, where both tonic and phasic firing are heightened, goal-directed behavior is optimal (i.e., an exploitive state). Here, organisms are most selective in their processing of stimuli. At the upper end of the curve, overactive tonic activity results in hyperarousal and indiscriminate responding (an exploration state). Sometimes hyperarousal can be good because it allows for the organism to seek or unexpectedly gain new information or rewards from the environment. However, this state of hyperarousal can hinder specific goal-related actions. There is ultimately a trade-off between these two states and the system must adapt to the current demands.

To evaluate a measure's construct validity, Cronbach and Meehl (1955) proposed that a nomological network for a given measure should be built and tested against data. The network should make predictions about how observable measures of the same, versus different, constructs relate to each other, how theoretical constructs relate to observed measures, and how scores on measures should change across different theoretically relevant contexts. Because the nomological network describes *relations* among constructs and observables, Cronbach and Meehl (1955) suggested examinations of correlation matrices, comparisons of groups, comparisons across time, and factor analytic approaches as methods for assessing and developing the nomological network and evaluating the validity of its measures (for an example of a construct-validation method using correlational data, see the multitrait-multimethod matrix; Campbell and Fiske, 1959).

The construct representation (experimental) approach to construct validity focuses, instead, on understanding the response processes that cause variation in scores on psychological tasks or measures (Embretson, 1983; Strauss & Smith, 2009). Embretson (1983) argued that such construct validation should focus on task decomposition, attempting to (a) identify and characterize the theoretical mechanisms (i.e., processes, strategies, and knowledge) that cause variation in task responses or performance, and then (b) design items or tests from this understanding (see also Borsboom et al., 2004). The construct representation approach primarily relies on experimental methods and computational models of the cognitive or affective processes that cause variation in performance outcomes. These techniques allow for researchers to pinpoint different task parameters or processes that can be manipulated across a task to see whether performance changes in ways that theory predicts (Embretson & Gorin, 2001).

Regarding sustained attention measures, a nomothetic span approach to construct validation could take the form of a latent variable study that includes multiple proposed measures of sustained attention and of other constructs that should appear within the nomological network. Supportive evidence from this approach would be that the proposed measures of sustained attention correlate strongly with each other and all load onto a sustained attention latent variable. Further, the sustained attention factor should correlate weakly, if at all, with measures that are not central to the nomological network (e.g., extraversion; word knowledge), while being moderately correlated with theoretically relevant factors (e.g., neuroticism; inhibitory control). From a construct representation approach, studies would aim to manipulate task parameters that should reduce (or increase) the need for, or success of, sustained attention mechanisms. Supportive evidence from the construct representation approach to construct validation would be to show that indicators of sustained attention can be impacted by changing aspects of the task that theory dictates should be sensitive to sustained attention (e.g., stimulus pacing; performance incentives).

Measuring Attention Consistency

In the following sections I describe two common approaches the field has used to assess sustained attention (in)abilities. I refer to these as *objective* (performance-based) and *subjective* (self-report based) indicators of attention consistency. Empirical Paper 2 (Welhaf & Kane, 2022a) provides a more detailed discussion of these indicators and so I only briefly review them here.

Objective Measures of Attention Consistency

Ideal tasks for assessing sustained attention should require subjects to endogenously maintain focus and respond in a timely and consistent manner; failing to do so should result in poor task performance. Additionally, tasks should be simple to understand and easy to accomplish at the trial level when subjects are paying full attention; that is, they should not require complex cognitive processing or place heavy demands on memory or reasoning ability. Finally, tasks should not be overly engaging; that is, people should not be so absorbed that the task exogenously captures and holds their attention. Variation in performance on tasks that meet these criteria should thus be due more to sustained attention than to other constructs.

Tasks like the psychomotor vigilance task (PVT; Lim & Dinges, 2008; Parasuraman et al., 1998), sustained attention to response task (SART; Robertson et al., 1997), metronome response task (MRT; Anderson et al., 2020; Laflamme et al., 2018; Seli, Cheyne, et al., 2013), continuous temporal expectancy task (CTET; O'Connell et al., 2009), and gradual onset continuous performance task (gradCPT; Rosenberg et al., 2013) all meet these criteria. These tasks require quick and consistent responding for successful performance, all present simple instructions (e.g., "respond when you see X on screen"), and all are repetitive enough that they are unlikely to completely engage subjects for their total duration. Table 1 provides brief descriptions of each of these tasks, how sustained attention contributes to successful performance in each task, and what cognitive processes beyond sustained attention might also influence their performance—as no psychological measures are process-pure.

Table 1. Descriptions of Commonly Used Sustained Attention Tasks

Task/Citation	Description	How is Sustained Attention necessary?	Non-Sustained Attention influences
Psychomotor Vigilance Task (PVT; Lim & Dinges, 2008)	Simple RT task that presents subjects with a set of 0s on-screen (like a stopwatch: "00.000") and requires them to respond as quickly as possible when they notice that the numbers begin counting up after a variable delay	Necessary to maintain task focus/engagement and intrinsic alertness during unpredictable periods between the start of the trial and stimulus onset. Failing to sustain attention would result in longer than normal RT	Processing speed; SOA guessing strategy; impulsivity
Sustained Attention to Response Task (SART; Robertson et al., 1997)	A go/no-go task that requires subjects to respond to frequently presented items from one category (~89% of the trials) and withhold responses to rare targets (~11% of the trials)	High "go" trial frequency can lead to mindless, habitual, responding. Sustained attention is needed to overcome the mindless, and potentially erratic, responding and maintain consistency. Rare "no-go" trials also require sustained attention in order to prevent commission errors that might occur because of habitual responding.	Response inhibition; response strategies (i.e., speed-accuracy tradeoff); processing speed; impulsivity; knowledge of stimuli used in task (i.e., knowledge of animals and fruits)
Metronome Response Task (MRT; Laflamme et al., 2018; Seli et al., 2013)	A continuous performance task, in which visual or auditory stimuli are presented at a constant rate, that requires subjects to respond in synchrony with the presentation of the stimuli	Repetitive presentation of low arousing stimuli for extended durations can elicit inconsistent responding. Sustained attention is needed to maintain consistent responding and not mis-time responses to stimuli.	Familiarity and skill with rhythm or music; time estimation; counting strategies
Continuous Temporal Expectancy Task (CTET; O'Connell et al., 2009)	Subjects view a series of abstract images that are perceptually similar to each other; their goal is to respond to rare target stimuli that are presented for longer-than-usual durations (1000-1200 ms) among frequent non-targets that are presented briefly (600-800 ms). Attention-capturing stimulus onsets/offsets are non-diagnostic to target detection.	Sustained attention is needed to focus and notice small temporal discrepancies among perceptually similar and repetitive stimuli.	Visual and temporal discrimination ability
Gradual Onset Continuous Performance Task (gradCPT; Rosenberg et al., 2013)	A go/no-go continuous performance task that presents subjects with frequent non-target stimuli and infrequent targets (similar to the SART). However, in the gradCPT, the stimuli gradually fade into one another, eliminating stimulus onsets/offsets which can capture attention.	High "go" trial frequency can lead to mindless, habitual, responding. Sustained attention is needed to overcome the mindless, and potentially erratic, responding and maintain consistency. Rare "no-go" trials also require sustained attention in order to prevent commission errors that might occur because of habitual responding.	Visual discrimination ability; response inhibition; processing speed; impulsivity; response strategies (i.e., speed-accuracy tradeoff)

Dependent Measures: Reaction Time Variability & Performance Accuracy

Historically, intra-individual variability in RT was mostly argued to reflect measurement error in simple tasks (Fiske & Rice, 1955). Several studies have suggested, however, that RT variability reflects an important source of information about subjects' cognitive state. Specifically, trial-to-trial RT variability and the frequency or duration of especially long RTs reflect, at least in part, the consistency (or inconsistency) of one's sustained attention. That is, if a subject is effectively sustaining focused attention across trials in a task, then their RTs should be similar from trial to trial. Note that this critical aspect of sustained attention is not well captured by central-tendency performance measures like mean or median RT, which instead better reflect general processing speed. As reviewed in Empirical Paper 2 (Welhaf & Kane 2022a) there are multiple ways to assess RT variability, all of which reflect behavioral instances of sustained attention failures, at least in part.

The least complex way to assess fluctuations of attention is to simply count the number of times that subjects produce relatively long RTs. These instances of "blocks," (e.g., Bills, 1931, 1935) or "lapses," (e.g., Lim & Dinges, 2008) capture sustained attention failures because they indicate instances in which subjects are not optimally focused on the task. Likewise, fluctuations in sustained attention can be assessed across the whole task by calculating within-subject trial-to-trial variability in responses using measures like intra-individual standard deviation of RT (RTsd), coefficient of variation (CoV) in RT, or Rhythmic Response Times (RRTs). These different measures reflect the (in)consistency of responding on a trial-to-trial basis rather than simple counts of attention lapses as reflected by blocks or lapses.

More complex approaches to assessing fluctuations in RT across a task include binning subjects' individual RTs or using distributional models (e.g., the ex-Gaussian model) to fit the full distribution of subjects' RTs (as exemplified in Empirical Papers 1 and 2 [Welhaf et al., 2020b; Welhaf & Kane, 2022a]). These approaches produce values that quantify subjects' slowest performing trials against their whole distribution. In the binning procedure, subjects' RTs are rank-ordered from shortest to longest and grouped into quantiles (e.g., shortest/fastest 20% to longest/slowest 20%), with the slowest quantile(s) partially reflecting sustained attention failures. Using the ex-Gaussian modeling approach, the *tau* parameter (which categorizes the tail of the RT distribution, as the mean and SD of its exponential component) is most often used as an indicator of sustained attention failures (but see Yamashita et al., 2021 for an argument that *sigma*, the standard deviation of the Gaussian component, better reflects sustained attention).

As subjects become mindlessly disengaged from a task, they may experience different types of errors that may (at least partially) reflect sustained attention failures. For example, subjects may respond when no response is required (i.e., a commission error) or fail to respond to one or more trials when it is required (i.e., omissions errors or "flat spots", see Cheyne et al., 2009; Unsworth et al., 2021). These errors might reflect different types, or severities, of sustained attention failures than those captured by RT fluctuations. Thus, it may be useful to consider both RT and accuracy-based measures to capture a full range of behavioral sustained attention failures (Cheyne et al., 2009; Unsworth et al., 2021).

Limitations of Objective Indicators of Attention Consistency

Studies of sustained attention that only use objective performance indicators may be tapping into sustained attention failures but may also be capturing measurement error. That is, objective indicators of sustained attention can be influenced by nuisance variables that confound its measurement. For example, longer-than-normal RTs, or performance errors, can be caused by failures of working memory (e.g., momentarily forgetting the stimulus that was just presented) or by subjects looking away between trials which causes them to miss the initiation of a trial or the trial completely. As well, as displayed in Table 1, commonly used sustained attention tasks require other non-sustained attention process for successful performance (e.g., the SART also requires inhibitory control for successful performance; experience with music and rhythms can contribute to MRT performance). We therefore cannot rely only on this one type of indicator as a process-pure manifestation of sustained attention ability.

Subjective Indicators of Attention Consistency

Failures of sustained attention are not limited to errors or fluctuations in performance measures. Some sustained-attention failures may be more overt, and perhaps overt enough to be easily reported by subjects when asked. Subjective measures of sustained attention aim to capture off-task thoughts and everyday attention failures. Below I review different methods of assessing subjective indicators of attention consistency and their limitations when used in isolation (see Empirical Study 2 [Welhaf & Kane, 2022a], for a more detailed review).

Diary Methods

Early studies on sustained attention lapses in daily life required subjects to record instances of attentional failures in a diary (Norman, 1981; Reason, 1984, 1990; Reason & Mycielska, 1982). While daily diary methods have their strengths (e.g., capturing salient lapses and recording rich details of attention failures; Reason & Lucas, 1984; Unsworth et al., 2012; Unsworth & McMillan, 2017), they have serious limitations: They rely on both prospective memory (remembering to write down any failures that occur) and retrospective memory (remembering *what* failures occurred), as well as meta-awareness (i.e., being aware that a failure occurred). Thus, many attention failures may go unreported.

Experience Sampling and Thought Probes

A more direct subjective assessment of sustained attention, both in the lab and in everyday life, is through experience-sampling methods. The technique used most frequently in the mind-wandering literature is thought probing, which has been used in a variety of tasks and contexts (for reviews see Kane, Smeekens, et al., 2021; Smallwood & Schooler, 2015). Studies using the thought-probe method will repeatedly and unpredictably interrupt subjects during one or more tasks or activities and have them report on the contents of their thoughts in the moment immediately preceding the probe appearance as being on-task or off-task (TUT), thus minimizing memory and meta-awareness contributions to reports.

Ample evidence suggests that TUT reports captured by thought probes are reliable and valid. TUT rates correlate substantially across tasks and contexts, and in latent variable studies, TUT rates from multiple tasks can be modelled as a single latent variable suggesting a trait-like propensity to experience TUTs (e.g., Kane et al., 2016; Rummel et al., 2021; Unsworth & McMillan, 2014; Unsworth et al., 2021). That is, people who tend to experience TUTs in one task (or session of an experiment) do so in other tasks (or sessions). Further, TUT rates are also associated with poorer task performance and external indicators of cognitive ability (see the *Evidence* section below for further discussion).

Limitations of Subjective Indicators of Attention Consistency

TUT reports are not process-pure indicators of sustained attention failures. First, subjects may not be able, or willing, to accurately report on their thoughts. This metacognitive and introspective demand may yield erroneous or biased reporting (Hurlburt & Heavy, 2001; Nisbett & Wilson, 1977). For example, subjects may feel like they need to report being on-task since an experimenter is watching them perform an activity; or subjects may be biased by the framing of the thought probe question (e.g., framing the question as "mind wandering" vs. "being on task", see Weinstein et al., [2018]).

Second, the frequency of thought probes, themselves, may change how TUTs are reported, if not experienced (see Welhaf et al., 2021 and Empirical Paper 3 [Welhaf & Kane, 2022b] for a discussion of the benefits and drawbacks of frequent probing). In general, more frequent probing can lead to lower TUT rates (Seli, Carriere, et al., 2013; Schubert et al., 2019; but see Robison et al., 2019), perhaps because they (re)orient attention back to the task. Thus, while probes are a useful way to access subjects conscious experience, this method may fundamentally alter how subjects' thoughts unwind during a task.

Third, thought probes might be biased by reactivity to performance. For example, in the SART (and other go/no-go tasks), subjects are often aware of the errors they make, especially on no-go trials. Previous work has found that TUT reports are more frequent following no-go errors compared to correctly withheld no-go trials (e.g., Kane, Smeekens et al., 2021; Schubert et al., 2019). This indicates that subjects may sometimes rely on their immediate performance to indicate where their thoughts might have been.

Sustained Attention Measurement Summary

Objective and subjective indicators provide different approaches to measuring attention (in)consistency. Because both have their limitations and independent sources of error, the combination of these two should best reflect the construct of sustained attention, independent of those sources of error. As such, I argue that the optimal way of capturing people's general sustained attention (in)ability, is to use variance that is common to both objective and subjective indicators.

Evidence for the Construct Validity of Attention Consistency Measures

Considerable research has examined relationships that sustained-attention indicators have with other theoretically relevant variables (the nomothetic span approach) and how these indicators change under different experimental conditions (the construct representation approach). Empirical Studies 2 and 3 (Welhaf & Kane, 2022a, 2022b) explain these findings in detail and so I review them only briefly below.

Nomothetic Span (Correlational) Studies

If attention consistency is best reflected by the covariation in objective and subjective indicators, then these two kinds of indicators should correlate moderately with each other. Findings at the between- and within-subject level indicate that there is a consistent association between objective and subjective indicators of sustained attention. Between-subject analyses (e.g., latent variable correlations) indicate moderate correlations between TUT rates and RT variability factors of .30–.40 (Kane et al., 2016; Unsworth, 2015; Unsworth et al., 2021; Welhaf et al., 2020b), suggesting these factors share some variance without being redundant: Subjects who report more off-task thoughts also show more inconsistent responding in simple attention and RT tasks.

Within-subjects analyses present parallel findings. Subjects are more likely to make errors and produce more variable RTs on the trials immediately preceding TUT reports compared to on-task reports (e.g., Bastian & Sackur, 2013; Kane, Smeekens et al., 2021; McVay & Kane, 2009; Schubert et al., 2019; Stawarczyk et al., 2011). RT variability on the trials leading up to probes may be the less biased way to examine such within-subject covariation between these measures because TUTs following errors might be a result of performance bias rather than actual sustained attention failures. That is, subjects are likely unaware of their consistency in responding (measured in fractions of a second) and how it relates to their performance, but they often can tell when they have made an error in a task (and may even make audible "oops" reactions following errors). In general, though, these RT and accuracy findings present initial construct validity evidence that these measures both reflect failures of sustained attention.

Given that performance and self-report indicators are subject to different non-sustained attention confounds, however, it is unsurprising that they are only moderately correlated. The lack of a strong correlation between objective and subjective measures of sustained attention is important because it suggests that these two indicators are not isomorphic ways of measuring sustained attention. Using either approach on its own may lead to incorrect conclusions about the ability to sustain attention. That is, studies can't simply swap out objective measures with subjective measures (or vice versa) and still be confident that they are measuring the ability to sustain attention. It may therefore be important, if not necessary, to measure sustained attention as the individual-differences covariation between these measurement types to make appropriate claims regarding sustained attention.

If there are stable individual differences in the ability to sustain attention, then a next step is to figure out who is especially susceptible to its failures. Previous research has identified multiple cognitive, contextual-state, and dispositional variables that can help explore this question. In terms of cognitive factors, RT variability is consistently related to working memory capacity (WMC), attention control (interference control), and processing speed, such that higher-ability subjects also show more consistent (i.e., less variable) performance (e.g., McVay & Kane, 2012a; Kane et al., 2016; Schmiedek et al., 2007; Unsworth et al., 2021).[2] As for TUT rates,

[2] Note that general processing speed, often represented by *M* RT, and RT variability are mathematically confounded. Slower RTs not only contribute to slower overall processing speed

again, WMC and attention control appear to be consistent correlates (e.g., McVay & Kane, 2012b; Kane et al., 2016; Unsworth & McMillan, 2014), but correlations with processing speed are less consistent (e.g., Unsworth et al., [2021] found a significant correlation, whereas Welhaf et al., [2020b] Empirical Study 1, did not). Collectively then, WMC and attention control abilities should be related to general sustained attention, but processing speed may or may not be. To better understand how these constructs correlate with sustained attention, one should model the individual-differences covariation between objective and subjective indicators as this is less influenced by processes unique to either indicator.

Contextual-state variables also appear to be related to both objective and subjective measures of attention consistency. The most frequently examined variables are self-reported motivation, alertness, and interest: People who report being more motivated, alert, or interested in a task show lower RT variability and report fewer TUTs (e.g., Hollis & Was, 2016; Kawagoe, 2022; Smith et al., 2022; Soemer & Schiefele, 2019; Unsworth et al., 2021). It is worth noting that the correlations between *subjective* measures of sustained attention with these contextual-state variables is often stronger than those for objective measures; this may not be surprising given that the contextual-state measures also rely on self-report, and so may partially reflect measurement-related variance. Thus, seeing how these contextual-state variables correlate with the shared variance between objective and subjective measures may give the field a better estimate of the actual strength of such correlations with sustained attention ability.

Finally, dispositional factors, like some personality traits, may also correlate with general sustained attention ability. On one hand, people who have higher levels of neuroticism typically

by increasing *M* RT, but they also increase the spread of the RT distribution which increases RT variability.

show increased RT variability and TUT rates and so may have poorer sustained attention ability, in general (Klein and Robinson, 2019; Robinson & Tamir, 2005; Robison et al., 2017; Unsworth et al., 2021). On the other hand, some personality traits like extraversion, conscientiousness, agreeableness, and openness, show inconsistent correlations with TUT rates and almost no correlation with objective sustained attention indicators. Only certain aspects of personality (e.g., neuroticism), then, may be related to general sustained attention ability.

A limitation that applies to all the previously discussed nomothetic span studies is that the correlations in question have looked at objective and subjective measures of sustained attention as separate outcomes. My research program has tested whether a more appropriate approach would be to see how these measures correlate with the shared variance between objective and subjective measures, as this measure would be less influenced by nuisance variables specific to either objective or subjective measures.

Construct Representation (Experimental) Studies

Below I briefly review relevant research that shows how using theoretically derived manipulations alters RT variability or TUT rates in ways that theory predicts. Empirical Study 3 (Welhaf & Kane, 2022b) provides a more detailed review of these studies and the implications of such experimental manipulations on sustained attention measurement.

One approach to altering the sustained attention demands of a task is to manipulate specific task parameters that might be critical, or responsive, to sustained attention, such as changing the pacing or expectancy of trials, or changing the response frequency during a task. Regarding task pacing, performance on faster-paced tasks (i.e., those with shorter, or more predictable or constant, interstimulus intervals) tend to have lower RT variability compared to slower-paced, or less predictable, tasks (e.g., Langner & Eickhoff, 2013; Unsworth et al., 2018).

The effects of these manipulations on TUT rates are less compelling. Faster paced tasks yield lower TUT rates than slower tasks (e.g., Antrobus, 1968; Giambra, 1995; Unsworth & Robison, 2020), but comparisons of different trial expectancies appear to have no effect on TUT rates (Hawkins et al., 2019; Massar et al., 2020).

Manipulating response or trial-type frequency appears to affect both objective and subjective indicators of attention consistency. When tasks require frequent repetitive responding, like in the SART, participants can build up a habitual, "mindless" response pattern. To minimize the sustained attention demands of such tasks, studies can reduce the response frequency, which gives subjects less opportunity to engage in extended periods of mindless, repetitive, responding. In go/no-go tasks, for example, increasing the proportion of no-go trials results in faster go RTs and increased no-go accuracy (Nieuwenhuis et al., 2003; Young et al., 2018). Note, however, that such changes in performance can also be attributed to changes in response strategy (i.e., speed-accuracy trade-offs) rather than sustained attention (e.g., Head & Helton, 2014; Mensen et al., 2022; Wilson et al., 2016). Of the few studies that have investigated the effect of trial-type manipulations on TUT rates (e.g., Giambra, 1995; Smallwood et al., 2007), the findings appear to support the idea that giving participants less opportunity to engage in prolonged sequences of repetitive work helps them better focus on the task at hand.

Another way to alter the sustained attention demands of a task is to provide monetary or performance incentives. Such incentives may be enough to keep subjects engaged in the current task and thus improve sustained attention indicators. This appears to be the case: Compared to control conditions, participants who are rewarded for their time and performance tend to show less RT variability and better accuracy, as well as lower TUT rates (e.g., Robison, et al., 2021; Seli et al., 2019; Smallwood et al., 2007).

This previously discussed experimental work provides some support for the construct validity of objective and subjective indicators of attention consistency. That is, theoretically derived manipulations affect mean levels of objective or subjective measures (or both), as predicted. An obvious limitation of these studies, though, is that these manipulations have only targeted changes in the mean levels of RT variability and/or TUT rates. If these manipulations are tied to sustained attention, then they should also reduce, or eliminate, the correlation between objective and subjective indicators, which I argue is a more construct valid way to measure sustained attention. Objective and subjective measures should be most strongly correlated in tasks with high sustained attention demands, as both variability in both measures is primarily caused by sustained attention processes. In contrast, when tasks place lower demands on sustained attention, the correlation between objective and subjective indicators should weaken because their variation is now primarily caused by non-sustained attention processes that are unique to each indicator type.

Aims

The goal of my research program is to better understand sustained attention and to improve its measurement. More specifically, this line of research builds on existing literature that has exclusively looked at objective and subjective indicators of sustained attention as separate, but correlated, constructs. Because each of these measurement types is influenced by different non-sustained-attention processes, relying solely on one of these indicator types as a primary measure of sustained attention may lead to erroneous claims. Rather, using the individual-differences covariation in these measures should allow for more accurate measurement of the ability to sustain attention, as this measure is not influenced by non-sustained-attention factors unique to either objective or subjective measures. Further, this individual-differences covariation should be correlated with theoretically relevant constructs and more sensitive to theoretically motivated experimental manipulations of demand than should either objective or subjective measures alone. By using the strengths of both the nomothetic span and construct representation approaches to construct validation, my research program can inform the field on the most appropriate way to assess sustained attention ability.

Empirical Paper 1 (Welhaf et al., 2020b)

Empirical Paper 1 (Welhaf et al., 2020b) investigated the robustness of the worst performance rule, and in doing so, explored the associations between performance and self-report measures of attention consistency. As noted earlier, the worst performance rule is the empirical finding that the correlation between subjects' ability level and their RTs increases from the shortest RTs (fastest/best responses) to the longest RTs (slowest/worst responses). In other words, higher- and lower- ability subjects don't necessarily differ in their fastest, or even average, performing trials, but lower-ability subjects are much slower on their slowest trials compared to their higher-ability counterparts. One theoretical account of the worst performance rule suggests that worst performing trials (i.e., ones with the longest RTs) occur, in part, because of lapses of attention (e.g., Larson & Alderton, 1990; Unsworth et al., 2010). On these trials, lower-ability subjects are more likely to be momentarily distracted, missing the onset of the trial and only regaining focus at the very end of the trial, producing an accurate, but slowed response.

A recent meta-analysis (Schubert, 2019) argued, instead, that the worst performance rule should be renamed the "not-best-performance-rule," as correlations between cognitive ability and RTs increased from fastest, or "best," RTs, to average RTs, but then remained stable to the worst RTs. That is, subjects' average and worst performance were both equally telling of one's ability level. We tested this claim by reanalyzing a previously published dataset (Kane et al.,

2016) and assessing the latent variable associations between WMC, TUT rate, and two different approaches to fitting RT data (RT binning and ex-Gaussian models).

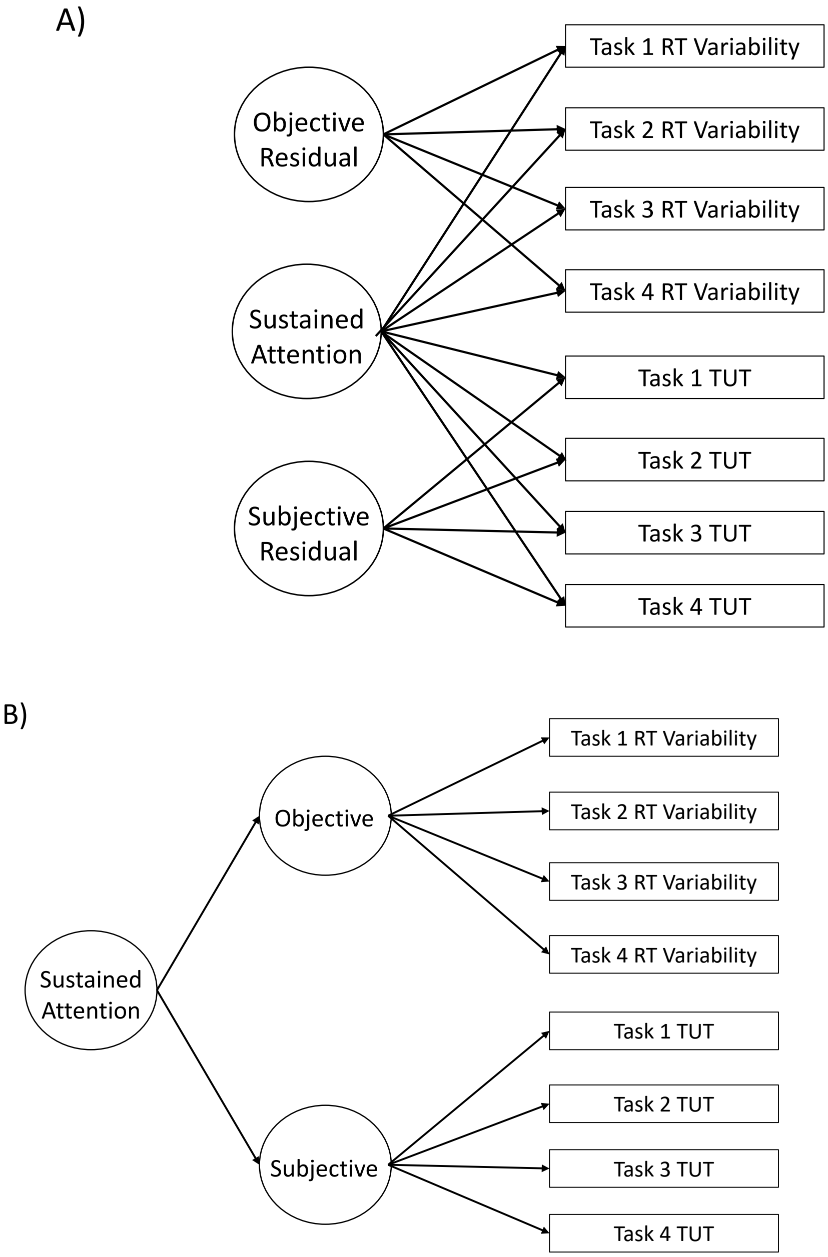
We found that the pattern of results described by both the worst performance rule and the "not-best-performance-rule" appeared, but it depended on the ability construct. Specifically, when using working memory capacity (WMC) as our ability measure, we found patterns of results consistent with Schubert's (2019) claims: Correlations between WMC and RT were weakest with subjects "best" performance and stronger with their "average" and "worst" performance, but these latter two correlations did not differ. However, when TUT rate was the (sustained attention) ability measure, we found traditional worst performance rule patterns: Correlations between TUT rates and RTs increased substantially across "best" to "worst" performance. These findings suggest a connection between subjects' slowest trials and their self-reported mind wandering; because both measures reflect, in part, the ability to sustain attention during a task, a common underlying ability may explain variation in both behaviors. However, because there are clear methodological differences (i.e., objective task performance vs. subjective self-reports), it may be necessary to look at what is *common* between these measures to best capture the ability to sustain attention as an individual-differences construct.

Empirical Paper 2 (Welhaf & Kane, 2022a)

Empirical Paper 2 (Welhaf & Kane, 2022a) took a nomothetic span approach to assessing the construct validity of sustained attention measures, and specifically the individual-differences covariation in objective and subjective indicators. We reanalyzed data from two large-*N* latent variable studies (Kane et al., 2016; Unsworth et al., 2021) that had multiple tasks from which we could derive different objective attention consistency indicators. Thought probes also appeared in multiple tasks in each study as subjective indicators.

We modeled the general ability to sustained attention in two ways: as a bifactor model and as a hierarchical model (see Figure 1 for a generic depiction of these models). In the bifactor model (panel A), we attempted to simultaneously model the variance common to all objective and subjective indicators (i.e., a common sustained attention factor) and the variance unique to the objective indicators and unique to the subjective indicators, while accounting for the general factor (i.e., residual objective-specific and subjective-specific factors). In the hierarchical model (panel B), the general factor was a second-order latent variable that was modeled as the shared variance between the first-order objective and subjective latent variables.

Figure 1. Proposed sustained attention factor structures



Note. Panel A depicts the bifactor model; Panel B depicts the hierarhcial model.

We found that the covariation in objective and subjective measures could be modeled using both bifactor and hierarchical approaches, indicating that there was an underlying general ability to sustain attention that explained variance in these measures. This general factor correlated weakly to moderately with theoretically relevant constructs like cognitive ability (e.g., WMC, attention control, and processing speed), contextual-state factors (self-reported motivation and alertness), and dispositional factors (self-reported cognitive failures and personality traits). Critically, the strengths of the associations between the nomological network constructs and the general sustained attention factor were as strong, if not stronger, than those with either the objective or subjective factors, providing convergent-validity evidence for the general factor. We also found evidence for discriminant validity. Specifically, some measures (e.g., conscientiousness and agreeableness) correlated with the subjective factor, but not the general factor. Thus, these measures might not be associated with general sustained attention but rather processes specific to self-reports (e.g., self-reporting biases).

Taken together, these findings suggest that previous research that has relied on only one type of sustained attention indicator may have under- or over-estimated correlations with nomological network constructs. Across both re-analyzed datasets, the hierarchical model, compared to the bifactor model, appeared to be more robust to different subject- and trial-level outlier treatments (see below), and a full bifactor model did not fit the Study 1 data well. We therefore argued that the hierarchical approach may be a more construct valid way to assess the general ability to sustained attention than either the bifactor approach or the separate objective or subjective factors.

Empirical Papers 1 and 2 also make a methodological contribution beyond understanding and improving sustained attention measurement. In both papers, many of the primary dependent measures were RT based. While relatively long RTs are often due to sustained attention failures, some may be caused by random behaviors of the subjects (e.g., sneezing, blinking, looking around the experiment room). How do we handle such outliers? Many papers don't report how trial-level outliers were treated, and among those that do, there are many different approaches. These different decisions can yield different results. To increase the transparency and test the robustness of our empirical claims, both Empirical Papers 1 and 2 used a "mini-multiverse" approach. Here, we examined how our main findings for each study changed as different, commonly employed subject- and trial-level outlier decisions were implemented on the raw data. As we note in both papers, the main results largely replicated across each strand of the multiverse, suggesting our findings were robust to different data analysis pipeline decisions. However, in cases where the findings did not replicate well across different variants, we use this as evidence that the approach or model may not be appropriate.

Empirical Paper 3 (Welhaf & Kane, 2022b)

Empirical Paper 3 (Welhaf & Kane, 2022b) took a construct representation approach, combining experimental with correlational methods to assess the construct validity of sustained attention measures. In two large-N studies conducted online, we assessed how theoretically derived experimental manipulations of sustained attention demands affected mean levels of RT variability and TUT rates in prototypical sustained attention tasks, and most critically, their correlation. Specifically, we asked whether implementing manipulations that should theoretically reduce the sustained attention demands of a task would result in lower RT variability and TUT rates (i.e., traditional experimental effects), and, critically, weaker correlations between these two indicators, compared to a task that placed a higher demand on sustained attention. If the individual-differences overlap in RT variability and TUT rates is a construct valid measure of sustained attention, then minimizing the demands of sustained attention should weaken this correlation because variation in each indicator will primarily be driven by other non-sustained attention processes unique to either RT variability or TUT rates.

The results of both studies indicated that our manipulations had a significant impact on mean levels of RT variability and TUT rate: In tasks that minimized the demands on sustained attention, both indicators of sustained attention failures were lower than in tasks that placed a high demand on sustained attention. However, contrary to predictions, these manipulations did *not* affect the individual-differences overlap in these measures: RT variability and TUT rate were significantly and similarly correlated with each other in the demand-maximized and demand-minimized tasks. Thus, from a construct representation approach, we found only some support for the construct validity of these measures. Specifically, our manipulations effectively reduced both mean levels of RT variability and TUT rates in both studies, supporting the idea that these manipulations are tied to sustained attention. However, in neither study did the correlation between RT variability and TUT rates (which, we argue, is a more valid measure of sustained attention) change because of these manipulations.

We reflected on these mixed results in multiple ways. First, we discussed that we could have been wrong about our sustained attention measurement approach in using the covariation between objective and subjective measures. We argued against this, however, noting the supportive findings from our nomothetic span study (Empirical paper 2; Welhaf & Kane, 2022a) and the clear experimental effects we found on both RT variability and TUT rates. Second, we suggest that despite our manipulations working to some degree, they may not have been strong enough to reduce the correlation between the indicators. Sustained attention is likely so fundamental to nearly any task that it may be extremely difficult, if not impossible, to reduce the shared variance between objective and subjective sustained attention measures enough to see any measurable between-person differences. We suggested that future studies should consider additional task-demand manipulations (e.g., probe frequency, motivation manipulations, more frequent and perhaps longer breaks) and other methodological considerations (e.g., testing in a controlled lab setting vs. online to reduce participant distraction) in experimentally testing our claim that the covariation between objective and subjective sustained attention measures is a more construct valid way to measure the ability to sustain attention.

CHAPTER II: WORST PERFORMANCE RULE, OR NOT-BEST PERFORMANCE RULE? LATENT-VARIABLE ANALYSES OF WORKING MEMORY CAPACITY, MIND-WANDERING PROPENSITY, AND REACTION TIME

Welhaf, M.S., Smeekens, B.A. Meier, M.E., Silvia, P.J., Kwapil, T.R., & Kane, M.J. (2020). The worst performance rule, or the not-best performance rule? Latent-variable analyses of working memory capacity, mind-wandering propensity, and reaction time. *Journal of Intelligence*, *8*, 25. https://www.mdpi.com/2079-3200/8/2/25

Abstract

The worst performance rule (WPR) is a robust empirical finding reflecting that people's worst task performance shows numerically stronger correlations with cognitive ability than their average or best performance. However, recent meta-analytic work has proposed this be renamed the "not-best performance" rule because mean and worst performance seem to predict cognitive ability to similar degrees, with both predicting ability better than best performance. We re-analyzed data from a previously published latent-variable study to test for worst vs. not-best performance across a variety of reaction time tasks in relation to two cognitive ability constructs: working memory capacity (WMC) and propensity for task-unrelated thought (TUT). Using two methods of assessing worst performance—ranked-binning and ex-Gaussian-modeling approaches—we found evidence for both worst and not-best performance rules. WMC followed the not-best performance rule (correlating equivalently with mean and longest RTs) but TUT propensity followed the worst performance rule (correlating more strongly with longest RTs). Additionally, we created a mini-multiverse following different outlier exclusion rules to test the robustness of our findings; our findings remained stable across the different multiverse iterations.

We provisionally conclude that the worst performance rule may only arise in relation to cognitive abilities closely linked to (failures of) sustained attention.

Introduction

Adults who score higher on intelligence tests also ted to respond faster in simple and choice response time (RT) tasks (Doebler & Scheffler 2016; Jensen 1992; Sheppard & Vernon 2008). However, different parts of the RT distribution are more predictive of cognitive ability: The *worst performance rule* (WPR; Coyle 2003a; Larson & Alderton 1990) describes the empirical finding that subjects' longest RTs (e.g., the slowest 20% of responses) correlate more strongly with cognitive ability than do their shortest or their average RTs. The WPR appears in a variety of RT tasks (Baumeister & Kellas 1968; Jensen 1982, 1987) and across the lifespan (Coyle 2001, 2003b; Fernandez et al. 2014).

A recent meta-analysis (Schubert 2019) indicated that the WPR is robust: Correlations between people's shortest RTs and intelligence ($r = -0.18$, [95% CI −0.27, −0.08]) were numerically weaker than those between their mean RT and intelligence ($r = -0.28$ [95% CI −0.38, −0.18]) and these were numerically weaker than between their longest RTs and intelligence ($r = -0.33$, [95% CI −0.41, −0.24]). Schubert noted, however, that the meta-analytic results suggested a logarithmic rather than linear association between measures of intelligence and RT. That is, the change between correlations was greatest between shortest and mean RTs, while the change from mean to longest RTs was small. Individual differences in shortest RTs were less strongly associated with ability than were *both* mean and longest RTs. Schubert thus suggested that the WPR be renamed the "not-best performance" rule.

Although the WPR is most often studied in relation to intelligence, related constructs show similar trends. Indeed, the WPR is sometimes explained as reflecting fluctuations of working memory (Larson & Alderton 1990; Larson & Saccuzzo 1989) or of focused attention to the task (Jensen 1992). Failing to maintain attention during a task may result in especially long RTs on those occasional trials where attention is focused elsewhere. People with lower working memory capacity (WMC) and lower intelligence are more prone to attentional lapses (Engle & Kane 2004; Kane & McVay 2012), and WMC appears to be especially related to subjects' slowest responses (McVay & Kane 2012a; Schmiedek et al. 2007; Unsworth et al. 2010; Unsworth et al. 2012; Unsworth et al. 2011; Wiemers & Redick 2018). The attention-control account of the WPR (Larson & Alderton 1990; Unsworth et al. 2010) thus proposes that people of lower ability are more susceptible to attentional lapses that disrupt goal maintenance in working memory than are those of higher ability.

On one hand, the attention-lapse account of the WPR is consistent with a prominent theory of intelligence, Process Overlap Theory (POT), which proposes that cognitive-task performance requires the contribution of many domain-specific processes and domain-general executive processes (Kovacs & Conway 2016). Central to POT is that, within a cognitive domain the overlapping processes may compensate for one another, but between domains they cannot; domain-general executive processes may thus act as a bottleneck for item solution when executive demands exceed executive ability. According to POT, then, the WPR arises partly because people with lower WMC/intelligence do not have the ability to meet the necessary executive demands of blocking distractions or sustaining focus on every trial, even though domain-specific processes may be up to the task. These occasions result in extremely slow responses that produce the WPR. On the other hand, POT does not require that the WPR better characterizes performance than does the not-best performance rule. Insofar as other executive processes also contribute to task performance, and these other executive processes tend to fail more frequently than rare attentional lapses (or fail with different thresholds), POT can accommodate either the WPR or not-best performance rule pattern. Indeed, POT might also predict that ability measures that best capture the propensity for occasional sustained attention failures should show a WPR pattern whereas ability measures that best capture other executive abilities might show a not-best performance rule pattern.

Two approaches have been used most frequently to quantify worst performance. The most common is the ranked-binning procedures, where subjects' individual RTs are ranked from shortest to longest and split into quantiles (e.g., 5 bins, from the shortest 20% of RTs to the longest 20%). A second approach models the shape of each subject's RT distribution. The ex-Gaussian model, for example, represents a subject's RT distribution—which is typically positively skewed—as a convolution of a Gaussian (normal) and exponential distribution, with three parameters: *mu*, *sigma*, and *tau*[3]. Mu and sigma reflect the mean and standard deviation of the Gaussian distribution, respectively, whereas tau represents the mean and standard deviation of the exponential component (i.e., the tail of the positively skewed distribution). The parameters of the ex-Gaussian model do not reflect isolated cognitive processes (Matzke & Wagenmakers 2009), but because the tau parameter frequently correlates with normal individual differences in WMC more strongly than do the other parameters, some have proposed that *tau* may sometimes reflect failure of goal-maintenance in the form of occasional attentional lapses (McVay & Kane 2012a; Unsworth et al. 2010, 2011, 2012).

[3] The ex-Gaussian model is but of many that adequately fit RT distributions including the Wald, Gamma, Weibull, and Lognormal functions (Heathcote, Brown, & Cousineau 2004; Matzke & Wagenmakers 2009; Ulrich & Miller 1993; Van Zandt 2000).

If failures of attentional focus can explain the WPR, at least in some task contexts, then assessing subjects' thought content during a task should also produce patterns consistent with the WPR. During laboratory tasks, as well as in everyday activities, peoples' thoughts sometimes drift from what they are doing to something unrelated, resulting in the phenomenon of "daydreaming," "mind wandering," or "task-unrelated thoughts" (TUTs; e.g., Christoff & Fox 2018; McVay & Kane 2010, Randall, Oswald, & Beier 2014; Smallwood & Schooler 2015). TUTs are typically assessed via experience sampling, where subjects are interrupted at unpredictable times during a task or activity and asked to report on their immediately preceding thoughts.

These probed TUT rates have been validated as predicting performance at both within-subject and between-subject levels. At the within-subject level, TUT reports are more frequent following task errors than correct responses (McVay & Kane 2009; Smallwood & Schooler 2006; Stawarczyk et al. 2011) and following relatively fast or variable runs of RTs (Bastian & Sackur 2013; McVay & Kane 2009, 2012a; Seli et al. 2013); TUT reports also vary with assessments of pupil size, an indirect and unobtrusive indicator of arousal and sustained attention (e.g., Unsworth & Robison 2016, 2018; Unsworth et al. 2018), and with particular neuroimaging signatures (e.g., Arnau et al. 2020; Baldwin et al. 2019; Christoff et al. 2009; Kam & Handy, 2013) At the between-subjects level, evidence indicates that TUTs reflect, in part, executive abilities to sustain attention. For example, individual differences in probed TUT rate are reliable across tasks and occasions, indicating a trait-like propensity for off-task thought during challenging activities (e.g., Kane et al. 2016; McVay & Kane 2012b; Robison & Unsworth 2018). Moreover, individuals who frequently report TUTs show worse performance (in accuracy, RT variability, or both) on a range of cognitive tasks including reading comprehension (McVay

& Kane 2012b; Schooler et al. 2004), working memory (Banks et al. 2016; Kane et al. 2007; Mason et al 2007; Mrazek et al. 2012; Unsworth & Robison 2015) and attention-control tasks (McVay & Kane 2012a, 2012b; Cheyne et al. 2009; Kane et al. 2016; McVay & Kane 2009, 2012a; Robison et al. 2017). Individual differences in TUT rate and attention-task performance also covary with those in pupil-size variability in cognitive tasks (e.g., Unsworth & Robison, 2017, 2018). These findings, together, indicate that, although it is a self-report measure, TUT rate reflects (at least in part) an ability to sustain attention during challenging tasks.

Several studies have shown that TUT rates correlate with intrasubject variability in RT (i.e., RT standard deviations or coefficients of variation; Bastian & Sackur 2013; McVay & Kane 2009, 2012a; Seli, Cheyne, & Smilek 2013; Unsworth 2015) but only one study has related TUT rates to characteristics of the RT distribution that might be indicative of the WPR. McVay and Kane (2012a) found modest correlations between TUT rates and ranked-bin RTs in a long-duration go/no-go task: Subjects with higher TUT rates had shorter RTs in the fastest bins and longer RTs in the slowest bin. From the ranked-bin approach, then, it is unclear whether TUT-variation follows a pure WPR pattern (go/no-go tasks may be unique in eliciting very fast but "mindless" go responses in addition to very slow ones). McVay and Kane also assessed the association between TUT rates and ex-Gaussian parameters, which provided evidence for the WPR: TUT rate was weakly associated with *mu* (r = -.18) and not related to *sigma* (r = -.07), but moderately associated with *tau* (r = .30); subjects who reported more mind wandering during the task also had more especially long RTs that were captured by the *tau* parameter.

The Present Study

The primary aim of the current study was to apply the meta-analytic findings of Schubert (2019) to a novel dataset, with a relatively large subject sample, across a variety of attention-control tasks, and in relation to two individual-differences constructs—WMC and TUT rate. While the meta-analysis conducted by Schubert (2019) coherently characterized existing "WPR" data, we assessed here whether it would similarly extend to a new, large dataset. Thus, we asked whether there is evidence for the traditional WPR or the "not-best" performance rule pattern (Schubert 2019)—or, perhaps, both, depending on the predictor construct. To do so, we reanalyzed data from a previously published latent-variable study (Kane et al. 2016), focusing on a subset of tasks where RT was a primary dependent measure (using only the non-conflict trials from those response-conflict tasks, in order to make closer contact with the WPR literature). We calculated both ranked-bins and ex-Gaussian parameters and assessed their associations with WMC and TUT rates, both at the individual-task level and at the latent-variable level.

As a secondary aim, we also examined the robustness of our findings to various treatments of outlier trials and outlier subjects via a "mini-multiverse" analysis (Silberzahn et al 2018; Steegen et al 2017). One of the main methodological considerations of the WPR, as discussed by Coyle (2003a), is the role of outliers. Given that outliers populate the slowest bins and affect the tau parameter, their inclusion or exclusion might substantially alter measurement of worst performance, and yet Schubert's (2019) meta-analysis found little consistency in outlier treatment. Here, then, we created different datasets based on different trial-level and subject-level outlier criteria based on commonly reported methods in the studies included in Schubert; we refer to this as a mini-multiverse because we explored a substantial number of reasonable combinations of prototypical outlier treatments without exploring the full universe of all possible treatments and their combinations (which, in terms of RT outlier criteria, are infinite).

Methods and Materials

Subjects

Kane et al. (2016) enrolled 545 undergraduates into their study from the University of North Carolina at Greensboro, a comprehensive state university (and Minority-Serving Institution for African-American students). Of these, 541 completed the first of three 2 hr sessions, 492 completed the second, and 472 completed all three. Full-information maximum likelihood (ML) estimation was used for missing data (see Kane et al. for details and demographics). By comparison, the average sample size of WPR studies included in Schubert (2019) meta-analysis was 164 (SD = 182), with only one included study testing more than 400 subjects (Dutilh et al. 2017).

Reaction Time (Outcome) Tasks

We focused our analyses on tasks where RT was the primary dependent measure from Kane et al. (2016): The Sustained Attention to Response Task (SART), Number Stroop, Spatial Stroop, Arrow Flanker, Letter Flanker, and Circle Flanker tasks. Below we briefly describe each task and how their RTs were derived; for analyses reported here, we used only the non-conflict trials from each task.

SART. In this go/no-go task, subjects pressed the space bar for words from one category (*animals*; 89% of trials) but withheld responding to another (*vegetables*; 11% of trials). Subjects completed 675 analyzed trials. RTs were taken from correct responses to “go” (animal) trials.

Number Stroop. Subjects reported the number of digits presented on each trial while ignoring the digits’ identity. Each trial presented 2 to 4 identical digits in a row and subjects responded with one of three labeled keys to indicate the number of digits on screen. There were

300 total trials, of which 80% were congruent (e.g., *4444*) and remaining 20% were incongruent (e.g., *2222*). Here, we took RTs from correct responses to congruent trials.

Spatial Stroop. Subjects reported the direction of a centrally presented arrow ("<" vs. ">") via keypress, with the arrow flanked horizontally by 4 distractors. Subjects completed two blocks of 96 trials: 24 neutral trials (target arrow presented amid dots), 24 congruent trials (all arrows pointing the same direction), 24 stimulus-response incongruent trials (central arrow pointing opposite direction of flankers), and 24 stimulus-stimulus incongruent trials (central arrow presented amid upward pointing arrows). Here, we used RTs from correct responses to both neutral and congruent trials.

Arrow Flanker. Letter Flanker Subjects reported whether a centrally presented "F" appeared normally or backwards via keypress, with that letter flanker horizontally by 6 distractors. Subjects completed 144 trials: 24 neutral trials (normal or backwards F presented amid dots), 48 congruent trials (target and distractor Fs all facing the same direction), 24 stimulus-response incongruent trials (target facing opposite direction of distractors), and 24 stimulus-stimulus incongruent trials (target presented amid right- and left- facing Es and Ts tilted at 90 and 270 degrees). Here, RTs were derived from correct responses to neutral and congruent trials.

Circle Flanker. Subjects reported whether a target letter was an X or N, via keypress, with the target flanked by two distractors. Targets appeared in one of eight possible locations in a circle, with distractors appearing to position one either side of the target; all other location were occupied by colons. Subjects completed 160 trials: 80 neutral trials (target letter surrounded by colons) and 80 stimulus-stimulus conflict trials (target flanked by two different distractors from the set H, K, M, V, Y, Z). Here we took RTs from correct responses to neutral trials.

Cognitive Predictor Measures

For a detailed description of the tasks used for the present analyses (as well as non-analyzed tasks and task order), see Kane et al. (2016). Here we used only two of their cognitive constructs as predictors in our statistical models—WMC and TUT rate (i.e., we did not analyze performance from attention-constraint or attention-restraint tasks here, other than the neutral and congruent RTs described from the tasks above as outcome measures).

Working Memory Capacity (WMC). In six tasks, subjects briefly maintained items in memory while engaging in secondary tasks or mental updating. Four complex span tasks presented sequences of verbal or visuospatial items that required immediate serial recall (Operation Span, Reading Span, Symmetry Span, Rotation Span); memory items were preceded by unrelated processing tasks requiring yes/no responses. Two memory-updating tasks (Running Span, Updating Counters) required subjects to maintain an evolving set of stimuli in serial order while disregarding previous stimuli. Higher scores indicated more accurate recall.

Thought Reports of TUT. Thought probes appeared randomly within 5 tasks (45 in SART, 20 in Number Stroop, 20 in Arrow Flanker, 12 in Letter Flanker, and 12 in an otherwise-unanalyzed 2-back task). At each probe, subjects chose among eight presented options that most closely matched the content of their immediately preceding thoughts. TUTs were comprised of response options 3-8 in Kane et al. (2016): “Everyday Things” (thoughts about normal life concerns, goals, and activities); “Current State of Being” (thoughts about one’s physical, cognitive, or emotional states); “Personal Worries” (thoughts about current worries); “Daydreams” (fantastical, unrealistic thoughts); “External Environment” (thoughts about things or events in the immediate environment); “Other.”

RT Data Cleaning Procedure

All data were cleaned and aggregated in R (R Core Team, 2017) using the *dplyr* package (Wickham, Francois, Henry, & Muller, 2018). Data from all RT tasks were cleaned in the same manner for primary analyses. We first identified and removed error and post-error trials (and, in tasks that included thought probes, post-probe trials). In tasks that included conflict trials, we removed all conflict trials to focus our analyses on non-conflict trials to remove potential interference effects. From the remaining trials, we eliminated likely anticipatory trials (i.e., faster than 200 ms). For all primary regression and latent variable models, we next identified trial outliers that were outside 3 times the interquartile range (3*IQR) of each individual subjects' mean RT for each task and replaced those trials with values equal to 3*IQR. This procedure affected <2% of trials in each task. Following all trial-level treatments and aggregation, RT variables were z-scored at the sample level. As we will discuss later, a mini-multiverse analyses repeated our primary latent variable analyses across various combinations of trial- and subject-level outlier decisions (see Mini-Multiverse Results).

Results

Data used for all analyses, as well as analysis scripts and output, are available via the Open Science Framework (https://osf.io/9qcmx/). For detailed description of data-analysis exclusions, scoring of predictor tasks, and treatment of outliers in predictor tasks, please see Kane et al. (2016). We modeled the cognitive predictor constructs (WMC and TUTs) identically to Kane et al., including any residual correlations among indicators.

In the following sections, we first report results from the ranked-bin approach. Regression analyses provide descriptive evidence of the WPR in each task separately. Our main results assess latent-variable models for RT ranked bins and their correlations with WMC and

TUTs. We follow these results with latent variable models using ex-Gaussian parameters to assess the WPR (via the tau parameter). Lastly, we present a mini-multiverse analysis to explore whether varying treatments of outliers influence the robustness of our primary latent-variable analyses.

Ranked Bin Analyses

Descriptive Statistics and Zero-Order Correlations. Table 2 presents descriptive statistics for all ranked-bin measures. M RTs increased substantially across bins for all tasks, and standard deviations suggest considerable between-subject variation (also increasing over bins). Supplemental Table S1 presents zero-order correlations among the predictor and RT-outcome measures. Correlations among RTs from the same bins across different tasks (e.g., SART bin 5, arrow flanker bin 5) were modest, suggesting convergent validity among ranked-bin RTs. It thus appears that we measured a reasonably trait-like pattern in RT distributions across subjects.

Table 2. Descriptive Statistics for Ranked Bin Measures for each reaction time task.

Variable	Mean	SD	Min	Max	Skew	Kurtosis
SART Bin1	337	72	213	639	0.726	0.562
SART Bin2	421	95	237	823	0.506	0.202
SART Bin3	491	107	258	979	0.265	0.177
SART Bin4	576	122	279	1048	0.367	1.002
SART Bin5	781	182	326	1419	0.870	1.412
Letter Flanker Bin1	437	59	292	627	0.608	0.417
Letter Flanker Bin2	498	75	339	773	0.651	0.372
Letter Flanker Bin3	547	91	367	864	0.778	0.574
Letter Flanker Bin4	611	118	405	1028	0.982	1.157
Letter Flanker Bin5	778	202	450	1488	1.168	1.467
Arrow Flanker Bin1	389	37	260	527	0.306	0.486
Arrow Flanker Bin2	437	45	311	584	0.552	0.195
Arrow Flanker Bin3	471	53	343	669	0.691	0.476
Arrow Flanker Bin4	515	67	373	750	0.776	0.536
Arrow Flanker Bin5	636	113	427	1048	0.949	0.744
Circle Flanker Bin1	426	46	293	595	0.668	0.930
Circle Flanker Bin2	489	57	351	699	0.629	0.516
Circle Flanker Bin3	536	69	389	799	0.776	1.010
Circle Flanker Bin4	600	96	421	941	1.131	1.794
Circle Flanker Bin5	768	180	466	1360	1.339	1.938
Number Stroop Bin1	411	40	309	557	0.590	0.940
Number Stroop Bin2	478	47	366	658	0.490	0.831
Number Stroop Bin3	523	53	405	724	0.480	0.687
Number Stroop Bin4	574	64	441	824	0.648	0.790
Number Stroop Bin5	716	125	502	1167	1.228	1.799
Spatial Stroop Bin1	516	95	293	880	1.013	1.507
Spatial Stroop Bin2	596	118	382	1010	1.151	1.716
Spatial Stroop Bin3	661	139	410	1133	1.216	1.714
Spatial Stroop Bin4	751	179	432	1333	1.334	1.900
Spatial Stroop Bin5	991	307	514	1955	1.408	1.765

Note. SART = Sustained Attention to Response Task. Bin 1 = subjects' fastest quintile of RTs; Bin 5 = subjects' slowest quintile of RT

Regression Evidence for the Worst Performance Rule. We first present two sets of regression analyses to assess descriptive evidence for either the WPR or the not-best performance rule (Schubert 2019) across the RT tasks. The first set of regressions tested whether WMC, TUT rates, or both, interacted with RT Quantile Bin to predict RT. The WPR would be reflected in associations with WMC and/or TUTs getting stronger across the bins. That is, WMC- and TUT-related differences should be largest in subjects' slowest RT bin (i.e., Bin 5). Alternatively, evidence for not-best performance rule would come in the form of associations with WMC and/or TUTs increasing across subjects' fastest and "mean" RT bins (i.e., Bin 1 and Bin 2), but the slopes from "mean" to slowest RT bins should look similar. As seen in Table 3 (under the Model 1 column), across tasks, Bin was a significant predictor of RT (as it should have been, by design); RTs were longer at the later than earlier bins. WMC was also a significant predictor of RT in all tasks, except the SART. However, all tasks exhibited a significant Bin × WMC interaction. Supplemental Figure S1 depicts this interaction for each task. The relation between WMC and RT in the SART was unique, in that extremely short RTs, which likely reflect habitual "go" responding, were positively related to with WMC. That is, higher-WMC subjects' shortest RTs were longer than lower-WMC subjects', consistent with prior research (McVay & Kane 2009). As can be seen in Supplemental Figure S1, across many of the tasks, the beta coefficients numerically increased across the bins. However, across the tasks, the 95% confidence intervals tended to overlap across many of the non-fastest bins (e.g., 2 though 5). This suggests that subjects' mean to longest RTs might not be statistically different in their association to WMC, perhaps inconsistent with the WPR. In interpreting these patterns, however, it is important to note that when RTs are highly correlated across bins (see Supplemental Table S1 for correlations) and variability increases across bins, the regression slopes must also increase across bins (Frischkorn et al. 2016). Thus, the slope increases we see across bins might be artifacts and not sufficient evidence for the WPR.

Table 3. Hierarchical Regressions examining the interaction between Cognitive Predictors and Bin each task in predicting RT

	Model 1 (WMC)		Model 2 (TUTs)	
SART	B (SE)	*ß*	B (SE)	*ß*
Bin	104.655 (1.743)	0.759***	104.655 (1.743)	0.759***
WMC	-3.677 (7.795)	-0.009		
Bin X WMC	-20.330 (3.556)	-0.169***		
TUT			-5.470 (9.427)	-0.011
Bin X TUT			24.743 (4.299)	0.171***
Letter Flanker				
Bin	79.654 (1.857)	0.665***	79.724 (1.824)	0.759***
WMC	-31.405 (7.331)	-0.089***		
Bin X WMC	-9.604 (3.851)	-0.090*		
TUT			50.355 (8.434)	0.123***
Bin X TUT			20.369 (4.403)	0.165***
Arrow Flanker				
Bin	57.324 (1.030)	0.748***	57.268 (1.044)	0.747***
WMC	-26.804 (4.547)	-0.120***		
Bin X WMC	-9.373 (2.115)	-0.139*		
TUT			17.039 (5.480)	0.064**
Bin X TUT			8.411 (2.571)	0.104**
Circle Flanker				
Bin	80.380 (1.594)	0.714***	80.441 (1.606)	0.714***
WMC	-47.011 (6.587)	-0.146***		
Bin X WMC	-15.938 (3.220)	-0.169***		
TUT			46.040 (7.929)	0.119***
Bin X TUT			19.843 (3.890)	0.170***
Number Stroop				
Bin	70.970 (1.098)	0.795***	70.998 (1.102)	0.795***
WMC	-32.507 (5.342)	-0.125***		
Bin X WMC	-12.222 (2.261)	-0.156***		
TUT			32.592 (6.274)	0.107***
Bin X TUT			16.167 (2.661)	0.176***
Spatial Stroop				
Bin	112.604 (2.980)	0.616***	112.817 (3.012)	0.618***
WMC	-59.276 (10.940)	-0.113***		
Bin X WMC	-21.361 (6.060)	-0.139***		
TUT			13.043 (13.129)	0.021
Bin X TUT			25.725 (7.301)	0.136***

Note. SART = Sustained Attention to Response Task. ^ p < .10; * p < .05; ** p < .01; *** p< .001

We next ran the same analyses using TUT rates as our ability predictor. As seen in Table 3 (under the Model 2 column), Bin again predicted RT across the tasks, as it must. TUT rates significantly predicted RT in all the tasks except for SART and Spatial Stroop. Of most importance, the TUT × Bin interaction was significant across the tasks (Supplemental Figure S2 visualizes the interaction for each task). Again, we find a unique pattern of results in the SART: higher TUT rates were associated with shorter RTs in subjects' fastest bins (e.g., bin 1 and 2), likely reflecting absentminded "go" responding. Consistent across the tasks, though, we found that higher TUT rates associated with longer RTs in subjects' slowest bins (e.g., Bins 3-5). In many of the tasks, Bin 5 and Bin 4 had overlapping confidence intervals. However, Bin 5 confidence intervals often failed to overlap with Bin 3, suggesting that the association between TUT rate and RT was strongest for the longest RTs versus the mean RTs. Thus, when using TUT rate as our measure of ability, we find stronger descriptive evidence for the WPR than we did for WMC.

In the next set of regression analyses, we investigated the predictive power of RT bins on WMC and TUTs. Hierarchical linear regressions tested whether RT bins for the slowest quintiles predicted variation in WMC and TUTs after accounting for the fastest RT quintiles. Given the strong correlations between adjacent bins in each task (e.g., Bin 1 and Bin 2), we focused these and all subsequent analyses on Bin 1, Bin 3, and Bin 5. This approach also parallels Schubert's (2019) focus on "Fast RT" (i.e., Bin 1), "Mean RT" (i.e., Bin 3), and "Slow RT" (i.e., Bin 5).

If the longest RTs are the ones that are especially related to WMC and TUT (i.e., typical WPR findings), then the slowest RT bins should account for unique variance in WMC and TUT rate after accounting for subjects' fastest and mean RT bins. Table 4 shows the results of hierarchical regressions on WMC, which suggest that the slower bins do not add much predictive power beyond the faster bins. That is, after adding in Bins 3 and 5 to the models, Bin 1 or Bin 3 (or both) were the main predictors of WMC, rather than Bin 5. (We note the evidence of suppressor effects in many of the final models of each task; Bin 1 negatively predicted WMC in the initial models for each task, but that effect sometimes changed sign once the slower bins are added into the models). Overall, then, when WMC serves as the outcome, it appears that we have better evidence for the not-best performance rule (Schubert, 2019) than for the WPR.

Table 5 shows the parallel regression analyses for the TUT rate outcome. Here, TUTs were solely predicted by the slowest RT bins in several of the tasks. These TUT-related finding are more in line with the WPR than with the not-best performance rule. At the task level, then, it appears that we find evidence suggestive of either the WPR or the not-best performance rule, depending on the cognitive ability being assessed (not-best performance for WMC associations, worst performance for TUT rate associations).

Table 4. Hierarchical Regressions of WMC regressed on Bins 1, 3, and 5, for each Task

	Model 1		Model 2		Model 3	
SART	B (SE)	*ß*	B (SE)	*ß*	B (SE)	*ß*
Bin 1	0.001 (0.000)	.159***	0.003 (0.001)	.394***	0.002 (0.001)	.220*
Bin 3			-0.001 (0.000)	-.276***	0.000 (0.000)	.018
Bin 5					-0.001 (0.000)	-.247*
R^2	.025		.047		.079	
Δ*R^2*			.022		.032	
Letter Flanker						
Bin 1	-0.001 (0.000)	-.105*	0.002 (0.001)	.244*	0.003 (0.001)	.329**
Bin 3			-0.002 (0.001)	-.381***	-0.003 (0.001)	-.619***
Bin 5					0.000 (0.000)	.182^
R^2	.011		.034		.041	
Δ*R^2*			.022		.007	
Arrow Flanker						
Bin 1	-0.002 (0.001)	-.138**	0.003 (0.001)	.217*	0.003 (0.001)	.202*
Bin 3			-0.004 (0.001)	-.407***	-0.003 (0.001)	-.356*
Bin 5					-0.000 (0.000)	-.041
R^2	.019		.058		.059	
Δ*R^2*			.039		.001	
Circle Flanker						
Bin 1	-0.002 (0.000)	-.230***	0.001 (0.001)	.049	0.001 (0.001)	.048
Bin 3			-0.002 (0.001)	-.317***	-0.002 (0.001)	-.315*
Bin 5					-0.000 (0.000)	-.002
R^2	.053		.076		.076	
Δ*R^2*			.023		.000	
Number Stroop						
Bin 1	-0.002 (0.001)	-.135**	0.005 (0.001)	.410***	0.006 (0.001)	.457***
Bin 3			-0.006 (0.001)	-.621***	-0.007 (0.001)	-.730***
Bin 5					0.000 (0.000)	.083
R^2	.018		.106		.108	
Δ*R^2*			.088		.002	
Spatial Stroop						
Bin 1	-0.001 (0.000)	-.128**	0.001 (0.001)	.149	0.000 (0.001)	.092
Bin 3			-0.001 (0.000)	-.300*	-0.001 (0.001)	-.185
Bin 5					-0.000 (0.000)	-.074
R^2	.016		.030		.031	
Δ*R^2*			.014		.001	

Note. SART = Sustained Attention to Response Task. ^ $p < .10$; * $p < .05$; ** $p < .01$; *** $p< .001$

Table 5. Hierarchical Regressions of TUTs regressed on Bins 1, 3, and 5, for each Task.

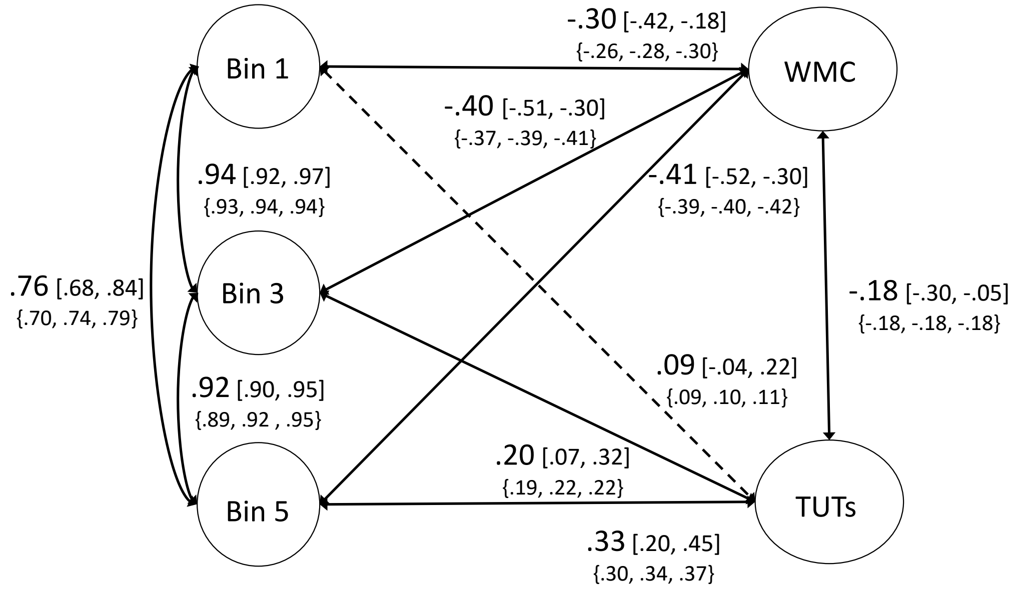
	Model 1		Model 2		Model 3	
SART	B (SE)	*ß*	B (SE)	*ß*	B (SE)	*ß*
Bin 1	-0.001 (0.000)	-.210***	-0.003 (0.000)	-.453***	-0.002 (0.001)	-.291**
Bin 3			0.001 (0.000)	.287***	0.000 (0.000)	.012
Bin 5					0.001 (0.000)	.231***
R^2	.044		.067		.094	
ΔR^2			.020		.027	
Letter Flanker						
Bin 1	0.001 (0.000)	.136**	-0.001 (0.001)	-.191^	-0.001 (0.001)	-.092
Bin 3			0.002 (0.001)	.358**	0.000 (0.001)	.080
Bin 5					0.000 (0.000)	.213*
R^2	.019		.039		.048	
ΔR^2			.020		.009	
Arrow Flanker						
Bin 1	0.000 (0.001)	.031	-0.003 (0.001)	-.241**	-0.002 (0.001)	-.180^
Bin 3			0.002 (0.001)	.312***	0.001 (0.001)	.112
Bin 5					0.001 (0.000)	.165
R^2	.001		.024		.029	
ΔR^2			.023		.005	
Circle Flanker	B (SE)	*ß*	B (SE)	*ß*	B (SE)	*ß*
Bin 1	0.001 (0.000)	.155***	-0.001 (0.001)	-.067	0.000 (0.001)	.031
Bin 3			0.001 (0.001)	.252**	-0.000 (0.001)	-.018
Bin 5					0.000 (0.000)	.220*
R^2	.024		.038		.051	
ΔR^2			.014		.013	
Number Stroop						
Bin 1	0.001 (0.00)	.089^	-0.003 (0.001)	-.295**	-0.001 (0.001)	-.101
Bin 3			0.003 (0.001)	.437***	-0.000 (0.001)	-.020
Bin 5					0.001 (0.000)	.345***
R^2	.008		.052		.080	
ΔR^2			.044		.028	
Spatial Stroop						
Bin 1	-0.000 (0.000)	-.107*	-0.003 (0.001)	-.672***	-0.002 (0.001)	-.513***
Bin 3			0.002 (0.000)	.612***	0.001 (0.001)	.286
Bin 5					0.000 (0.000)	.209^
R^2	.011		.068		.075	
ΔR^2			.057		.007	

Note. SART = Sustained Attention to Response Task. ^ p < .10; * p < .05; ** p < .01; *** p < .001.

Confirmatory Factor Analyses of Ranked Bins. We next assessed how binned RTs correlated with our cognitive predictors at the latent variable level. Like the above regression models, we included only RT bins 1, 3, and 5 to best parallel Schubert's (2019) meta-analytic findings (and to circumvent problems from extremely strong correlations between adjacent RT bins). A measurement model for just RT bins 1, 3, and 5 fit the data well, χ^2 /df = 2.40, CFI = .977, TLI = .970, RMSEA = .051 [.043-.059], SRMR = .052, indicating consistent individual differences in RT bins across our tasks. Even after dropping adjacent bins, however, some of the bins were highly correlated with each other, especially the closer bins ($\varphi_{bin1,3}$ = .94 ; $\varphi_{bin3,5}$ = .92). The correlation between Bin 1 and Bin 5 ($\varphi_{bin1,5}$ = .76) was still strong, but was numerically weaker than those of the closer bins.

Next, we asked how these factors correlated with WMC and TUT rates. Prior work on the WPR would suggest that cognitive abilities should correlate more strongly with the slowest RT bins than with the rest of the RT distribution. However, Schubert's (2019) meta-analysis suggested that an individual's cognitive ability is equally correlated with their mean RT and longest RTs, with both correlations stronger than with subjects' shortest RTs. A confirmatory factor analysis with WMC, TUTs, and RT bins (1, 3, 5) fit the data well, χ^2/df = 2.03, CFI = .964, TLI = .957, RMSEA = .044 [.039-.048], SRMR = .062. Figure 2 presents the full model. WMC was significantly negatively correlated with each RT bin. Of most importance, WMC appeared to be less strongly correlated with Bin 1 (φ = -.30), than with Bin 3 or Bin 5 (φs = -.40 and -.41, respectively). To test whether these estimates were statistically different from each other, we ran another CFA where the paths from WMC to Bin 1 and Bin 3 were set to be equal. Although this model fit the data well, χ^2/df = 2.24, CFI = .962, TLI = .956, RMSEA = .048 [.044-.053], SRMR = .065, it fit significantly worse than the model with all paths freely estimated, $\chi^2_{\text{diff}} = 19.99$, $df_{\text{diff}} = 1$, $p < .001$. WMC correlated less strongly with Bin 1 RTs than with the others, thus demonstrating the not-best performance rule.

Figure 2. Confirmatory factor analysis of ranked-bin model



Note. WMC = Working memory capacity. TUTs = Task-unrelated thought rate. Path estimates are presented in largest size font. 95% Confidence Intervals are presented in brackets. Values in the braces below represent the lowest, median, and highest estimate from the mini multiverse analysis. For clarity, factor loadings are not presented here; see Supplemental Table S2 for factor loadings for all models included in the primary analyses.

For TUT-rate correlations, in contrast, we find a pattern more consistent with the WPR. TUTs were not significantly related to subjects' fastest RT bin ($\varphi = .09$, $p > .05$), but they were to subjects' middle RT bin ($\varphi = .20$, $p < .05$) and slowest RT bin ($\varphi = .33$, $p < .01$). Here, we tested whether fixing the paths from TUTs to Bin 3 and Bin 5 to be equal significantly hurt model fit. In fact, fixing these correlations to be equal significantly hurt model fit, $\chi^2_{\text{diff}} = 8.49$, $df_{\text{diff}} = 1$, $p < .005$. Therefore, the pattern of correlations does appear to get stronger across the RT bins, consistent with traditional WPR findings. These results complement the task-based regression analyses and suggest that evidence for the WPR and not-best-performance rule depend on the cognitive ability construct being measured. Those abilities that are most closely tied to attentional lapses (i.e., TUTs) show more consistent evidence for the WPR, whereas those less strongly related to lapses (i.e., WMC) tend to show the not-best-performance pattern[4].

Ex-Gaussian Analyses

Descriptive Statistics and Zero-Order Correlations. As a second methodological approach to characterizing RTs (and worst performance), we used ex-Gaussian models to estimate three parameters from subjects' RT distributions for each of the tasks, mu, sigma, and tau. We conducted ex-Gaussian modeling with the retimes package (Massidda, 2015). Table 6 provides the descriptive statistics for the ex-Gaussian parameter estimates for each task. Supplemental Table S3 shows the bivariate correlations among the cognitive predictors and ex-Gaussian parameter estimates. Each parameter appeared to be modestly correlated across tasks, suggesting convergent validity, and in most cases each parameter correlated more strongly with its counterparts across tasks than with the other parameters across tasks, suggesting discriminant validity. Thus, as with RT bins, it appears that we measured trait-like patterns in ex-Gaussian RT distributions across subjects.

[4] As a secondary approach we attempted to fit latent growth curve models to the ranked bin data (Duncan et al 2006; Preacher et al 2008), but we were unable to fit the data with these models, likely a result of the high collinearity between the bin factors.

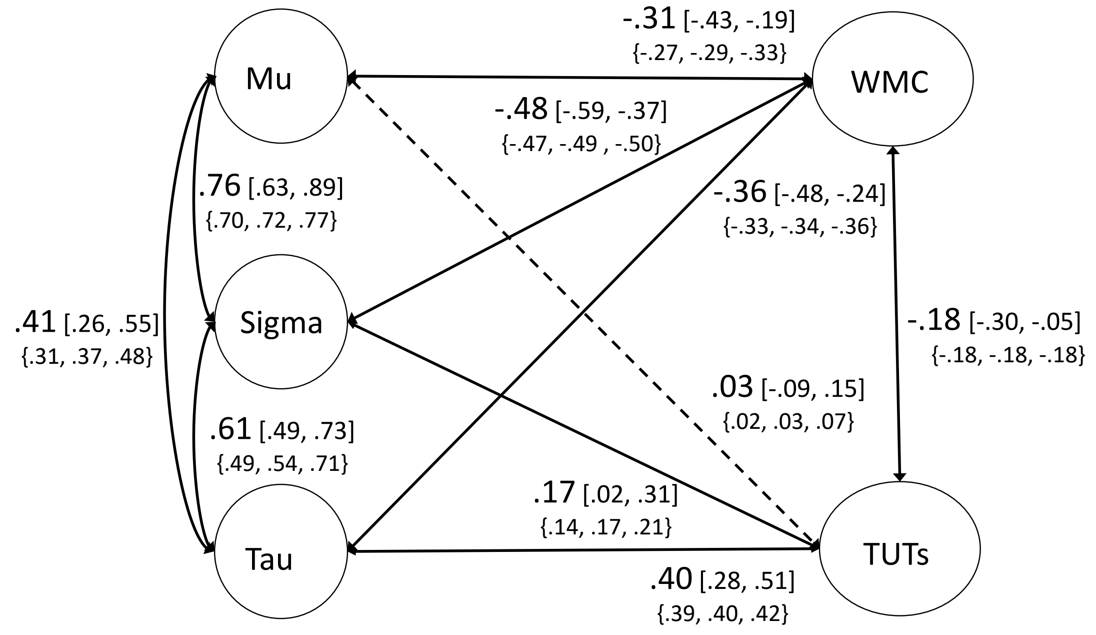
Table 6. Descriptive Statistics for ex-Gaussian Measures.

Variable	Mean	SD	Min	Max	Skew	Kurtosis
SART *Mu*	447	49	330	646	0.678	0.927
SART *Sigma*	49	19	0	119	0.309	0.571
SART *Tau*	114	61	1	303	1.263	1.704
Letter Flanker *Mu*	376	108	200	871	0.465	0.109
Letter Flanker *Sigma*	69	41	0	252	0.758	0.460
Letter Flanker *Tau*	144	71	3	386	1.147	1.670
Arrow Flanker *Mu*	446	42	356	608	0.446	0.639
Arrow Flanker *Sigma*	58	14	22	108	0.523	0.456
Arrow Flanker *Tau*	94	42	5	237	1.335	1.801
Circle Flanker *Mu*	407	39	309	533	0.506	0.300
Circle Flanker *Sigma*	39	15	0	91	0.917	1.366
Circle Flanker *Tau*	82	36	3	203	0.968	0.796
Number Stroop *Mu*	534	105	329	917	1.046	1.593
Number Stroop *Sigma*	58	30	0	167	1.024	1.852
Number Stroop *Tau*	165	98	3	486	1.353	1.816
Spatial Stroop *Mu*	456	65	310	702	0.645	0.444
Spatial Stroop *Sigma*	47	21	0	127	0.730	0.879
Spatial Stroop *Tau*	115	64	2	345	1.210	1.851

Note. SART = Sustained Attention to Response Task.

Ex-Gaussian Structural Models. We next attempted to model latent variables from the ex-Gaussian variables. Model fit was acceptable, χ^2 /df = 2.74, CFI = .940, TLI = .920, RMSEA = .052 [.044-.059], SRMR = .066. Positive correlations among the ex-Gaussian factors were moderate to strong, in line with prior work using this technique (e.g., Schmiedek et al. 2007). We next added both WMC and TUTs into the model as a confirmatory factor analysis. This model fit the data adequately, χ^2/df = 2.15, CFI = .920, TLI = .905, RMSEA = .046 [.042-.051], SRMR = .065). As seen in Figure 3, WMC correlated significantly negatively with each parameter estimate, not just with tau. These estimates do not follow a worst-performance-rule pattern (i.e., the correlation with mu is substantial, and the strongest WMC correlation is with sigma rather than tau). We tested whether fixing the paths between WMC and mu and tau significantly hurt model fit; it did not, χ^2_{diff} = 0.12, df_{diff} = 1, p > .05. TUT rates showed a different pattern. TUT rate was not significantly correlated with mu (φ = .03) and was weakly associated with sigma (φ = .17). Importantly, however, TUT rate was moderately correlated with tau (φ = .40). As we did with WMC, we tested whether fixing the paths between TUTs and mu and tau hurt model fit, and here it did, χ^2_{diff} = 27.64, df_{diff} = 1, p < .001. This suggests that subjects who were more prone to lapses of attention associated with mind wandering also had more behavioral lapses (i.e., especially long RTs) captured by the tau parameter. Thus, it again appears that TUT-rate variation shows the worst-performance rule pattern.

Figure 3. Confirmatory factor analysis of ex-Gaussian model



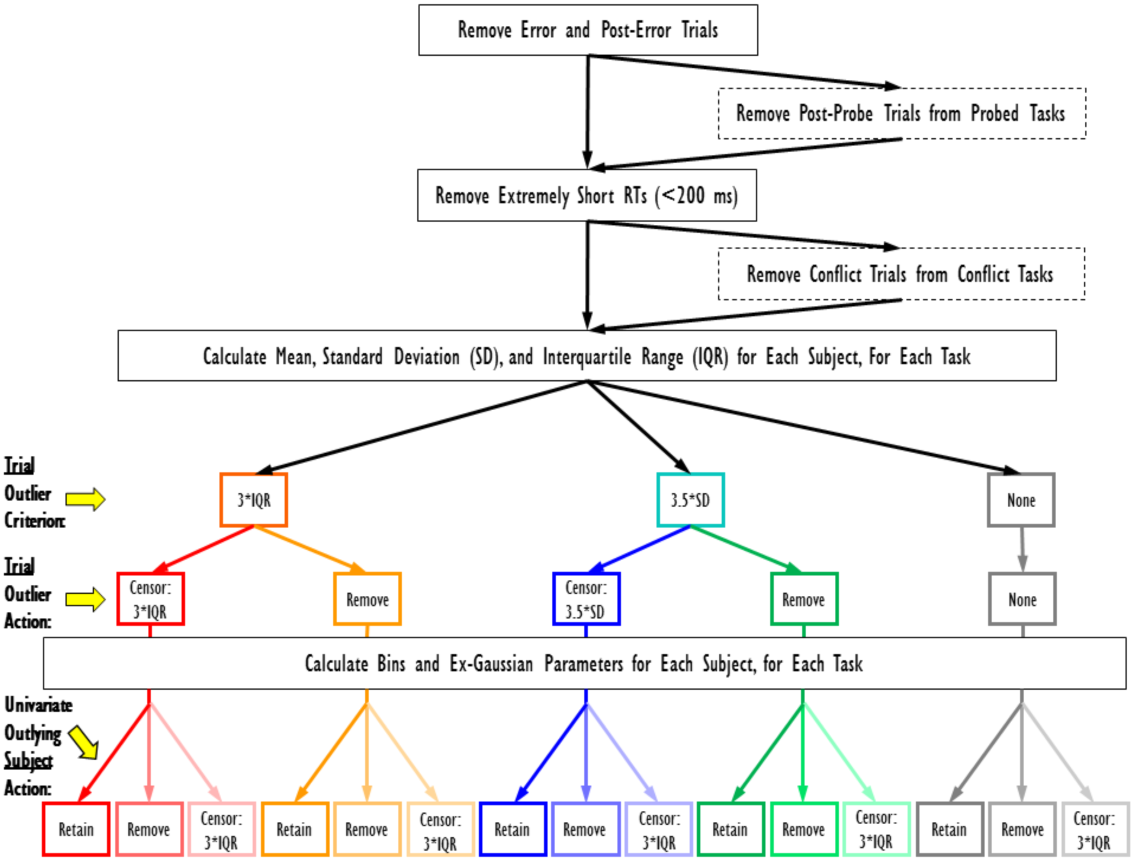
Note. WMC = Working Memory Capacity. TUTs = Task-Unrelated Thoughts. Path estimates are presented in largest size font. 95% Confidence Intervals are presented in brackets. Values in the braces below represent the lowest, median, and highest estimate from the mini multiverse analysis. For clarity, factor loadings are not presented here; see Supplemental Table S2 for factor loadings for all models included in the primary analyses.

Mini-Multiverse Analysis of WPR findings

Researchers that conduct binning and ex-Gaussian analyses of RTs have many degrees of freedom in how they treat the data corresponding to the upper limit of the RT distribution. While some relatively long RTs may be characteristic of an attentional lapse, it is possible that other, perhaps outlying, RTs result from idiosyncratic or unplanned events (e.g., sneezes, looking away from the monitor, checking a phone) that aren't characteristic of a subject's performance or ability. How should the data analyst handle these long or outlying RTs, particularly when WPR-related phenomena are driven by exactly those longer-than-average RTs? There is no single answer. While many WPR studies report some RT outlier treatment, there are almost as many treatment variations as there are studies. In just the 23 studies included in Schubert's (2019) meta-analysis, 9 papers did not describe any RT outlier treatment and the remaining 14 each had different criteria and protocols. Some of these treatments were simple (e.g., removing the slowest RT trial), while others were more complex (e.g., an iterative process that removed outlying trials until none remained). The most common approach was that of defining a cut-off based on each subjects own RT distribution (e.g., Mean RT + 3.5*SD) and discarding trials that were slower than this criterion.

Differences in cutoff values for outlying RTs might alter RT distributions, and their correlations with cognitive abilities, across studies. To examine this possibility, we created a mini-multiverse of potential datasets based on various outlier cutoff criteria and consequence (see Steegen et al. 2017); we describe this as a mini-multiverse because we did not assess every possible combination of possible (or plausible) data treatments. The processing of data is an active process in which many decisions can be made (e.g., outlier cutoffs). Thus, the raw dataset that researchers begin with can ultimately yield different datasets based on different outlier decisions (i.e., multiverses). To increase transparency and test the robustness of our main latent-variable findings, we created variations of the original dataset based on different RT cutoff values for outliers (e.g., mean RT + 3*IQR; mean RT + 3.5*SD) and whether trials outside of those cutoffs were either (a) removed completely or (b) censored to the cutoff value before aggregating. We also created versions that took into account the potential impact of univariate outlier subjects after aggregating the data. This univariate outlier rule was based on 3*IQR and was used across all multiverse paths. Figure 4 depicts our decisions in creating the multiverse. Again, these decisions are not exhaustive, and an infinite set of other cutoffs could be plausibly chosen (e.g., Mean RT + 2.5*SD, Mean RT + 2.75*SD, Mean RT + 3*SD, etc.). To foreshadow, our findings were impressively consistent across different iterations of the multiverse, suggesting that deviations across our decisions did not affect our outcomes and conclusions. Whether this is generally true, at least in studies with large sample sizes that take a latent-variable approach across multiple RT indicators, remains to be determined by multiverse analyses of other studies.

Figure 4. Mini-Multiverse Decision Tree

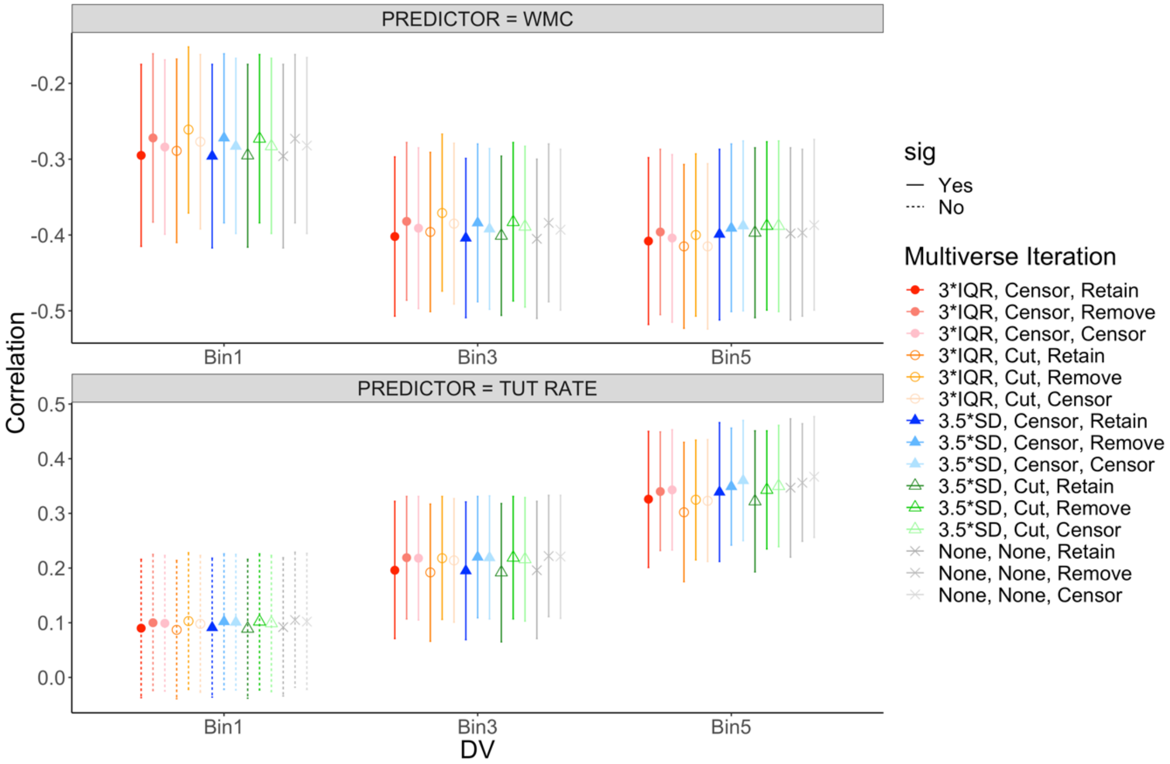

Remove Error and Post-Error Trials
Remove Post-Probe Trials from Probed Tasks
Remove Extremely Short RTs (<200 ms)
Remove Conflict Trials from Conflict Tasks
Calculate Mean, Standard Deviation (SD), and Interquartile Range (IQR) for Each Subject, For Each Task
Trial Outlier Criterion:
3*IQR
3.5*SD
None
Trial Outlier Action:
Censor: 3*IQR
Remove
Censor: 3.5*SD
Remove
None
Calculate Bins and Ex-Gaussian Parameters for Each Subject, for Each Task
Univariate Outlying Subject Action:
Retain
Remove
Censor: 3*IQR
Retain
Remove
Censor: 3*IQR
Retain
Remove
Censor: 3*IQR
Retain
Remove
Censor: 3*IQR
Retain
Remove
Censor: 3*IQR

Note. Solid black boxes represent decisions that were made in every task in every multiverse iteration. Dashed black boxes include decisions that were made in some tasks (i.e., those with thought probes or conflict trials) in every multiverse iteration. Retain = kept outlier in data set. Remove = remove outlier (trial or subject) from data set. Censor = change outlying value to specified cut-off.

Mini-Multiverse Results

Supplemental Table S4 presents the latent correlations among WMC, TUT rates, and our Bin factors across the various multiverse iterations. These results are visually depicted in Figure 5. Estimates of these associations are remarkably stable across iterations, with correlations within a range of +/- .06. Thus, changing the outlier cutoff for individual trials, cutting, censoring, or retaining those outlier trials, and deciding whether or not univariate outliers should be included, cut, or censored did not substantively alter the estimates of the relations between our cognitive ability factors and RT bins. As in our main analysis reported above, WMC was negatively related to each RT Bin, and this pattern reflected the not-best performance rule: WMC showed weaker correlations with subjects' shortest RTs and numerically similar estimates for subjects' mean and longest RTs. As well, the association between TUT rate and the RT bins followed an identical pattern to the main analyses: TUT rates were not related to subjects' shortest RTs, were weakly associated with subjects' mean RTs, but were more strongly related to subjects' longest RTs. Thus, across the mini-multiverse, we see evidence for the WPR only when examining TUT propensity as our cognitive ability measure.

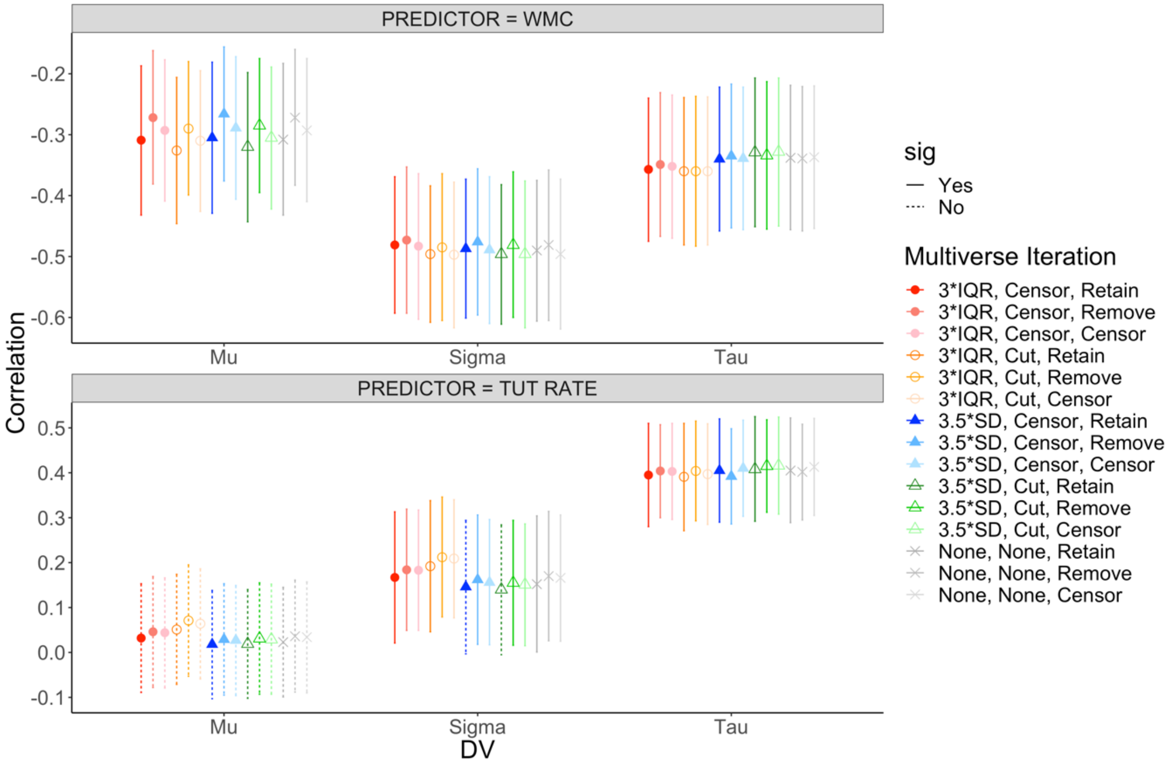
Figure 5. Mini-Multiverse of Ranked-Bin Correlations



Note. The top panel presents correlations with Working Memory Capacity (WMC). The bottom panel presents correlations with rate of Task-Unrelated Thought (TUT). Points reflect the correlation with error bars representing the 95% confidence interval (CI) around the estimate. Circles represent iterations where outlying trials were defined by interquartile ranges (IQR), triangles represent iterations where outlying trials were defined by standard deviations (SDs), and xs represent iterations where no criteria were applied to outlying trials. Filled shapes reflect iterations where outlying trials were censored to the respective cut-off value before aggregating and open shapes reflect iterations where outlying trials were removed before aggregating. Colors presented in this Figure match those illustrating the multiverse iterations in Figure 6. Solid CIs represent significant correlations, dashed CIs represent non-significant correlations at p = .05.

We next examined the impact of mini-multiverse decisions on the associations with the ex-Gaussian parameter estimates. Supplemental Table S5 provides the latent variable correlations between WMC, TUTs, and the ex-Gaussian parameter estimates across multiverse iterations. These results are visually depicted in Figure 6. Again, the range of estimates across the multi-verse was small, +/- .07, suggesting high reliability across iterations. The correlations between WMC and the ex-Gaussian parameters were consistent with our main analysis presented earlier: WMC was modestly (and equivalently) correlated with *mu* and *tau* and more strongly correlated with *sigma*. The patterns for TUT rates were also consistent with our main analysis. TUTs were not significantly associated with *mu* in any iteration of the multi-verse. The association with *sigma*, however, did vary somewhat, and in two cases did not reach significance ($p > .05$). However, given that this estimate was the weakest to begin with, it is not surprising that some multiverse paths were not significant. TUT rate's strong positive correlation with *tau* was consistent across the multiverse. Our multiverse analyses of the ex-Gaussian parameters, then, found patterns consistent with both the not-best-performance rule and the WPR, depending on our measure of cognitive ability.

Figure 6. Mini-Multiverse of ex-Gaussian Correlations



Note. The top panel presents correlations with Working Memory Capacity (WMC). The bottom panel presents correlations with rate of Task-Unrelated Thoughts (TUT). Points reflect the correlation with error bars representing the 95% confidence interval (CI) around the estimate. Circles represent iterations where outlying trials were defined by interquartile ranges (IQR), triangles represent iterations where outlying trials were defined by standard deviations (SDs), and xs represent iterations where no criteria were applied to outlying trials. Filled shapes reflect iterations where outlying trials were censored to the respective cut-off value before aggregating and open shapes reflect iterations where outlying trials were removed before aggregating. Colors presented in this Figure match those illustrating the multiverse iterations in Figure 6. Solid CIs represent significant correlations, dashed CIs represent non-significant correlations at p = .05.

Discussion

We reanalyzed data from a large latent-variable study (Kane et al., 2016) to test the robustness of the WPR (or the not-best performance rule; Schubert 2019) across a variety of demanding attention-control tasks. We used two approaches, ranked RT bins and ex-Gaussian estimation, to describe the RT distributions across tasks. In doing so, we assessed latent variables and tested their associations with two cognitive ability constructs, WMC and propensity for

TUTs. Our primary findings complement both traditional findings of the WPR and recent meta-analytic claims that cognitive ability is equally predictive of mean and longest RTs, compared to shortest RTs (Schubert, 2019). Specifically, WMC showed consistent patterns, at both the task level and latent-variable level, of the not-best performance rule: WMC least strongly predicted subjects' shortest RTs, but was more strongly—and equally—correlated with their mean and longest RTs; ex-Gaussian analyses showed that WMC correlated at least as strongly with the Gaussian parameters of *sigma* and *mu* as it did with *tau*. TUT rate, on the other hand, showed trends more consistent with the WPR. TUTs were not related to subjects' shortest RTs (or the *mu* parameter) and were weakly associated with mean RTs; instead, TUT rate correlated most strongly with subjects' longest RTs (i.e., with both RT Bin 5 and the *tau* parameter). Thus, our results suggest that claims about cognitive ability and worst performance may depend on the ability construct in question. Cognitive abilities that are strongly related to attentional lapses and sustained attention (i.e., propensity for TUTs as assessed by in-task thought probes) may show patterns consistent with the WPR, whereas those that are less strongly related to attentional lapses (i.e., WMC) may show the not-best performance rule.

It is important to note, however, that WMC was not *unrelated* to long RTs (i.e., Bin 5) or *tau*. In fact, the WMC correlations here were of similar magnitude to those of the TUT rate. Instead, WMC correlated with worst *and* mean performance to a similar degree (and best performance to a lesser degree), while TUTs primarily correlated only with worst performance. What might contribute to these different patterns? The association with worst performance is likely driven in part by attention-control ability, which is central to both WMC and TUT propensity. Specifically, the TUT-RT findings are largely supportive of the attentional control theory of WPR. Individuals with poor attentional control, and thus higher likelihood of mind-wandering, will experience more attentional lapses than those with better control ability. These occasional attentional lapses result in occasional extremely long RTs that are reflected in the tail of that individuals RT distribution (i.e., *tau* and the slowest RT bin). However, WMC and TUTs are multidetermined constructs, and so combinations of other processes likely also contribute to their associations with RT variables. There are likely many cognitive processes (executive and otherwise) that are associated with WMC, but not TUTs, that also contribute to average RT—such as stimulus-response binding (Wilhelm & Oberauer 2006), speed-accuracy trade-off (Unsworth & Engle 2008), working memory load (Shahar et al. 2014), encoding ability (Unsworth & Spillers 2010), and evidence-accumulation processes (Schmiedek et al. 2007)—and variation in these additional processes contribute to the not-best performance rule pattern for WMC. Thus, the processes that contribute to performance on fast and average RT trials seem to overlap more with WMC processes (and executive processes related to WMC) than with TUT-related processes (Kovacs & Conway 2017).

A methodological issue that arises when assessing the WPR (or any RT or performance phenomenon in psychological science) is how to treat outlier trials and outlying subjects. As noted in the introduction, reporting of such outlier treatment was scarce in the articles included in Schubert's (2019) meta-analysis of the WPR. This is unfortunate. Bakker and Wicherts (2014) investigated whether simply reporting the removal of outliers was related to weaker evidence in a set of RT studies. Although they found no difference in the strength of evidence between studies that did versus did not report outliers, they did find that there were issues in reporting and suggested there was a common failure to report exclusions or missing data. Bakker and Wicherts argued for greater transparency in reporting of outliers and statistical analyses, and we agree (see also Ley et al. 2019 for a discussion on how to identify and handle outliers in a study).

To explicitly probe the issue of outlier treatment—which prior WPR studies have not considered systematically—we created a mini-multiverse of outlier treatments at both trial and subject levels that are common to the literature (including no treatment). We then re-ran our primary confirmatory factor analyses across these iterations to investigate whether they altered associations between cognitive-ability constructs and aspects of the RT distributions. They did not. That is, the results of our primary analyses replicated across multiverse iterations. Thus, in a study that collects RTs across multiple tasks per subject, and does so for hundreds of subjects, outlier treatment does not significantly affect the assessment of worst performance and individual differences therein. Our multiverse findings cannot say whether outlier decisions are equally irrelevant to conclusions drawn from smaller-N studies using single tasks.

We must acknowledge the study's limitations, however. First, although we analyzed RTs from only non-conflict trials from six tasks, all the tasks presented some conflict trials, thus creating an "attention control" context; our findings thus might not generalize to simple or choice RT tasks without conflict trials included. Second, although our RT tasks created an attention-control context, they did not impose significant memory demands. Prior work suggests that such memory demands (i.e., more choices in choice-RT tasks, or arbitrary response mappings) may make the WPR more apparent (Meiran & Shahar 2018; Shahar et al. 2014). For example, Rammsayer and Troche (2016) found a stronger link between WPR and psychometric *g* in 1- and 2-bit versions of the Hick task, compared to the simpler 0-bit version. More complex tasks, such as problem-solving tasks, might also elicit stronger WPR patterns than not-best performance rule patterns (Kranzler 1992; Ratcliff, Tahpar, & McKoon 2010); at the same time, the more complex a task becomes, the more executive processes may become involved in successful performance, which might yield stronger evidence for the not-best performance rule. Whether one finds evidence for the WPR or the not-best performance rule might therefore vary with both the nature of the cognitive ability construct and the cognitive demands of the RT tasks. An additional limitation of this study is that our assessment of sustained attention ability relied solely on self-reported TUTs. Although these reports have generally be found to be valid indicators of ones' propensity (and, presumably, ability) to sustain attention, they are not pure indicators of ability. Future WPR research should therefore assess performance measures of sustained attention ability, such as RT variability, vigilance decrements, or even pupil size, to test whether the WPR versus not-best performance rule patterns reported here also obtain with objective rather than self-report measures.

CHAPTER III: A NOMOTHETIC SPAN APPROACH TO THE CONSTRUCT VALIDITY OF SUSTAINED ATTENTION MEASUREMENT: RE-ANALYZING TWO LATENT-VARIABLE STUDIES OF PERFORMANCE VARIABILITY AND MIND-WANDERING SELF-REPORTS

Abstract

Sustained attention is frequently assessed using either objective behavioral measures, such as reaction-time (RT) variability, or subjective self-report measures, such as rates of task-unrelated thought (TUT). The current studies examined whether the *individual-difference covariation* in these measures provides a more construct valid assessment of sustained attention ability than does either alone. We argue that performance and self-report measures mutually validate each other; each measurement approach has its own sources of error, so their shared variance should best reflect the sustained attention construct. We reanalyzed two latent-variable studies where RT variability and TUTs were measured in multiple tasks (Kane et al., 2016, *Journal of Experimental Psychology: General, 145*, 1017-1048; Unsworth et al., 2021, *Journal of Experimental Psychology: General, 150*, 1303-1331), along with several nomological-network constructs to test the convergent and discriminant validity of a general sustained attention factor. Confirmatory factor analyses assessing bifactor (preregistered) and hierarchical (non-preregistered) models, suggested that sustained attention can be modeled as the shared variance among objective and subjective measures. This sustained attention factor was related to working memory capacity, attention control, processing speed, state motivation and alertness, and self-reported cognitive failures and positive schizotypy. Multiverse analyses of outlier decisions suggested that bifactor models of general sustained attention ability are less robust than hierarchical models; exploratory latent profile analyses provided converging evidence that poorer

sustained attention was associated with lower scores on many of the constructs of interest. The results provide evidence for the general ability to sustain attention and suggestions for improving its measurement.

Introduction

People sometimes strive to keep their attention directed on their current task and goals but do so with varying success. We may neglect to attach a file to an e-mail message, forget to stop at the grocery store on the way home from work, or even fail to check our surroundings before driving our car in reverse. Everyday observations suggest that some people better sustain their attention than do others, showing more consistent performance with fewer behavioral lapses, and experiencing fewer instances of their thoughts being captured by personal concerns. What might account for these individual differences?

Despite attentional lapses being partially responsible for real-world errors, the ability to sustain attention has been less thoroughly studied by psychologists than have other components of attention, such as selective, divided, and switching attention (Esterman & Rothlein, 2019). And, despite sustained attention supporting the regulation and control of other cognitive processes and behavior, it has been understudied relative to the executive functions of inhibition, updating, and switching (Miyake & Friedman, 2012). Research has nonetheless identified distinct, yet correlated, empirical measures that reflect sustained attention failures—variability in task performance and self-reports of mind wandering—but it has not yet considered that the overlap in these measures might be the most valid reflection of sustained attention (in)ability. The goal of the current study is to investigate and evaluate the construct validity of sustained attention measurement from a nomothetic span, or individual differences, perspective. We investigate whether there are stable individual differences in sustained attention failures, as indicated by performance and self-report measures, and if so, ask what other psychological or contextual factors might predict them.

The construct of sustained attention is not new, of course, and several models have explored how and why attention fluctuates (for reviews see Esterman & Rothlein, 2019; Fortenbaugh, et al., 2017). Traditionally, sustained attention has been viewed—and studied empirically—as the ability to maintain performance over many task trials (and many minutes). Failures of sustained attention from this view correspond to the so called "vigilance decrement" (e.g., Lim & Dinges, 2008; Mackworth, 1950; Parasuraman, 1986). Here, performance—be it reaction time or accuracy, usually in detecting rare target signals—worsens as time on task increases.

Our approach to sustained attention focuses instead on the moment-to-moment stability of attention, or "attentional consistency" (Unsworth & Miller, 2021), which may be a (partially) distinct form of sustained attention from that reflected in the general worsening of performance over time (Thomson et al., 2015). Specifically, we define sustained attention as *the purposeful act of maintaining optimal task focus to successfully, and consistently, perform goal-relevant actions*.

Sustained Attention as the Covariation of Objective and Subjective Measures

The cognitive psychology literature has taken two approaches to measuring attention consistency. In the following sections we describe both, which we refer to as *objective* (based on performance data) and *subjective* (based on self-report data). Each section describes how these measures reflect sustained attention (in)ability and their limitations when used in isolation. In doing so, we will argue that a combination of objective and subjective indicators, and specifically their shared variance, will provide the most valid assessment of sustained attention ability and its individual-differences variation.

We build this argument on the precedent that in some traditional sustained attention tasks, different performance indicators seem to reflect different types or degrees of sustained attention failures (Cheyne et al., 2009; Unsworth et al., 2021). For example, Cheyne et al. (2009) found that three performance measures all predicted unique variance in no-go accuracy in a go/no-go sustained attention task. Each of these measures also mapped on to three distinct, hypothetical attentional states with increasing levels of disengagement. RT variability reflects State 1 (*focal inattention)*, which is characterized by brief periods of attentional instability and stimulus processing, that produces errors, near misses, and variable performance; anticipations reflect State 2 (*global inattention)*, where top-down attention is disengaged from the current task to the point where automatic, "mindless" behaviors take over. Finally, omissions reflect State 3 (*behavioral/response disengagement),* where subjects' attention is withdrawn from the task to the point where they fail to engage in any task-appropriate responding. Although more theoretical and empirical work needs to be done to convincingly establish an inattention or disengagement continuum (Tay & Jebb, 2018), modeling the overlap of various measures of sustained attention performance (and different methodological approaches) may be a more construct valid way of assessing sustained attention ability than relying solely on any one type of error-prone measure.

Objective (Performance-Based) Measures of Attention Consistency

Reaction Time Variability

Optimal sustained attention performance can be measured as the magnitude of a subject's RT variability across a task, or as the rate or durations of a subject's relatively long RTs within a task (Bunce et al., 1993, 2004; West et al., 2002). That is, if a subject is effectively sustaining focused attention across a task that makes consistent cognitive demands across trials, then their RTs should be similar from trial to trial. RT variability and extremely long RTs reflect how *consistently* (or *inconsistently*) a subject performs a repetitive task.

Early work by Bills (1931, 1935) showed that after extended periods of continuous work on a task, subjects started to occasionally show very long RTs (e.g., twice the mean; "blocks"), which were often followed by more variable or erroneous performance (see also Bertelson & Joffe, 1963, Fiske & Rice, 1955; Sanders & Hoogenbroom, 1970). In modern tasks, like the psychomotor vigilance task (PVT), subjects must maintain focus for some variable and unpredictable duration (typically 1–10 s) before the stimulus numbers begin counting upwards, which is the signal for subjects to hit a key to stop the clock. Here, the number of "lapses" (i.e., RTs > 500 ms) is frequently used as a dependent measure that represents sustained attention (in)ability (e.g., Lim & Dinges, 2008; Unsworth & Robison, 2016). Like blocks, the number of lapses reflects variation in sustained attention because they seem to capture instances where subjects are not optimally task-focused (i.e., they're not optimally ready to respond to the target digits beginning to count upward).

Attention consistency is also assessed via trial-to-trial variability in RT in some tasks. Here, subjects with better sustained attention should show lower RT variability, with few very short or very long RTs. Common approaches to measuring RT variability include intra-individual standard deviation (RTsd), coefficient of variation (CoV), or Rhythmic Response Times (RRTs). RTsd simply takes the standard deviation of RTs across correct trials for each subject. CoV expresses an individual's RTsd as a function of their mean RT (CoV = [SD / *M*] * 100). RRTs also reflect consistency of responding but are calculated as the difference between response and stimulus onsets (Laflamme et al., 2018; Seli et al., 2013), so they can be positive (responding after the stimulus appears) or negative (responding before the stimulus appears); RRTs are often taken across a set number of trials (e.g., 5) to create a moving window of response variability calculated across the entire task.[5]

Finally, fluctuations in sustained attention can be assessed by fitting a subject's RTs to a distributional model. Ex-Gaussian models, for example, a convolution of a Gaussian (normal) and an exponential distribution, provide three parameters: μ and σ represent the mean and standard deviation of the Gaussian component, respectively, while τ represents the mean and standard deviation of the exponential component (i.e., the tail). In general, τ reflects increased variability in RTs in the form longer-than-average RTs and may capture sustained attention (in)ability to some degree (although ex-Gaussian parameters are not purely mapped onto any one or several psychological processes; Matzke & Wagenmakers, 2009).

Performance Accuracy

Accuracy-based measures may also reflect sustained attention lapses, at least in part. In the Sustained Attention to Response Task (SART), errors of omission (i.e., not responding to a "go" trial) and errors of commission (i.e., erroneously responding to a "no-go" trial) might reflect even greater task disengagement than is captured by variable responding (Cheyne et al., 2009; Unsworth et al., 2021). That is, errors of omission might reflect a complete disengagement from the task whereas errors of commission might reflect being captured enough by monotonous responding that individuals keep making repetitive responses when they are not supposed to.

[5] The heart rate literature provides variability indicators that might be profitably considered in sustained-attention research (Pham et al., 2021). Difference-based indices like the root mean square of successive difference (RMSSD) capture differences between successive intervals and capture short-term variations in heartrate.

Likewise, during continuous tracking tasks, in which subjects attempt to closely follow an object onscreen with a stylus or cursor, subjects may occasionally exhibit “flat spots,” or brief instances where they fail to respond to the stimuli (Peiris et al., 2006; Unsworth et al., 2021).

Limitations of Objective Indicators of Attention Consistency

Like other cognitive ability measures, objective indicators of attention consistency are not process-pure. Longer-than-normal RTs can certainly be caused by attention lapses. But subjects can also experience long RTs simply because they are generally slower than other subjects, or because they momentarily changed their response strategy (e.g., speed-accuracy trade-offs or post-error slowing). Long RTs can also result from involuntary actions (e.g., sneezing or yawning) or cases where subjects take intentional “rest breaks” during a trial. Further, task-specific processes, unrelated to sustained attention, may affect performance variability, especially if a task presents trial types with differing cognitive demands (e.g., Stroop tasks or Sternberg item-recognition tasks).

Thus, when assessing attention consistency using solely objective indicators, any one performance measure won’t fully capture all types or all instances of disengagement (and it will capture extraneous sources of measurement error). Rather, the performance variance that is common, or shared, across several of these objective measures should better reflect sustain attention abilities. Moreover, combining additional, non-performance indicators of sustained attention with performance assessments may provide for still more valid measurement of sustained attention, as we discuss below.

Subjective (Self-Report-Based) Indicators of Attention Consistency

Objective indicators, like RTsd, may capture relatively subtle fluctuations in sustained attention. However, some sustained-attention failures may be more obvious, and perhaps conspicuous enough to be easily reported by subjects when asked. Self-report measures of sustained attention aim to capture off-task thought experiences that are characteristic of everyday attention failures.

One commonly used, subjective approach to assessing attention consistency, both in the lab and in everyday life, is the thought-probe method. This technique is most frequently used to capture subjects' mind-wandering (or task-unrelated thought; TUT) experiences as they occur, and has been used in a variety of tasks and contexts, including attention tasks (e.g., Hutchison et al., 2020; Kane et al., 2016; McVay & Kane, 2012a), reading tasks (e.g., Franklin et al., 2014; McVay & Kane, 2012b; Unsworth & McMillan, 2014), driving simulations (e.g., Albert et al., 2018; Baldwin et al., 2017; He et al., 2011), live classroom or virtual learning environments (e.g., Hollis & Was, 2016; Kane, Carruth et al., 2021; Wammes, Seli et al., 2016), and in everyday life (e.g., Kane et al., 2007; 2017; Killingsworth & Gilbert, 2010; Marcusson-Clavertz et al., 2016). Here, subjects are repeatedly and unpredictably interrupted during a task or activity and asked to report on the contents of their thoughts in the moment immediately preceding the probe appearance. Subjects typically indicate whether they were focused on the task or were experiencing TUTs.

Thought-Probe Methods and Measures

Various aspects of mind wandering have been investigated using thought-probe methods (see Seli et al., 2018; Weinstein, 2018). In some studies, subjects answer a simple "yes/no" question about whether they were focused on the task or mind wandering (e.g., Franklin et al., 2014; Song & Wang, 2012; Szpunar et al., 2013). Other studies present thought-choice menus that allow subjects to select among categories or qualities of thoughts, such as thought content (e.g., worries, fantastical daydreams), temporal orientation (e.g., past events, future goals), emotional valence (e.g., positive, negative), or intentionality (e.g., deliberate, spontaneous; Banks et al., 2016; Smallwood et al., 2009; Stawarczyk et al., 2011; 2013; Unsworth & McMillan, 2014). Still others have used Likert scales to rate depth or intensity of mind wandering (e.g., Allen et al., 2013; Christoff et al., 2009). Thus, much like there are different tasks in which performance variability is measured as objective indicators of sustained attention, there are a variety of ways to subjectively assess sustained attention failures that are experienced as mind wandering.

The typical measure derived from thought probes is TUT rate (number of TUT reports/number of probes) which estimates the frequency with which subjects are not focusing on the task at hand. Subjects report being off-task 30–60% of the time, on average, suggesting that TUTs occur frequently across artificial and authentic contexts. As will be discussed in subsequent sections, TUTs are associated with poorer task performance, further validating that they reflect momentary failures of sustained attention.

Limitations of Subjective Indicators of Attention Consistency

Like performance measures of attention consistency, TUT reports come with confounds and concerns to consider (Kane, Smeekens et al., 2021). Most obviously, as these self-reports rely on introspection, we must consider the potential influence of reporting biases and errors (Hurlburt & Heavey, 2001; Nisbett & Wilson, 1977).

Thought reports to probes might be impacted by the frequency with which probes occur in the task. Too frequent probing might provide reminders to stay on task or not give enough time for subjects' minds to wander between probes, whereas too infrequent probing may miss instances of mind wandering that occur between probes (Welhaf et al., in press). Only a few studies have explicitly examined this possibility and the findings are mixed. Robison et al.

(2019) found that more frequent probing (13% vs. 7% of total trials) did not influence TUT rates. However, studies by both Seli, Carriere et al. (2013) and Schubert et al. (2019) found that more frequent probing (across ranges of 1%–6% of trials) resulted in lower TUT rates, suggesting that frequent probes act as on-task reminders or thought-flow disruptors.

An additional concern about probing during a task is that responses to probes might be biased by reactivity to performance. That is, when subjects make an error and a thought probe follows that error, subjects may use their performance as evidence for where their thoughts were focused. Few studies have examined this possibility (Head & Helton, 2018; Kane, Smeekens et al., 2021; Schubert et al., 2019), but they suggest some reactivity in tasks that elicit salient errors (e.g., go/no-go tasks). Schubert et al. (2019), for example, found that TUT reports in a SART were more frequent following “no-go” compared to “go” trials and that TUT reports were more frequent following “no-go” errors compared to correct “no-go” trials.

Although thought probes vary across studies, recent work suggests that some findings are robust across different thought-probe variations. For example, Kane, Smeekens et al. (2021) found similarities in *M* TUT rate, TUT rate reliability across tasks, within-person associations between TUTs and go/no-go performance, and between-person associations with theoretically relevant constructs (e.g., executive-control ability) across four different probe types. These findings provide generally supportive evidence for acceptable construct validity of the thought probe method.

Kane, Smeekens, et al. (2021) also noted some concerns, however, about specific probe types (i.e., asking about intentionality or depth of mind wandering). For example, one common finding is that “no-go” accuracy in the SART is worse on the trials before TUT reports compared to on-task reports. Kane, Smeekens, et al. (2021) replicated this finding but found that it was more pronounced for probes asking about the intentionality or depth of mind wandering (versus its content), suggesting that these TUT reports might be especially influenced by reactivity to performance. Thus, it's possible that not all TUT reports equally reflect sustained attention failures or are equally affected by sources of measurement error.

Just as the field should not rely solely on any one objective measure of attention consistency in any one task, it also should not rely solely on any one self-report measure from any single task. Rather, the variance that is common across subjective indicators from multiple contexts and tasks (and perhaps across different types of thought-probes) should yield a more accurate sustained attention measure. And, further, as argued previously, variance that is common across multiple subjective indicators and multiple objective indicators should provide an optimally construct valid assessment of general sustained attention ability.

Correlations between Objective and Subjective Measures

Objective and subjective indicators provide starkly different approaches to measuring sustained attention abilities. If they are, nonetheless, both influenced by a general sustained attention ability, then then they should be consistently correlated. Indeed, at the between-person level of analysis, latent variable correlations between RT-variability and TUT-rate factors typically range from .30–.50 (Kane et al., 2016; Unsworth, 2015; Unsworth et al., 2021; Welhaf et al., 2020; for similar RT variability–TUT correlations in single experimental tasks, see Löffler et al., 2021; Stawarczyk et al., 2014; Yamashita et al., 2021). These factors are thus only *moderately* correlated: Subjects who report more off-task thoughts also show more inconsistent responding in simple attention and RT tasks, but the association between these measures is not strong. Despite this moderate correlation, the shared variance between subjective and objective measures of sustained attention should provide the most construct valid measure of sustained attention ability.

Indeed, we further argue that *because* of this moderate correlation, using the shared objective–subjective variance to measure attention consistency is especially important. These factors are not redundant—objective and subjective indicators cannot simply be used interchangeably. Performance and self-report measures may not only capture different dimensions or depths of sustained attention failures (e.g., Cheyne et al., 2009), but these different approaches are also subject to different non-sustained attention confounds, which uniquely influence their measurement. Relying on only one type of indicator as *the* measure of attention consistency in a study may lead to incorrect conclusions about how other theoretically relevant constructs correlate with sustained attention ability. Instead, using what is common between these measures should be a more construct valid way to measure sustained attention than using either in isolation: Researchers should assess the covariation between performance and self-report measures of attention consistency not *despite* their moderate correlation, but *because of* it.

At the within-person level of analysis, one would also expect poorer performance (i.e., more errors) and greater RT variability in the moments preceding TUT reports compared to on-task reports. Indeed, commission errors on the SART, where subjects erroneously press a key on “no-go” trials, are more likely to occur prior to TUTs than to on-task reports (e.g., Kane, Smeekens et al., 2021; McVay & Kane, 2009, 2012a; Smallwood & Schooler, 2006; Stawarczyk et al., 2011). These findings are potentially supportive of construct validity, but also ambiguous, because TUT reports that follow errors might be reactively biased by subjects’ knowledge of their performance, as noted earlier (Schubert et al., 2019). Because subjects are likely less aware of their RT variability on the trials leading up to thought probes, however, examining RTs that precede thought reports should provide a less biased assessment of behavioral correlates of TUT experiences. In fact, RTs preceding TUTs are more variable than those preceding on-task reports (Bastian & Sackur, 2013; Kane, Smeekens et al., 2021; Seli, Carriere et al., 2013).

Summary of Sustained Attention Measurement

Objective measures allow researchers to examine subtle fluctuations in attention (i.e., RT variability) or instances of attentional lapses that produce inappropriate responding (i.e., commission errors, omission errors, and flat spots), of which subjects are not necessarily consciously aware. In contrast, subjective measures (i.e., namely TUT reports) capture instances of sustained attention failures that are so apparent to subjects that they can readily report on them. These two types of measures frequently correlate with each other at the between- and within-subject level, suggesting they may both be impacted by a common underlying ability. At the same time, these correlations are of only moderate strength because each may capture different degrees of sustained attention failure, and each has independent limitations and sources of error. Under these conditions, then, the combination of these two assessment types should best reflect the construct of sustained attention, independent of those sources of error. We therefore argue that the optimal way of capturing people's general sustained attention abilities is to quantify the individual-differences variance that is common to both objective and subjective indicators.

Evidence for the Construct Validity of Sustained Attention Measures

Considerable research has examined associations that sustained-attention indicators have with other theoretically relevant variables, taking a "nomothetic span" approach to construct validation (Cronbach & Meehl, 1955; Embretson, 1983). Studying these theoretically relevant variables, as part of the nomological network, provides evidence for the convergent and discriminant validity of attention consistency measures.

Correlations with Executive Attention Ability

Executive Attention theory (e.g., Burgoyne & Engle, 2020; Engle & Kane, 2004) argues that working memory capacity (WMC) broadly predicts performance on higher-order tasks (e.g., language comprehension, reasoning) because it reflects, in part, how effectively people can maintain ready access to goal-relevant information in the face of distraction or interference. According to this view, lower-WMC subjects have poorer goal-maintenance ability, and so they should show more frequent attention lapses compared to higher-WMC subjects. Indeed, WMC measures (and related attention-control measures, such as Stroop and antisaccade performance) correlate moderately with objective sustained attention measures, like RT variability, across a variety of tasks and measurement approaches (Kane et al., 2016; McVay & Kane, 2009, 2012a; Schmiedek et al., 2007; Schweizer & Moosbrugger, 2004; Unsworth, 2015; Unsworth et al., 2010; 2012; 2021). Higher-WMC subjects are less variable in performance than are lower-WMC subjects.

WMC and attention-control abilities are also frequently negatively associated with TUT rates in lab tasks (e.g., McVay & Kane, 2012b; Meier, 2019; Rummel & Boywitt, 2014; Unsworth & McMillan, 2017; Unsworth et al., 2012, 2021) and in certain everyday-life contexts (Kane et al., 2007, 2017). In latent-variable studies that use multiple tasks to test construct-level correlations, the association between TUT rate and WMC often yields *r* =|.20–.30|, whereas associations between TUT rate and attention-control performance is stronger, *r* =|.35–.45|. WMC and attention-control abilities reliably predict sustained-attention ability, whether derived from objective or subjective measures. At the same time, WMC and attention-control ability appear to predict these different indicators of sustained attention to differing degrees. Thus, the field must examine how these constructs correlate with a variable reflecting the shared variance between objective and subjective indicators to better understand their relationships with sustained attention.

Correlations with Processing Speed

An important consideration with any RT measure, including RT variability, is that it may capture individual differences in general processing speed rather than the cognitive abilities of interest. Regarding sustained attention measurement, individuals may be more prone to extremely long or variable RTs because they have an overall slower processing rate. RT variability and speed can be highly collinear across experimental conditions and tasks (r ~ .90; Jensen, 1987a, 1992; Wagenmakers & Brown, 2007). As well, both mean RT and RT variability are influenced by long RTs that might reflect attentional lapses. Thus, it is possible that the apparent inability to sustain attention might simply be due to poor processing speed.

Measures of processing speed and objective indicators of attention consistency correlate substantially. For example, Unsworth et al. (2021) operationalized processing speed as subjects' fastest 20% of trials within three attention tasks. A latent variable of objective attention consistency indicators (PVT lapses, mouse-tracking flat spots, SART CoV) correlated with the speed latent variable (r = .47): Subjects with slower processing also exhibited poorer sustained attention. However, speed was also highly correlated with other cognitive ability measures like attention control, so structural equation models tested whether speed predicted any unique variance in objective attention consistency measures. It did not: After accounting for shared variance with other measures (like attention control) processing speed did not significantly predict the objective sustained attention latent variable. Attention control (but not WMC) predicted unique variance in objective sustained attention after accounting for shared variance among all the predictor constructs, suggesting that attention control, and not speed of processing or WMC, might be critical in explaining variation in attention lapses.

Additionally, latent-variable studies provide mixed evidence regarding correlations between processing speed and self-report indicators of attention consistency. Unsworth et al. (2021) found that processing speed measures were weakly associated with TUT rates (*r* = .24), whereas Welhaf et al. (2020) did not find a significant association between subjects' shortest RTs and TUT rates (*r* = .09). Given that processing speed is more strongly related to objective than subjective indicators, measuring sustained attention as a latent variable reflecting the covariation between objective and subjective measures should best distinguish sustained attention ability from processing speed.

Correlations with Cognitive Self-Report Variables

People who are more prone to cognitive failures in daily life should also show poorer sustained attention in lab tasks. Indeed, scores on retrospective self-report measures of everyday attention failures like the Cognitive Failures Questionnaire (CFQ; Broadbent et al., 1982) and Attention-Related Cognitive Errors Scale (ARCES; Cheyne et al., 2006) correlate positively with both performance measures of sustained attention (e.g., Cheyne et al., 2006; McVay & Kane, 2009; Smilek et al., 2010, Steinborn et al., 2016) and TUT rates (e.g., McVay & Kane, 2009, 2013; Smallwood et al., 2004; Unsworth et al., 2021). In general, such self-reported cognitive failures appear to be slightly more strongly correlated with subjective measures of attention consistency (*r*s ~ .20) than with objective measures (*r*s ~ .15). Thus, measuring sustained attention as the individual-differences overlap in objective and subjective measures should provide a better estimate of the correlation between sustained attention ability and everyday cognitive failures.

Correlations with Contextual-State Variables

People who are more motivated or interested in a task should exhibit better sustained attention in that task; being more willing to expend effort to focus on the task, or finding it rewarding to do so, should allow them to perform more optimally. Indeed, objective measures of attention consistency correlate with post-task self-report measures of motivation, interest, and arousal, with greater RT variability associated with lower state reports ($r = -.65$ to $-.30$; Robison & Unsworth, 2018; Seli et al., 2015; Unsworth et al., 2021). As well, subjects who are more motivated, interested, or aroused report fewer TUTs across a variety of tasks and activities ($r = -.60$ to $-.27$; Brosowsky et al., 2020; Hollis & Was, 2016; Kane, Carruth et al., 2021; Robison & Unsworth, 2015; 2018; Unsworth & McMillan, 2013, Unsworth et al, 2021; but see Rummel et al., 2021). While the ranges of these contextual variable correlations are quite similar, correlations with TUTs are often stronger than those with objective measures, perhaps due to similar self-report biases at play. Thus, by measuring sustained attention as the overlap in objective and subjective measures, we should better assess the relation between sustained attention and these contextual factors.

Correlations with Personality Traits

Individual differences in certain personality traits may affect or reflect sustained attention ability. People who experience high levels of anxiety (i.e., neuroticism), for example, may show worse sustained attention ability due to ruminative tendencies or intrusive worries. People who are more willing to work toward goals or follow task instructions (i.e., high in conscientiousness or agreeableness), in contrast, may exhibit better sustained attention, perhaps especially in mundane tasks.

In terms of objective indicators of attention consistency, individuals who are high in neuroticism tend to show more variable RTs and more frequent lapses in simple tasks (Klein & Robinson, 2019; Robinson & Tamir, 2005; Unsworth et al., 2021). Other “big-5” personality factors, however, do not appear to be related to performance measures (e.g., Unsworth et al. 2021). In terms of subjective indicators, correlations are less consistent. Students high in neuroticism frequently report more TUTs in the lab (Jackson et al., 2013; Kane, Gross et al., 2017; Robison et al., 2017; Unsworth et al., 2021), whereas students who are more goal-oriented (i.e., high in conscientiousness) report fewer TUTs in some studies (Jackson & Balota, 2012; Robison et al., 2020; Unsworth et al., 2021), but not in others (Jackson et al., 2013; Kane, Gross et al., 2017). Likewise, students who are more likely to comply with task instructions (i.e., high in agreeableness) reported fewer TUTs in one study (Unsworth et al. 2021), but not in another (Kane, Gross et al., 2017). Finally, openness to experience often fails to predict TUT rates in the lab (Smeekens & Kane, 2016; Unsworth et al. 2021), but does predict TUTs in daily life (Kane, Gross et al., 2017). Thus, neuroticism, which is unique in consistently correlating with both objective and subjective measures (in the lab, at least), might be related to a general sustained attention ability representing their shared variance.

Nomothetic Span Summary

Correlational studies provide evidence of convergent validity of attention consistency measures. Constructs that should predict sustained attention ability do so: People with (a) better cognitive abilities, such as WMC and attention control, (b) higher motivation and interest in performing well, and (c) lower dispositional tendences to experience sustained attention failures, all show less variable responding and lower TUT rates in simple tasks.

Although it is—and should be—rare to find constructs with *no* association (i.e., a null correlation) with attention consistency, given how fundamental sustained attention should be to so many domains of performance and experience, *relative* differences in correlation magnitudes can provide evidence for discriminant validity. First, attention control ability (typically measured with response-competition or interference-control tasks) frequently correlates more strongly with attention consistency measures (RT variability and TUT rate) than does WMC; one possible explanation for this difference is that WMC tasks are influenced by processes like memory storage or strategy choices that are less relevant to attention regulation. Second, RT variability indicators do not share unique variance with processing speed after accounting for other cognitive abilities, and TUT rates correlate weakly (if at all) with processing speed, suggesting that sustained attention is not simply a speed factor. Third, and lastly, some personality traits, such as agreeableness, conscientiousness, extraversion, and openness, are not correlated with objective attention consistency measures, but are weakly (and inconsistently) correlated with subjective measures, suggesting they may not be related to general sustained attention ability. Neuroticism and self-reported cognitive failures, however, correlate with both types of attention consistency measures, suggesting they may also be associated with general sustained attention ability.

As previously argued, the modest correlations between objective and subjective indications of attention consistency indicates a need to use the covariation between these indicators as a more construct valid approach to assessing attention consistency than either measure on their own. Our perspective follows from how very different behavioral performance measures and self-report measures are, with each possibly reflecting different degrees of attentional disengagement (à la Cheyne et al., 2009) and each affected by unique sources of measurement error, both of which drive down their correlation.

A competing argument, however, is that objective and subjective indicators do not correlate strongly enough to indicate convergent validity and so they must instead reflect two *different* constructs (i.e., they provide discriminant validity evidence for one another). At least implicit to this argument is that only one of these indicator types is a construct-valid measure of attention consistency. We think this argument is not compelling. First, objective and subjective indicators of sustained attention consistently correlate with each other at both the within- and the between-subject level, suggesting that both reflect, at least partially, a failure to sustain attention. Further, each type of indicator correlates with other nomological network constructs in ways in which theory would predict. For example, people with higher WMC and attention control abilities show better scores on objective and subjective attention consistency measures. Likewise, certain dispositional characteristics (e.g., agreeableness) show reliable null associations with *both* indicator types, suggesting that constructs that should not correlate with sustained attention do not, regardless of the indicator used. Proposing that these two types of indicators reflect two different constructs implies that one of these two literatures is simply wrong about the body of relevant evidence and their claims that their measures (i.e., objective performance measures or subjective self-reports) reflect the ability to sustain attention.

Goals of the Current Studies

Many studies have investigated the nomological network of sustained attention, or how objective and subjective measures correlate with each other or with theoretically relevant variables. If these two forms of measurement are both presumed to reflect variation in sustained attention, albeit imperfectly, then their covariation should best reflect the general ability to sustain attention: Each indicator of sustained attention may reflect different degrees of attention failure and each has its own source of measurement error that may impact attention consistency measurement if used on its own, but what they measure in common should reflect the sustained attention construct especially well.

The present studies' goals were (1) to assess whether there indeed exists a general sustained attention construct that reflects the individual-differences overlap in objective and subjective measures and, if so, (2) to examine how theoretically relevant constructs like cognitive ability (e.g., WMC, attention control, and processing speed), contextual-state variables (e.g., task-specific motivation and alertness), and dispositional characteristics (e.g., everyday cognitive failures and personality traits) correlate with this common sustained attention factor. We reanalyzed data from two large latent-variable studies that had (a) multiple tasks with objective performance measures of attention consistency and (b) probed self-report assessments of TUTs within multiple tasks (Kane et al., 2016; Unsworth et al., 2021). These datasets allowed us to use confirmatory models to test whether there was enough variance shared between the objective and subjective measures to model a general factor of sustained attention, and to model influences unique to both objective and subjective measures in the form of bifactor models.

Study 1

Methods

We analyzed data from Unsworth et al. (2021), a study on individual differences in attention lapses. Details of our preregistration are available on the Open Science Framework (https://osf.io/xeu63/).

Subjects

Three hundred fifty-eight subjects from the University of Oregon were individually tested in a 2-hour session.

Tasks and Materials

Objective Sustained Attention Indicators. For each objective indicator task, we present multiple dependent measures that theoretically should reflect variation in sustained attention; we describe our preregistered procedures for selecting among these dependent variables for analysis below, under "Objective Indicator Variable Selection." For each task, we first list our *a priori* measure, while also considering different measurement approaches and dependent variables across tasks (i.e., not choosing RTsd as the primary measure for all tasks). We set these *a priori* measures as the primary indicator for each task and assessed reliability, distribution characteristics, and bivariate correlations of the secondary measures against them; that is, we planned to use only the *a priori* measure for each task if all other measures were redundant with it. We preregistered that any measures correlated ≥ .70 would be considered redundant and thus would only retain the *a priori* measure for each task. Non-redundant measures would be retained and included in structural models, as they may reflect different types or degrees of sustained attention failures.

Psychomotor Vigilance Task (PVT). Subjects were presented with a row of zeros onscreen. After an unpredictable period (from 2–10 s), the zeros began counting-up in 17 ms intervals. The goal of the task was to press the spacebar as quickly as possible to stop the numbers. The RT was displayed for 1 s to provide feedback. The task lasted for 10 min (roughly 75 trials). The potential dependent variables derived from this task will be average RT of the slowest 20% of trials, number of lapses (RTs > 500 ms), intra-individual standard deviation of all

RTs (RTsd), intra-individual median absolute deviation of all RTs (RTmad), and the τ estimate from an ex-Gaussian model of all RTs.

Semantic Sustained Attention to Response Task (SART). Subjects were instructed to respond quickly by pressing the spacebar to frequently presented non-target stimuli from one category (animals, presented on 89% of trials) while withholding responses to infrequent target stimuli from a different category (vegetables, presented on 11% of trials). Stimuli were presented for 300 ms followed by a 900 ms mask. There were 315 trials, 35 of which were no-go targets. The potential dependent variables derived from this task will be intra-individual RTsd to correct "go" trials, intra-individual RTmad to correct "go" trials, omission errors on "go" trials, average RT of the slowest 20% of correct "go" trials, RMSSD to correct "go" trials, the τ estimate from an ex-Gaussian model using correct "go" trials, and fastest 20% of correct "go" trials.[6]

Choice RT (CRT). Subjects responded as quickly as possible to a stimulus (a white cross) in one of four horizontally spaced locations onscreen. The cross appeared after a random interval (300–550 ms in 50 ms increments) and could not appear in the same location on consecutive trials. Subjects indicated the location of the cross by pressing one of four keys on the keyboard (F, G, H, J) mapped to the four locations. Subjects completed 15 practice trials and 210 real trials. The potential dependent variables derived from this task will be the τ estimate from an ex-Gaussian model of correct trials, the number of "blocks," defined as RTs that were twice each

[6] The SART presents a unique case for assessing lapses of attention. Namely, because of the high frequency of "go" responses that build up habitual, mindless responding, extremely fast responses might also be indicative of lapses of sustained attention. Indeed, prior research has found that TUT rates in a SART are significantly correlated with the fastest 20% of responses, in addition to the slowest; that is, individuals who mind wander more in that SART also have shorter "short" RTs along with longer "long" RTs (McVay & Kane, 2012a; Welhaf et al., 2020). Thus, the fastest 20% of SART RTs has been included as a possible indicator of sustained attention ability.

individual's mean RT (Bills, 1931a, 1931b, 1935; see also Bertleson & Joffe, 1963), intra-individual RTsd to correct trials, intra-individual RTmad to correct trials, average RT of slowest 20% of correct trials, and RMSSD to correct trials.

Continuous Tracking. Subjects saw a small black circle moving against a gray background onscreen. The goal was to follow the black circle as closely as possible with the mouse cursor. Each block began with a screen saying, "Please focus on the dot," for 3 s. The circle moved in a pseudorandom fashion within a centered 400 × 440 pixel region. The circle moved at a constant speed in vertical, horizontal, or diagonal directions. Subjects completed a 30 s practice block, followed by (in a random order) one 30 s and one 120 s block, and two 60 and 90 s blocks. The potential dependent variables derived from this task will be tracking distance variability (calculated as a moving window average tracking error in pixels of 5 trials), the number of flat spots (instances where subjects stopped responding for at least 1.5 s), overall average tracking error (i.e., the distance between the cursor and the circle in pixels on each trial across each block), and intra-individual standard deviation in tracking error (calculated as the standard deviation of the distance, in pixels, between the circle location and the cursor location). Tracking distance variability, overall average tracking error, and intra-individual standard deviation in tracking error will be calculated at the block level first and then averaged for each subject to account for tracking duration differences of each block.[7]

[7] Unsworth et al. (2021) also measured lapses in a whole report working memory task, as the number of trials where subjects recalled ≤ 1 item. We did not include this lapses measure because some may have reflected working memory failures. Additionally, we did not include self-reports of sleep quality, boredom proneness, or mindfulness, the latter due to multicollinearity problems (reported in Unsworth et al., 2021).

Subjective Sustained Attention Indicators. Subjects responded to thought probes in four tasks: the PVT (15 probes), the SART (21 probes), a working memory task (8 probes), and Stroop task (12 probes). The probes asked subjects to classify their immediately preceding thoughts into one of five categories. Subjects reported via keypress whether their conscious experience was: (1) *I am totally focused on the current task, (2) I am thinking about my performance on the task, (3) I am distracted by sights/sounds/physical sensations, (4) I am daydreaming/my mind is wandering about things unrelated to the task,* or *(5) My mind is blank.* Consistent with Unsworth et al. (2021), we operationalized TUTs as the proportion of responses 3–5.

Working Memory Capacity (WMC) Tasks. Subjects completed three complex span tasks of WMC. For each complex span task, subjects completed three practice stages: the first provided practice in memorizing small sets of the memoranda for each task (e.g., letters or grid locations); the second practice was for processing-only (e.g., math equations, symmetry decisions, sentence comprehension). RTs were recorded during this processing only practice for each subject. During the real trials, if a processing decision was not made within 2.5 SDs of the processing-only mean, that trial was counted as a processing error; the third practice consisted of both the memory and processing task combined (as in the real trials).

Operation Span. Subjects verified whether math operations were true or false while trying to remember a set of letters. After each math operation, a letter was presented for 1 s, and then the next math operation was presented. At the end of the set, subjects were asked to recall the letters from the set by clicking the letters onscreen in the presented serial order. Subjects were granted credit only if the item letters were recalled in the correct serial position. Set sizes ranged from 3 to 7 items and each set size was presented twice (for a max score of 50). Higher scores reflected better recall.

Symmetry Span. Subjects verified whether an abstract image presented in an 8 × 8 matrix was symmetrical along the vertical axis. Following the verification, they were presented with a red square for 650 ms in a 4 × 4 grid for memory. At the end of each set, subjects recalled the location of each red square presented; subjects earned credit for items recalled in correct serial position. Set sizes ranged from 2 to 5 items and each set size was presented twice (for a max score of 28). Higher scores reflected better recall.

Reading Span. Subjects decided whether sentences made sense or not while remembering a set of letters. Sentences were made nonsensical by altering one word. After deciding whether a sentence made sense, subjects saw the to-be-remembered letter for 1 s. After the final letter of the set, subjects recalled the set; subjects earned credit for items recalled in correct serial position. Set sizes ranged from 3 to 7 items, and each set size was presented twice (for a max score of 50). Higher scores reflected better recall.

Attention Control Tasks. Subjects completed three tasks measuring attention control.

Antisaccade. Subjects completed 60 trials in which they were told to direct their focus away from a flashing cue (a white flashing "=") to identify a masked letter (B, P, or R) presented briefly to the opposite side of the screen. The flashing cue and target letter location were 12.7 cm to the left or right of central fixation. The target stimuli appeared onscreen for 100 ms and then were masked (by an *H* for 50 ms then an *8*, which remained onscreen until response). Subjects pressed the corresponding key on the numeric keyboard (4, 5, and 6 were used for B, P, and R, respectively) to identify the target letter. Before completing the antisaccade trials, subjects completed 10 response-mapping trials and 10 prosaccade trials (where the flashing cue and letter appeared on the same side). The dependent variable was the proportion of correct antisaccade trials.

Cued Visual Search. Subjects decided whether a target F located in a 5 × 5 array of 25 letters (comprised of distractors including forward and backward Es and rotated Ts) was either normal facing (by pressing the "/" key) or mirror-reversed (by pressing the "Z" key). Subjects first completed 8 response-mapping trials. On each trial, subjects received a central arrow cue (500 ms) indicating which two or four possible locations (of eight) the target F could appear in. Following the cue, a blank screen (50 ms) appeared before the 5 × 5 grid of 25 possible locations appeared as dots for 1500 ms, followed by another 50 ms blank screen. Finally, the array of 25 letters was shown, at which time subjects responded to the target F (the array was shown until response, but no longer than 4000 ms). Other Fs also appeared in uncued, nontarget locations as distractors, and so to respond correctly, subjects must selectively maintain focus on the cued locations. Subjects completed 8 practice trials followed by 80 scored trials. Cue type, target direction and location were all randomly and equally presented during the scored trial block. The dependent measure was mean RT for correct responses.

Stroop. Subjects were presented with a color word (red, green, or, blue) in one of three different font colors (red, green, or blue). The goal of the task was to indicate the font color as quickly and accurately as possible via key press (1 = red, 2 = green, 3 = blue). Subjects completed 15 response-mapping practice trials and 6 practice trials of the real task. Subjects then completed 100 scored trials (67 congruent trials [e.g., the word "red" was presented in red font color]; 33 incongruent trials [e.g., the word "green" presented in blue font color]). The dependent measure was the Stroop RT effect (correct incongruent RT – correct congruent RT).

Processing Speed. As in Unsworth et al. (2021), we assessed processing speed in three tasks where RT was one of the primary measures recorded (the PVT, Stroop, and CRT). In these tasks, RTs were ranked from fastest to slowest and the fastest 20% of trials were used as indicators. Here, however, in addition to using RT for the fastest 20% of trials, processing speed will also be calculated using the μ parameter from the ex-Gaussian model from the SART, PVT, Stroop, and CRT (reflecting the mean of the Gaussian component). Additionally, median RT of the 10 prosaccade practice trials in the antisaccade task will also be used as a measure of processing speed. Selection of processing speed measures from each task for analyses followed a similar approach to the selection of objective sustained attention measures (see "Objective Sustained Attention Dependent Variables Selection Procedure").

Cognitive Failures Questionnaire – Memory and Attention Lapses (CFQ-MAL). Subjects responded to 40 questionnaire items about their everyday memory and attention lapses. Subjects indicated via keypress that they experienced such failures on the following scale: 1) *never,* 2) *rarely*, 3) *once in a while*, 4) *often*, 5) *very often*. The dependent variable was an item sum score.

Non-Cognitive Predictor Measures. Subjects completed the following self-report scales.

Motivation and Alertness. Following the completion of four tasks (PVT, CRT, continuous tracking, and antisaccade) subjects responded to one question each about their motivation and alertness on a 1–6 scale (higher scores meaning more motivated or alert). Specifically, they were asked "How motivated were you to perform well on the task?" and "How alert do you feel right now?"

Big Five Inventory (BFI). Subjects completed a 44-item version of the Big Five personality inventory. Extraversion was assessed by eight items, agreeableness by nine, conscientiousness by nine, neuroticism by eight, and openness by 10. Each item asked the subject to respond based on how well it described them using a 5-point scale (1 = disagree strongly, 5 agree strongly). The dependent variable was the average rating across items for each factor.

Procedures

After providing informed consent, subjects completed the cognitive battery in the following order: operation span, symmetry span, reading span, antisaccade, cued visual search, PVT, Stroop, SART, choice RT, continuous tracking, and whole report visual WM. Following completion of the cognitive tasks, subjects completed questionnaire measures in the following order: BFI, Boredom Proneness Scale, CFQ-MAL, Mindful Attention Awareness Scale, and self-reported sleep quality and quantity.

Results

Below we report the results of our preregistered analyses and note where we deviated from the preregistered plan. Data and Rmarkdown files for all analyses are available on the Open Science Framework (https://osf.io/xeu63/).

Data Analysis Exclusions

Consistent with Unsworth et al. (2021), we excluded the same subject data from the psychomotor vigilance (n = 2), Stroop (n = 1), and Choice RT tasks (n = 1) for having extremely long M RTs in each task (in the PVT, one subject had M RT > 1200 ms and one had M RT > 18 s; for Stroop, the subject had M RT > 2400 ms; for the Choice RT task, the subject had M RT > 1200 ms). Also following Unsworth et al. (2021), we dropped 16 subjects' data from the SART for having > 50% omission errors. Prior to calculating any of the DVs for the current study, we calculated "go" trial accuracy for the remaining subjects and identified 3 who had "go" trial accuracy < 70%. We therefore deviated from Unsworth et al. (2021) and *deviated from our preregistration* by excluding SART data from these subjects, too, as such low accuracy might indicate a failure to understand or comply with task instructions, rather than failures of sustained attention. As preregistered (but deviating from Unsworth et al., 2021), we also dropped subjects' TUT data from a task if their performance data were dropped from that probed task. Finally, and although *not preregistered* (and not specified in Unsworth et al., 2021), we also dropped Choice RT task data from two subjects with 0% accuracy, indicating they did not follow or understand task instructions.

RT Cleaning Procedures

In tasks where RT was the primary measure of interest (e.g., objective attention consistency tasks and processing speed tasks), we implemented a preregistered multistep procedure for trial-level cleaning. First, we identified and removed RTs for error and post-error trials (and, in tasks that included thought probes, post-probe trials). Next, we removed RTs for trials that were likely anticipations (i.e., RTs < 200 ms). From the remaining trials, we next calculated for each subject, in each task, a value equal to their Median RT + 3*IQR. Any trials outside of this value were replaced with this value. Details on the average number of trials cleaned in each task can be found in Supplemental Table S6.

Finally, we calculated the number of usable trials each subject had following our RT cleaning protocol. As preregistered, we dropped task data for subjects who did not have at least 40 trials and thus could not reliably contribute to our primary measures of interest. This resulted in 6 additional subjects' data being dropped for the PVT only.

Selection of Objective and Processing Speed Indicators

After trial-level cleaning, we calculated all possible DVs of interest for each task. As preregistered, for each DV, we used the Median + 3*IQR rule that we had applied to trial-level data to censor outlying subjects (replacing outlying subjects' data with a value equal to the Median + 3*IQR for each DV). Supplemental Tables S7 and S8 presents the descriptive statistics for each possible DV, for each task, as well as the number of subjects censored for each measure. Note that many of the measures originally used in Unsworth et al. (2021) had potentially problematic skewness and kurtosis, but after our trial- and subject-level cleaning procedures, these values were acceptable for all potential dependent measures.

As preregistered, our first step in selecting which DVs from each task to include in our structural models focused on examining the univariate distributions for possible issues of skewness and kurtosis. Per the guidelines suggested by Kline (2011), problematic skew was identified as > 3.0 and problematic kurtosis was identified as > 10.0. No variables were removed from consideration for problematic distributions.

We next examined the reliability of the measures. Consistent with Unsworth et al. (2021), we calculated split-half reliability for each measure where applicable. (Note that in the PVT and Stroop, splitting the task resulted in < 40 trials in each grouping, which prohibited reliable estimation of ex-Gaussian models; we thus do not report reliability for PVT τ, PVT μ, or Stroop μ). We preregistered that any measures with poor split-half reliability (< .50) would not be considered for models, but no variables needed to be removed from consideration for poor reliability.

We next examined within-task bivariate correlations to see whether any measure combinations provided non-redundant information with the *a priori* measure (preregistered criterion for redundancy: $r \geq .70$). As seen in Supplemental Table S9, many proposed measures were redundant. For the PVT, CRT, and Continuous Tracking Task, we retained only the *a priori* measure (*M* RT of the slowest 20% for the PVT, τ for the CRT, and the Tracking Variability measure for Continuous Tracking). For the SART, however, correlations suggested that several measures reflected differing degrees or types of sustained attention failures (Cheyne et al., 2009; Unsworth et al., 2021). Thus, by our selection criteria, we would retain not only SART RTsd (*a priori*), but also τ, *M* RT from the fastest 20% of trials, and Omissions. We had not expected to find evidence for four nonredundant measures from the SART, particularly while finding no additional non-*a priori* measures from the other tasks. To avoid oversaturating the objective sustained attention latent variable with SART measures—with four SART indicators but only one indicator each from the other tasks—we retained only SART RTsd, τ, and Omissions for all structural models (*deviating from preregistration*).[8]

For processing speed, our proposed measures showed good split-half reliability and distributional characteristics. To diversify speed-factor indicators (and prevent overlap with other DVs for other factors), we selected the following: μ from the PVT, *M* RT of the fastest 20% of trials for the CRT and the Stroop, and median RT of the Prosaccade practice trials; *this deviated from the preregistration*, which indicated using *M* RT of the fastest 20% of trials for the PVT. Supplemental Table S10 provides the bivariate correlations among the possible speed of processing measures for each task. (Note that we also *deviated from preregistration* by not including SART μ, given the large number of SART indicators we included as objective

[8] We dropped the fastest 20% variable rather than other SART measures as it correlated only weakly with RTsd and Omissions (*r*s < .20). RTsd, τ, and Omissions all correlated with each other at $r \geq .25$.

sustained attention indicators and given its poor zero-order correlations with the other processing speed measures.)

Multivariate Outliers

As preregistered, once we established our primary indicators, we checked for multivariate outliers in the final dataset. To do this, we used the *Routliers* package (Leys et al., 2019) to calculate Mahalanobis distance for each observation. This analysis indicated there were 11 multivariate outliers in the dataset (~3% of subjects). These subjects' data were removed case-wise before conducting structural modeling.

Descriptive Statistics and Correlations for Final Dataset

Table 7 provides descriptive statistics for the final dataset; Table 8 reports bivariate correlations among all measures of interest. Consistent with Unsworth et al. (2021), measures from the same putative construct (e.g., WMC, Attention Control, Motivation) all correlated more strongly with each other than with measures of other constructs. Importantly, our newly selected objective attention consistency indicators also showed evidence of convergent validity (median $|r| = .30$), suggesting that subjects who showed more variable responding in one task also tended to do so in other tasks.

Table 7. Descriptive statistics for Study 1 measures

Construct/Measure	Mean	SD	Min	Max	Skew	Kurtosis	N
Objective Sustained Attention							
PVT Bin 5	455.01	94.34	307.08	789.01	1.35	2.34	333
SART RTsd	132.35	48.48	44.07	299.83	1.09	1.80	322
SART Omissions	18.48	14.23	0.00	70.00	1.36	2.17	322
SART Tau	100.64	67.60	0.00	328.59	0.95	1.23	322
CRT Tau	90.06	36.29	26.32	229.52	1.22	2.11	335
Continuous Tracking Variability	1.14	0.41	0.24	2.56	0.70	0.62	322
Subjective Sustained Attention							
PVT TUTs	0.43	0.29	0.00	1.00	0.23	-0.92	333
SART TUTs	0.44	0.33	0.00	1.00	0.24	-1.19	322
WRWM TUTs	0.53	0.37	0.00	1.00	-0.12	-1.43	271
Stroop TUTs	0.21	0.28	0.00	1.00	1.42	1.01	341
Working Memory Capacity							
OPERSPAN	38.00	8.04	10.00	50.00	-0.70	0.11	345
READSPAN	37.41	8.48	1.00	50.00	-1.04	1.36	346
SYMSPAN	18.85	5.19	2.00	28.00	-0.48	-0.12	346
Attention Control							
Antisaccade Accuracy	0.60	0.15	0.25	0.93	0.03	-0.60	337
Cued Visual Search RT	1276.70	290.03	596.24	2316.29	0.64	0.42	344
Stroop RT	147.81	97.70	-224.04	509.78	0.60	1.13	341
Processing Speed							
PVT Mu	286.22	28.01	227.26	380.01	0.61	0.09	333
CRT Bin 1	293.81	35.52	227.19	434.92	1.03	1.48	335
Stroop Bin 1	437.86	63.46	304.45	675.58	0.93	1.35	340
Prosaccade M RT	703.05	240.33	305.50	1628.00	1.04	1.54	311
Alertness							
PVT	3.30	1.28	1.00	6.00	0.13	-0.55	341
CRT	3.31	1.33	1.00	6.00	0.08	-0.71	338
Continuous Tracking	2.31	1.43	1.00	6.00	0.79	-0.48	314
Antisaccade	3.64	1.29	1.00	6.00	-0.01	-0.78	337
Motivation							
PVT	4.00	1.31	1.00	6.00	-0.40	-0.53	341
CRT	4.03	1.37	1.00	6.00	-0.54	-0.35	338
Continuous Tracking	2.70	1.60	1.00	6.00	0.39	-1.16	314
Antisaccade	4.00	1.35	1.00	6.00	-0.39	-0.61	337
Dispositional Measures							
Openness	3.57	0.578	1.80	4.90	-0.12	-0.10	274
Conscientiousness	3.61	0.63	1.22	4.89	-0.49	0.47	274
Extraversion	3.22	0.87	1.14	4.57	-0.38	-0.80	274
Agreeableness	3.92	0.64	1.67	5.00	-0.68	0.57	274
Neuroticism	3.16	0.86	1.00	5.00	-0.00	-0.76	274
Cognitive Failures	111.01	26.16	52.00	191.00	0.26	-0.17	274

Note. PVT = Psychomotor Vigilance Task. SART = Sustained Attention to Response Task. CRT = Choice Reaction Time Task. WRWM = Whole-Report Working Memory Task. TUTs = Rate of Task-Unrelated Thoughts in specified task. OPERSPAN = Operation Span. READSPAN = Reading Span. SYMSPAN = Symmetry Span. Bin 5 = Mean RT of Slowest 20% of correct trials. RTsd = intra-individual RT variability. Bin 1 = Mean RT of Fastest 20% of correct trials.

Measurement Models of Sustained Attention

As preregistered, our first set of analyses attempted to simply replicate the latent variable correlation between objective and subjective indicators of sustained attention reported by Unsworth et al. (2021). We first tested a 2-factor sustained attention model with separate latent variables for objective (i.e., RT variability and omissions) and subjective (i.e., TUT reports) indicators; these latent variables were allowed to correlate. As seen in Table 9, the model adequately fit the data. *Although not preregistered*, we included residual correlations among any performance and TUT indicators from the same task (e.g., PVT Slowest 20% with PVT TUTs); a model without these residuals did not adequately fit the data and we retained these residual correlations for all subsequent models. Our measures and analysis conceptually replicated the lapse–TUT correlation in Unsworth et al. (2021), although this relationship was slightly weaker here (r = .32 vs .44 in the original study; see Table 10 for factor loadings).[9] Again, this *moderate*—but not strong—correlation confirms that these two indicator types of sustained attention are not redundant. Instead, as we've argued, each indicator type may reflect different degrees of disengagement and each is likely influenced by non-sustained attention processes that are unique to that measurement type, so modeling the shared variance among the indicators may provide a more construct valid measure of sustained attention than either objective or subjective measures alone.

[9] *Although not preregistered*, we also tested whether a single factor Sustained Attention model fit the data, by specifying a model with all the objective and subjective indicators loaded onto a single latent variable. This model fit the data poorly, $\chi^2(28) = 237.438$, CFI = .745, TLI = .590, RMSEA [90% CI] = .147 [.130-.165], SRMR = .117. A chi-square differences test also indicated that the two-factor model fit significantly better than the one-factor model ($\Delta\ \chi^2$ (1) = 183.08, p < .001).

Table 8. Fit statistics for latent variable models for Study 1

Model	χ^2 (df)	χ^2 /df	CFI	TLI	RMSEA [90% CI]	SRMR
Measurement Models						
2-Factor	54.362 (27)	2.01	.967	.944	.054 [.033-.075]	.042
True Bifactor	--	--	--	--	--	--
Bifactor Subjective-Residual	51.369 (24)	2.14	.967	.937	.058 [.036-.079]	.040
Bifactor Objective-Residual	44.357 (22)	2.02	.973	.944	.054 [.031-.077]	.035
Hierarchical Model	54.362 (27)	2.01	.967	.944	.054 [.033-.075]	.042
Confirmatory Factor Analyses						
2-Factor	760.400 (442)	1.72	.907	.882	.046 [.040-.051]	.055
True Bifactor	--	--	--	--	--	--
Bifactor Subjective-Residual	757.589 (439)	1.73	.907	.881	.046 [.040-.051]	.055
Bifactor Objective-Residual	748.092 (437)	1.71	.909	.883	.045 [.040-.051]	.054
Full Hierarchical	850.309 (453)	1.88	.884	.856	.050 [.045-.056]	.067
Reduced Predictor Hierarchical	324.822 (190)	1.71	.914	.886	.045 [.037-.054]	.057

Our next preregistered measurement model was a bifactor model, which attempted to account for common variance across all the objective and subjective indicators of sustained attention ability, while also modeling residual shared variance that was unique to each indicator type. Unfortunately, there were signs of misfit (e.g., warnings of negative error variances), so we could not successfully fit a full bifactor model.

As preregistered, then, we next attempted to fit separate bifactor models where each had only one residual factor modeled (e.g., a common sustained attention factor plus a residual objective-indicator factor, with no residual TUT factor). As seen in Table 9, each of these models adequately fit the data. Table 10 presents the factor loadings for each model. In the Subjective-Residual model, all indicators aside from one (TUT rate from the WM task) significantly loaded onto the general Sustained Attention factor, although the TUT-rate loadings were weak. Additionally, there was enough remaining shared variance in TUT reports to successfully model a Subjective-Residual factor. In the Objective-Residual model, many of the indicators significantly loaded onto the general Sustained Attention variable, but none of the objective SART indicators did, and all performance indicator loadings were weak. After accounting for general sustained attention ability, there was still enough shared variance left over to successfully model an objective residual latent variable. Thus, in these two separate models, we were able to assess general sustained attention ability as the individual-differences overlap among objective and subjective measures.

Table 9. Standardized factor loadings (and standard errors) for latent variable measurement models for Study 1

Construct and Measure	Model Name			
	Two Factor Measurement	Bifactor Sub-Res Measurement	Bifactor Obj-Res Measurement	Hierarchical Measurement
Working Memory Capacity				
OPERSPAN				
READSPAN				
SYMSPAN				
Attention Control				
Antisaccade				
Cued Visual Search				
Stroop				
Processing Speed				
CRT Bin 1				
PVT μ				
Stroop Bin 1				
Prosaccade M RT				
Alertness				
PVT				
CRT				
Continuous Tracking				
Antisaccade				
Motivation				
PVT				
CRT				
Continuous Tracking				
Antisaccade				
General Sustained Attention				
PVT Bin 5		.62 (.05)	.24 (.06)	
SART RTSD		.46 (.06)	.06 (.07)	
SART Omissions		.49 (.06)	.05 (.07)	
SART τ		.37 (.06)	.11 (.07)	
CRT τ		.58 (.05)	.18 (.06)	
Continuous Tracking Variability		.66 (.05)	.28 (.06)	

Table 9 (continued). Standardized factor loadings (and standard errors) for latent variable measurement models for Study 1

Construct and Measure	Model Name			
	Two Factor Measurement	Bifactor Sub-Res Measurement	Bifactor Obj-Res Measurement	Hierarchical Measurement
PVT TUTs		.17 (.07)	.64 (.05)	
SART TUTs		.28 (.07)	.74 (.04)	
WRWM TUTs		.14 (.07)	.65 (.04)	
Stroop TUTs		.24 (.06)	.67 (.04)	
Objective/Objective^{resid}				
PVT Bin 5	.62 (.05)		.57 (.05)	.62 (.05)
SART RTSD	.46 (.06)		.48 (.06)	.46 (.06)
SART Omissions	.48 (.06)		.51 (.06)	.48 (.06)
SART τ	.37 (.06)		.36 (.07)	.37 (.06)
CRT τ	.58 (.05)		.54 (.05)	.58 (.05)
Continuous Tracking Variability	.66 (.05)		.60 (.05)	.66 (.05)
Subjective/Subjective^{resid}				
PVT TUTs	.61 (.05)	.59 (.05)		.61 (.05)
SART TUTs	.76 (.04)	.70 (.05)		.76 (.04)
WRWM TUTs	.66 (.05)	.66 (.05)		.66 (.05)
Stroop TUTs	.67 (.04)	.62 (.05)		.67 (.04)

Note. Bifactor Sub-Res = bifactor model with a subjective-indicator residual factor; Bifactor Obj-Res = bifactor model with an objective-indicator residual factor; OPERSPAN = operation span; READSPAN = reading span; SYMMSPAN = symmetry span; PVT Bin 1 = Mean RT of the fastest 20% of trials in the PVT; PVT Bin 5 = Mean RT of the slowest 20% of trials in the PVT; SART RTSD = intrasubject standard deviation in RT from SART; PVT = Psychomotor Vigilance Task; SART = Sustained Attention to Response Task. CRT = Choice Reaction Time Task; WRWM = Whole Report Working Memory task; TUTs = TUT rate from task

It is worth emphasizing, however, that the loadings on the general factor were heavily favored by the "absent-residual" factor in each model. That is, in the Subjective-Residual model, the general factor reflected mostly variance from the objective indicators, and in the Objective-Residual model, the general factor mostly reflected variance from the subjective indicators. This imbalance of factor-loading weights will likely impact correlations between the "general" sustained attention factor and other constructs (see below), and they suggest that these reduced bifactor models might inadequately describe the data, despite reasonable global fit indices (Bornovalova et al., 2020).

Confirmatory Factor Analyses of Individual Differences in Sustained Attention

Our next set of preregistered analyses assessed the correlations between our nomological-net predictor constructs with our sustained attention factors. While our focus was on the bifactor models, we first present the correlations between our predictors and the 2-factor sustained attention model to attempt replication of the correlations from Unsworth et al. (2021). A model with latent variables for WMC, Attention Control, Speed of Processing, Motivation, Alertness, and manifest variables for openness, conscientiousness, extraversion, agreeableness, neuroticism, and cognitive failures adequately fit the data (Table 9) and all predictor indicators loaded onto their respective constructs (see Table 11; as in Unsworth et al. [2021], we fixed the loadings of the dispositional manifest variables equal to one). We note, however, that the TLI for this model, and for all subsequent structural models that included predictor constructs, was just below the minimum cut-off for adequate fit. We therefore interpret these CFA models with some caution and discuss implications of these findings in the Study 1 Discussion.

Table 10. Standardized factor loadings (and standard errors) for latent variable confirmatory factor analysis (CFA) models for Study 1

Construct and Measure	Model Name			
	Two Factor CFA	Bifactor Sub-Res CFA	Bifactor Obj-Res CFA	Hierarchical CFA
Working Memory Capacity				
OPERSPAN	.71 (.05)	.71 (.05)	.71 (.05)	.71 (.05)
READSPAN	.64 (.05)	.64 (.05)	.64 (.05)	.65 (.05)
SYMSPAN	.62 (.05)	.62 (.05)	.62 (.05)	.61 (.05)
Attention Control				
Antisaccade	.55 (.05)	.55 (.05)	.55 (.05)	
Cued Visual Search	-.56 (.05)	-.56 (.05)	-.56 (.05)	
Stroop	-.18 (.06)	-.18 (.06)	-.18 (.06)	
Processing Speed				
CRT Bin 1	.68 (.04)	.68 (.04)	.68 (.04)	.67 (.05)
PVT μ	.47 (.05)	.47 (.05)	.47 (.05)	.46 (.05)
Stroop Bin 1	.82 (.04)	.82 (.04)	.82 (.04)	.85 (.04)
Prosaccade M RT	.27 (.06)	.27 (.06)	.27 (.06)	.23 (.06)
Alertness				
PVT	.84 (.03)	.84 (.03)	.84 (.03)	
CRT	.68 (.04)	.68 (.04)	.68 (.04)	
Continuous Tracking	.55 (.04)	.55 (.04)	.55 (.04)	
Antisaccade	.58 (.04)	.58 (.04)	.58 (.04)	
Motivation				
PVT	.82 (.03)	.82 (.03)	.82 (.03)	
CRT	.68 (.04)	.68 (.04)	.68 (.04)	
Continuous Tracking	.53 (.04)	.53 (.04)	.53 (.04)	
Antisaccade	.55 (.05)	.55 (.05)	.55 (.05)	
General Sustained Attention				
PVT Bin 5		.65 (.04)	.29 (.06)	
SART RTSD		.46 (.06)	.06 (.06)	
SART Omissions		.46 (.06)	.05 (.06)	
SART τ		.36 (.06)	.11 (.06)	
CRT τ		.59 (.05)	.17 (.06)	
Continuous Tracking Variability		.65 (.04)	.26 (.06)	

Table 10 (Continued). Standardized factor loadings (and standard errors) for latent variable confirmatory factor analysis (CFA) models for Study 1

Construct and Measure	Model Name			
	Two Factor CFA	Bifactor Sub-Res CFA	Bifactor Obj-Res CFA	Hierarchical CFA
PVT TUTs		.25 (.06)	.74 (.04)	
SART TUTs		.23 (.06)	.65 (.06)	
WRWM TUTs		.12 (.07)	.61 (.05)	
Stroop TUTs		.25 (.06)	.66 (.04)	
Objective/Objective$^{\text{resid}}$				
PVT Bin 5	.65 (.04)		.58 (.05)	.62 (.05)
SART RTSD	.47 (.05)		.48 (.06)	.41 (.05)
SART Omissions	.46 (.05)		.48 (.05)	.48 (.06)
SART τ	.37 (.06)		.35 (.06)	.32 (.06)
CRT τ	.59 (.05)		.56 (.05)	.60 (.05)
Continuous Tracking Variability	.65 (.04)		.59 (.05)	.64 (.05)
Subjective/Subjective$^{\text{resid}}$				
PVT TUTs	.73 (.04)	.68 (.04)		.63 (.04)
SART TUTs	.67 (.04)	.64 (.04)		.75 (.04)
WRWM TUTs	.61 (.05)	.61 (.05)		.65 (.05)
Stroop TUTs	.67 (.04)	.61 (.04)		.70 (.04)

Note. Bifactor Sub-Res = bifactor model with a subjective-indicator residual factor; Bifactor Obj-Res = bifactor model with an objective-indicator residual factor; OPERSPAN = operation span; READSPAN = reading span; SYMMSPAN = symmetry span; PVT Bin 1 = Mean RT of the fastest 20% of trials in the PVT; PVT Bin 5 = Mean RT of the slowest 20% of trials in the PVT; SART RTSD = intrasubject standard deviation in RT from SART; PVT = Psychomotor Vigilance Task; SART = Sustained Attention to Response Task. CRT = Choice Reaction Time Task; WRWM = Whole Report Working Memory task; TUTs = TUT rate from task.

Although not preregistered, we also included residual correlations between PVT μ with other PVT measures in all models that included processing speed indicators, following from a post-hoc residual correlation added in Unsworth et al. (2021). Table 12 displays correlations with the two sustained-attention factors. In general, we replicated the correlations reported in Unsworth et al. (2021), even after changing the data-processing pipeline and some indicators. For example, the objective factor correlated strongly with attention control, (r = −.86 vs. r = −.69 in Unsworth et al., 2021) and processing speed (r = .51 vs. r = .47. in Unsworth et al., 2021), and the subjective factor correlated strongly with alertness (r = −.78 vs. −.78 in Unsworth et al., 2021) and motivation (r = −.67 vs. −.65 in Unsworth et al., 2021).

Table 11. Latent variable correlations from Study 1 two-factor model

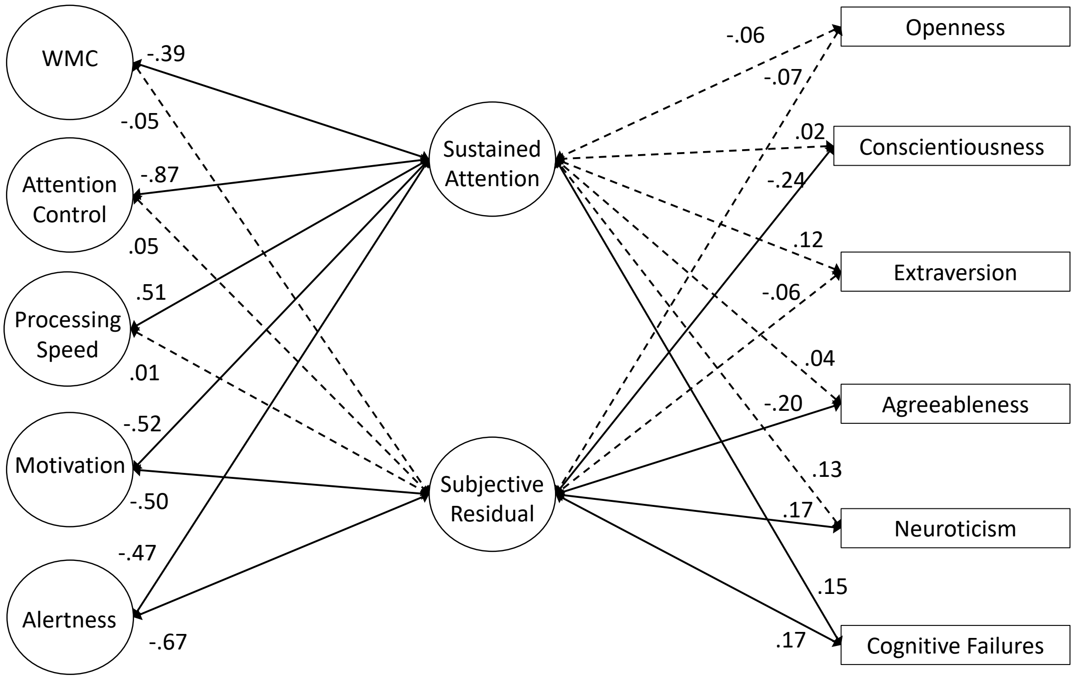
Construct/Measure	1	2	3	4	5	6	7	8	9	10	11	12
1) Objective Sustained Attention												
2) Subjective Sustained Attention	.33											
3) WMC	-.39	-.18										
4) Attention Control	-.86	-.24	.50									
5) Processing Speed	.51	.18	-.33	-.67								
6) Alertness	-.47	-.78	.19	.42	-.17							
7) Motivation	-.52	-.64	.21	.33	-.14	.77						
8) Openness	-.06	-.09	.01	-.01	-.05	.15	.04					
9) Conscientiousness	.02	-.22	-.07	-.02	.04	.14	.14	-.04				
10) Extraversion	.13	-.01	-.02	.02	-.00	.05	.03	.15	.10			
11) Agreeableness	.04	-.18	.00	-.07	.02	.19	.13	-.01	.25	.11		
12) Neuroticism	.12	.21	-.18	-.25	.06	-.13	-.08	-.03	-.19	-.26	-.32	
13) Cognitive Failures	.15	.21	-.01	-.09	-.04	-.12	-.08	.01	-.40	-.05	-.19	.40

Note. WMC = Working Memory Capacity

We next ran two separate nomological-network CFAs with each of the reduced bifactor models. We first report results for the Subjective-Residual-only model, and then from the Objective-Residual-only model. In each model, we allowed correlations among the predictor variables to be estimated, and they were consistent across the models and similar to those presented in the two-factor CFA model (exact correlations among predictors in these models can be found in Supplemental Table S11).

The Subjective-Residual-Only model is presented in Figure 7 (for clarity, Table 10 presents individual indicators and their factor loadings). Several correlations appeared consistent with our predictions. First, individual differences in WMC and attention control ability both were negatively correlated with general sustained attention (in)ability, with a stronger correlation for attention control. That is, subjects with greater WMC and attention control showed fewer sustained attention failures. As well, subjects with slower processing speed exhibited poorer sustained attention. Finally, subjects who reported higher motivation and alertness also showed fewer sustained attention failures.

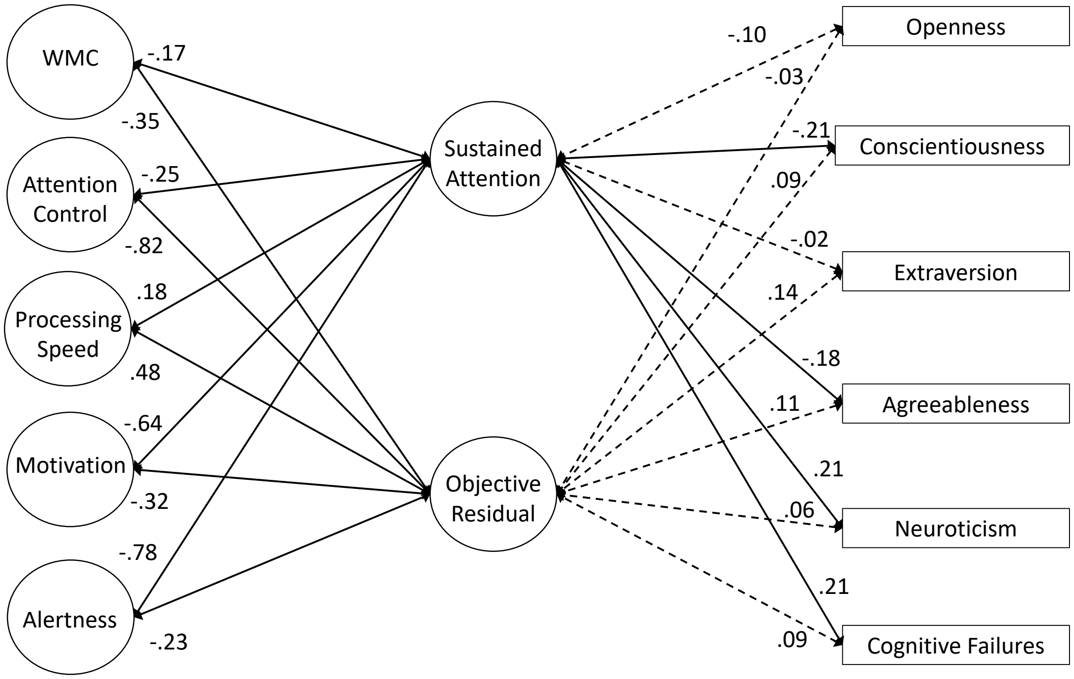
Figure 7. Confirmatory factor analysis of the Subjective-Residual Model



Note. WMC = Working Memory Capacity. Standardized path estimates are presented. For clarity, factor loadings are not presented here; see Table 5 for factor loadings for all models included in the primary analyses**.**

In terms of dispositional constructs, only self-reported cognitive failures significantly (but weakly) correlated with general sustained attention (in)ability: Subjects who reported having more daily memory and attention failures also showed poorer sustained attention in the lab. None of the personality measures significantly correlated with the general sustained attention factor. Finally, only the self-report measures (motivation, alertness, agreeableness, neuroticism, and cognitive failures) correlated with the subjective-residual factor. These correlations are perhaps unsurprising, as the subjective-residual factor likely captures some variance related to self-assessments and self-beliefs, self-reporting biases, and socially desirable responding that might also influence responding to the contextual and dispositional self-rating measures.

Figure 8. Confirmatory factor analysis of the Objective-Residual Model



Note. WMC = Working Memory Capacity. Standardized path estimates are presented. For clarity, factor loadings are not presented here; see Table 5 for factor loadings for all models included in the primary analyses.

The Objective-Residual Model is displayed in Figure 8. Here, the cognitive individual-differences variables again correlated with general sustained attention ability, albeit more weakly. Self-reported alertness and motivation again strongly correlated with general sustained attention. Lastly, conscientiousness, agreeableness, neuroticism, and cognitive failures all correlated significantly (but modestly) with general sustained attention ability, most in the hypothesized directions: Higher conscientiousness and agreeableness were related to better sustained attention, while higher neuroticism and greater cognitive failures were related to worse sustained attention ability. Correlations with the objective-residual factor were limited to our cognitive and contextual variables; none of the dispositional variables correlated significantly with the objective-residual factor.

In general, the correlations across these two separate models appear to follow the trends of the two-factor model. When general sustained attention ability is captured primarily by objective indicators (i.e., when the bifactor model includes a subjective residual factor), associations with *cognitive and contextual variables* are more aligned with predictions. On the other hand, when general sustained attention ability is primarily captured by variance in subjective indicators (i.e., when the bifactor model includes an objective residual factor), associations with *contextual and dispositional variables* are more in line with predictions. We will return to the complexities of interpreting the general factors from these reduced bifactor models below.[10]

Exploratory Hierarchical Model of Sustained Attention

Because we had to conduct the bifactor models as separate reduced models (each with a different residual factor), the "general" factor did not clearly represent a general sustained attention (in)ability. This could be seen in the factor loadings of each model. Specifically, the general factor was primarily a reflection of objective measures in the subjective-residual model, and a primary reflection of subjective measures in the objective-residual model, suggesting that the reduced bifactor models did not adequately describe the data (Bornovalova et al., 2020). Thus, this imbalance in the loadings on the general factor impacted the correlations with the constructs within the nomological network.

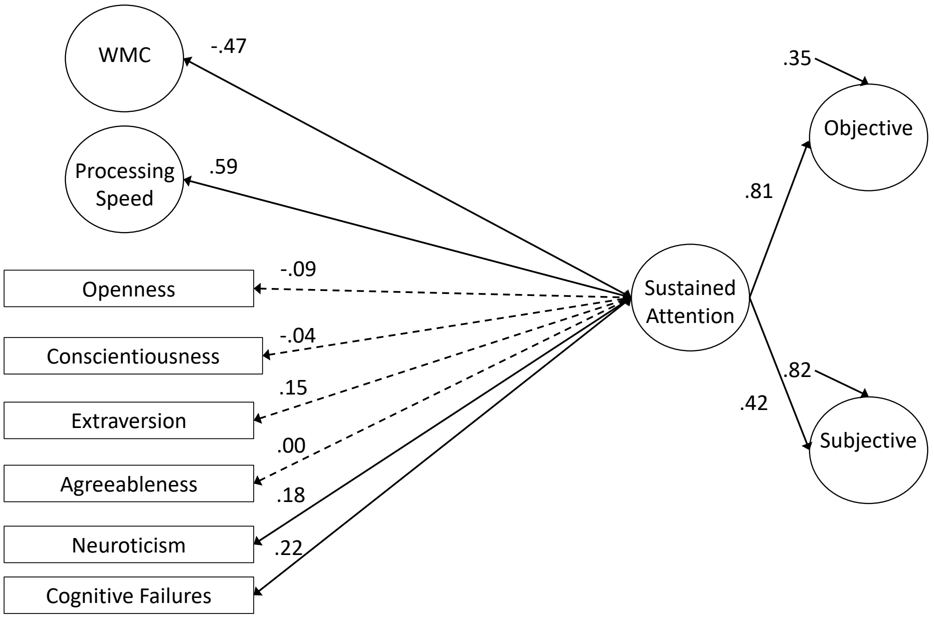
[10] As a preregistered exploratory set of analyses, we also investigated what the objective-residual factor might reflect. However, the results here were especially complicated by the need to use the reduced bifactor model with a general factor biased toward the TUT indicators. Our original hypothesis was that the objective-residual factor might primarily represents processing speed, given that many of the contributing indicators are RT based. Or, that it may reflect strategy choice in sustained attention tasks (i.e., speed-accuracy trade-off). Although the reduced bifactor structure of the model prevented us from drawing clear conclusions, we present the results of these preregistered analyses in Supplemental Tables S13 and S14.

To remedy this, we ran an *exploratory (non-preregistered)* hierarchical model to represent the general factor. Our intention was to see whether a second-order sustained attention factor that was equally loaded by first-order objective and subjective latent variables would provide some clarity about the associations between the general sustained attention factor and other constructs. We first ran a measurement model with general sustained attention ability as a second-order factor (with first-order factors loaded by the objective and subjective indicators), rather than a first-order general factor across the individual indicators. In order to identify a hierarchical model with only two first-order factors, we set the unstandardized paths of both the objective and subjective factors to 1 (Kline, 2011). This hierarchical model adequately fit the data (Table 9). Again, all individual indicators loaded onto their respective first-order latent variables (see Table 10). Additionally, the first-order latent variables were both predicted by a second-order sustained attention latent variable (Objective β = .73, Subjective β = .44). Note, however, that the residual variances for the first-order factors were large (Objective ζ = .46, Subjective ζ = .81). Despite the model fitting the data well, there was still variance that could not be explained in each first-order factor by the higher-order factor (as expected from the two-factor model showing only a moderate correlation between objective and subjective factors).

We next ran a CFA including the individual-differences constructs of interest to assess their correlations with the general sustained attention factor. When including all constructs of interest in the model, the data did not provide adequate fit (see Table 9). Inspection of the model summary indicated that the paths from attention control, alertness, and motivation to the general sustained attention factor were all $|r| > 1.0$. We thus ran a second model without these nomological network constructs included. Model fit was improved and consistent with our previous models (although the TLI was still slightly below threshold). We again allowed the predictor constructs to correlate (see Supplemental Table S12).

As seen in Figure 9, WMC and processing speed both significantly correlated with the general sustained attention factor: Individuals with better WMC and faster processing speed showed fewer sustained attention failures. Only two dispositional variables showed significant (but modest) correlations with the general sustained attention variable: Individuals high in neuroticism and those who report more everyday attention and memory failures had more sustained attention failures. Given the exploratory nature of the measurement model, and the selectivity of this nomological-network model, we interpret the results with caution. Future work should consider this hierarchical structure of sustained attention as a possible model (as we will do in Study 2) and preregister analyses to investigate such associations.

Figure 9. Confirmatory factor analysis of the reduced hierarchical model



Note WMC = Working Memory Capacity. Standardized path estimates are presented. For clarity, factor loadings are not presented here; see Table 5 for factor loadings for all models included in the primary analyses**.**

Mini-Multiverse Analyses

Extremely long RTs may reflect occasional lapses of sustained attention (and will necessarily increase variability in RTs when aggregated at the subject level). It is also possible, however, that these RTs might sometimes result from behaviors or events completely unrelated to failures of sustained attention (e.g., sneezing, or intentionally looking away to check the time). Thus, researchers face the challenging question of how to handle especially long RTs. Researcher degrees of freedom for treating outlying RTs are infinite, and so there is no single answer.

Choices about the cutoff values and consequences for outlying RTs (and subjects) can alter RT distributions and bias estimates of RT measures and their correlations with other performance variables. To investigate the robustness of our findings, then, we next conducted a preregistered mini-multiverse analysis (Steegen et al., 2017; see also specification curve analysis, Simonsohn et al., 2015). Our previous work (Welhaf et al., 2020) has found that in a large-scale dataset with many hundreds of subjects and multiple tasks per construct (i.e., the Kane et al. [2016] dataset examined in present Study 2), decisions about outlying RT treatments and outlying subject treatments yielded negligible changes in estimates for correlations between latent variables for sustained-attention lapses and cognitive ability.

For the current study's primary analyses above, we based RT outlier decisions on a cutoff value equal to each subject's median RT + 3*IQR (for each task). Here, however, for each subject, and each task, we created different datasets that either (a) retained outlying RTs, (b) censored outlying RTs to the cutoff value, or (c) cut outlying RTs. We also extended this process to univariate outlier subjects after aggregating the objective sustained attention measures (i.e., retained outlying subjects, censored outlying subjects' scores to the cutoff value, or dropped outlying subjects' data).

We focused our analyses on the Subjective-Residual bifactor, Objective-Residual bifactor, and hierarchical structural models, as these models provided estimates of general sustained attention ability. The results are visually depicted in Supplemental Figure S3, where it's clear that the findings were robust to varying outlier treatments. For the Subjective-Residual bifactor model, many of the cognitive, contextual, and dispositional correlations with the general sustained attention factor were significant and all correlations were within .06 of the primary model correlations. There was some variability in significance in the correlations between the general sustained attention factor, neuroticism, and self-reported cognitive failures. All correlations with the subjective residual followed the pattern of the primary models and were all within .10 of the primary correlation.

For the Objective-Residual bifactor Model, all correlations with the sustained attention general factor again followed the primary model results. Correlations with the general factor were even more stable, with all iterations within .03 of the primary model estimates. For the objective residual factor, correlations were overall more variable, but still within .06 of the primary estimates, and largely consistent in terms of significance. The only difference in significance occurred with one path for each extraversion and cognitive failures with the objective residual factor.

Finally, for the hierarchical model, we dropped the attention control, alertness, and motivation constructs (as in the primary analyses), and the remaining correlations were largely consistent, but variability across the models was more apparent. Most correlations were within .10 of the primary model estimates. The largest inconsistencies in significance were isolated to associations with neuroticism, which was significantly correlated with the second-order sustained attention factor in half of the iterations. Overall, the results of our mini-multiverse analyses largely indicate that associations between our predictor variables and sustained attention factors were robust to outlier-definition and outlier-treatment criteria.

Discussion

Our reanalysis of Unsworth et al. (2021) provided preliminary evidence for the construct validity of objective and subjective sustained attention measures, and their covariation. First, objective measures of attention consistency showed high levels of within-task redundancy, with most potential indicators correlating with each other above $r = .70$. Second, our chosen objective and subjective indicators of attention consistency all loaded onto their respective latent variables, which were moderately correlated with each other (*r* = .33). This objective–subjective correlation indicated the potential for modeling a common sustained attention factor, an ostensibly superior way to validly assess the sustained attention construct.

We attempted to fit a full bifactor model to the data but were unsuccessful. Alternative bifactor models that separately modeled the individual residual factors did fit the data. These reduced models allowed us to examine the associations between both a common sustained attention factor and the residual factors, with other constructs in the nomological network. Results from CFAs were somewhat in line with our predictions: Individual differences in cognitive ability (e.g., WMC, attention control, and processing speed) were significantly correlated with the common factor, such that individuals with better abilities showed better sustained attention. Individual differences in contextual factors (e.g., self-reported alertness and motivation) also correlated with the common factor, with higher levels of each being associated with better sustained attention. Finally, dispositional characteristics provided some evidence for convergent and discriminant validity of the general sustained attention factor. Specifically, as predicted, cognitive failures consistently correlated with the common sustained attention factor in each model indicating convergent validity. Openness and extraversion, in contrast, were consistently uncorrelated with the general factor, suggesting some evidence for discriminant validity. Correlations with other personality traits such as agreeableness, conscientiousness, and neuroticism were inconsistent across models and future work should is needed to determine how these variables fit in the nomological network of sustained attention measures.

In each alternative bifactor model, we also examined associations with the residual factors. As hypothesized, the cognitive ability factors did not correlate with the subjective residual (which might capture processes like self-beliefs and socially desirable responding), but they did with the objective residual (which might reflect constructs like processing speed or strategy choice/implementation). The contextual self-report factors of alertness and motivation correlated with both the subjective and objective residual factors, but more strongly with the subjective. Finally, many trait-dispositional factors correlated with the subjective residual, which is unsurprising given they rely on similar self-report methods and might both be influenced by response biases and beliefs. Dispositional factors did not correlate with the objective residual.

Given that the full bifactor model (with separate residual factors for both the objective and subjective indicators) did not converge, we could not adequately assess the common sustained attention factor. Each reduced bifactor model had a concerning degree of bias that muddied our interpretation of the correlations with the general factor (see Bornovalova et al., 2020). Specifically, in the subjective-residual model, the general factor was heavily weighted by objective indicators. Thus, to better investigate the general sustained attention factor, we conducted a series of *exploratory (non-preregistered)* analyses using a hierarchical sustained attention model. Here, our intention was to see whether a second-order factor that had equally loading first-order objective and subjective factors could clarify any associations between a general sustained attention ability and other constructs.

The second-order factor significantly correlated with WMC and processing speed (a model with attention control and the contextual factors yielded correlations > 1.0) and with neuroticism and cognitive failures, which suggests some consistency across the models and with the literature (e.g., Kane et al., 2016; Unsworth et al., 2021). Future research should consider this hierarchical model as a worthy approach to assessing individual differences in general sustained attention ability, especially given the potential challenges of fitting bifactor models with both objective and subjective indicators residuals.

We must highlight some areas of concern for the current study. Despite the measurement models adequately fitting the data, none of the structural bifactor models that included predictor constructs fit adequately across all indices (i.e., TLI values were slightly below conventional cut-offs). As well, because our exploratory hierarchical model had only two first-order factors (one for performance indicators and one for TUT rates), we had to constrain their unstandardized loadings onto the general factor to be equal to allow an identified model. All our CFA models exploring the nomological network of sustained attention should therefore be interpreted with caution until replications can support their conclusions. With that said, the results of Study 1 are largely robust to different outlier treatments and so we have confidence that the models can provide some preliminary evidence of the structure of sustained attention ability.

Study 2

Study 1 provided preliminary evidence for the construct validity of a general sustained attention construct measured across objective and subjective indicators. However, concerns about measurement and fit of the structural models warrant caution in interpreting the results. Study 2 therefore serves as a conceptual replication using an independent dataset (Kane et al., 2016) to see whether: (a) the proposed bifactor structure of attention consistency measures can be modeled, and (b) sustained-attention factors correlate with theoretically relevant constructs.

Study 2 assessed, in addition to WMC, a new construct—positive schizotypy—to further investigate convergent and discriminant validity. Prior work has demonstrated that positive schizotypy, reflecting the proneness to have unusual beliefs and perceptual experiences (and a risk factor for schizophrenia and related disorders), is related to both objective and subjective indicators of attention consistency. Specifically, subjects with higher positive schizotypy scores (from self-report questionnaires) show more variable RTs in basic attention tasks (Kane et al.

2016; Schmidt-Hansen & Honey, 2009) as well as higher TUT rates (Kane et al. 2016), than do those with lower scores. Thus, positive schizotypy should be negatively correlated with an assessment of general sustained attention ability.

Methods

Below we describe the general procedure and materials from Kane et al. (2016). We provide detailed descriptions of the tasks and measures selected for the current study in their respective sections.

Subjects

Kane et al. (2016) enrolled 545 undergraduates into their study from the University of North Carolina at Greensboro. Of these, 541 completed the first of three 2-hr sessions, 492 completed the second, and 472 completed all three. Full-information maximum likelihood (ML) estimation was used for missing data (see Kane et al., 2016, for details and sample demographics).

Materials

Objective Attention Consistency. As in Study 1, for each objective indicator task, we assessed multiple dependent measures that theoretically should reflect variation in sustained attention. We focus the present analyses on tasks where RT was the primary outcome in Kane et al. (2016). Our procedure for picking a primary indicator for each task was identical to that for Study 1. Again, we list the possible dependent measures for each task with the *a priori* measure listed first. For many of the tasks in Study 2, however, there was no measure that has traditionally been used to reflect sustained attention (aside from the SART), so we tried to balance RTsd, τ, and slowest 20% of trials across the tasks such that each was the primary indicator for at least one of the “attention restraint” tasks (i.e., SART and Stroop-like tasks), and one was the primary measure among the “attention constraint” tasks (i.e., flanker tasks).

Semantic SART. In this go/no-go task, subjects pressed the space bar for words from one category (*animals*; 89% of trials) but withheld responding to words from another category (*vegetables*; 11% of trials). Stimuli were presented for 300 ms followed by a 1500 ms mask. There were 675 trials, 75 of which were no-go targets. The potential dependent variables derived from this task were intra-individual RTsd, intra-individual RTmad, omission errors, mean RT of the slowest 20% of trials, mean RT of the fastest 20% of trials, the τ estimate from an ex-Gaussian model, and RMSSD, all for correct “go” trials.

Number Stroop. Subjects reported the number of digits presented on each trial while ignoring the digits’ identity. Each trial presented 2 to 4 identical digits in a row and subjects responded with one of three labeled keys to indicate the number of digits on screen. There were 300 total trials: 240 were congruent (e.g., “*333*”) and 60 were incongruent (e.g., “*222*”). The potential dependent variables derived from this task were the *τ* estimate from an ex-Gaussian model, RTsd, RTmad, and mean RT of the slowest 20% of trials, all for correct congruent trials.

Spatial Stroop. Subjects reported the relative position of a word to an asterisk (left, right, above, below), with the word and asterisk both presented to the left or right, or above or below, fixation; subjects ignored both the identity of the word (“*LEFT*,” “*RIGHT*,” “*ABOVE*,” “*BELOW”*) and absolute location of the word and asterisk on-screen. Subjects responded to the relative position of the word to the asterisk by pressing the corresponding arrow on the numeric keypad arrow keys. Subjects completed a total of 120 trials: 60 presenting words congruent for absolute and relative location, 30 presenting words congruent for absolute location but incongruent for relative location, and 30 presenting words incongruent both for absolute and relative location. Here, the potential dependent variables derived from this task were mean RT of the slowest 20% of trials, RTsd, RTmad, and the τ estimate from an ex-Gaussian model, all for correct responses to trials where words were congruent for both absolute and relative position.

Arrow Flanker. Subjects reported the direction of a centrally presented arrow ("<" vs. ">") via keypress, with the arrow flanked horizontally by 4 distractors. Subjects completed two blocks of 96 trials: 24 neutral trials (target arrow presented amid dots), 24 congruent trials (all arrows pointing the same direction), 24 stimulus-response incongruent trials (central arrow pointing opposite direction of flankers), and 24 stimulus-stimulus incongruent trials (central arrow presented amid upward-pointing arrows). Here, the potential dependent variables derived from this task were mean RT of the slowest 20% of trials, RTsd, RTmad, and the τ estimate from an ex-Gaussian model, all for correct responses to both neutral and congruent trials.

Letter Flanker. Subjects reported whether a centrally presented "F" appeared normally or backwards via keypress, with that letter flanker horizontally by 6 distractors. Subjects completed 144 trials: 24 neutral trials (normal or backwards F presented amid dots), 48 congruent trials (target and distractor Fs all facing the same direction), 24 stimulus-response incongruent trials (target facing opposite direction of distractors), and 24 stimulus-stimulus incongruent trials (target presented amid right- and left- facing Es and Ts tilted at 90 and 270 degrees). Here, the potential dependent variables derived from this task were RTsd, RTmad, mean RT of the slowest 20% of trials, and the τ estimate from an ex-Gaussian model, all for correct responses to neutral and congruent trials.

Circle Flanker. Subjects reported whether a target letter was an X or N, via keypress, with the target flanked by two distractors. Targets appeared in one of eight possible locations in a circle, with distractors appearing on either side of the target; all other locations were occupied by colons. Subjects completed 160 trials: 80 neutral trials (target letter surrounded by colons) and 80 conflict trials (target flanked by two different distractors from the set H, K, M, V, Y, Z). Here, the potential dependent variables derived from this task were the τ estimate from an ex-Gaussian model, RTsd, RTmad, and mean RT of the slowest 20% of trials, all using correct responses to neutral trials.

Subjective Attention Consistency Measures. Thought probes were randomly presented in 5 tasks (45 in SART, 20 in Number Stroop, 20 in Arrow Flanker, 12 in Letter Flanker, and 12 in an otherwise-unanalyzed 2-back task). Each probe presented subjects with eight categories of thoughts they might have just experienced. Subjects selected their options by pressing the number on the keyboard that most closely matched the content of their immediately preceding thoughts. The options were: 1) "The task" (thoughts about the stimuli or responses); 2) "Task experience/performance" (thoughts about how one was performing on the task); 3) "Everyday things" (thoughts about normal life concerns, goals, and activities); 4) "Current state of being" (thoughts about one's physical, cognitive, or emotional states); 5) "Personal worries" (thoughts about current worries); 6) "Daydreams" (fantastical, unrealistic thoughts); 7) "External environment" (thoughts about task-unrelated things or events in the immediate environment); 8) "Other." TUTs were assessed as the proportion of responses with options 3-8, as in Kane et al. (2016).

WMC Tasks. Subjects completed four complex span tasks and two updating tasks. As in Study 1, for each complex span task, subjects completed three practice stages: the first provided practice in memorizing small sets of the memoranda (e.g., letters, grid, or arrow locations); the second practice was for processing-only (e.g., math equations, symmetry decisions, sentence comprehension, letter direction). RTs were recorded during this processing only practice for each subject. During the real trials, if a processing decision was not made within 2.5 SDs of the processing-only mean, that trial was counted as a processing error; the third practice consisted of both the memory and processing task combined (as in the real trials).

Operation Span. Same as Study 1. Here, however, the set sizes of 3 to 7 were presented three times in a random order rather than twice (max score of 75).

Reading Span. Same as Study 1. Here, however, the set sizes of 2 to 6 were presented three times in a random order rather than twice (max score of 60).

Symmetry Span. Same as Study 1. Here, however, the set sizes of 2 to 5 were presented three times in a random order rather than twice (max score of 42).

Rotation Span. Subjects were presented with random sequences of large and small arrows to remember, radiating from a center location in one of 8 possible directions. Between presentation of each arrow, a rotated letter (F, G, J, or R) was presented facing its normal direction or mirror-reversed (50% of the time) and subjects had to verify its direction. At the end of the set, subjects recalled the arrows from the set by clicking the location onscreen in the presented serial order. Subjects were granted credit only if the arrow was recalled in the correct serial position. Set sizes ranged from 2–5 items; each set size was presented three times (for a max score of 42). Higher scores reflected better recall.

Running Span. Subjects were presented with a sequence of letters and were asked to recall only the final 3–7 letters from the trial. Trials were unpredictably 0, 1, or 2 items longer than the set size (e.g., set size 5 had list lengths of 5, 6, and 7 items in the task). Each trial started with a number to indicate the set size (i.e., the number of items to be recalled at the end of the list). At the end of the list, all 12 possible letters appeared on-screen along with the corresponding set size and subjects selected via mouse-click the appropriate letters from the set (in serial position). Subjects completed 15 total trials. Credit was granted for items that were recalled in the correct serial position (for a max of 75). Higher scores reflected better recall.

Updating Counters. Subjects recalled the numerical values presented in boxes, some of which were updated from their original values. Each trial began with 3–5 boxes presented horizontally on-screen. There were three phases for each trial: (1) the learning phase, where a digit (1 thru 9) was presented in a random order in each box; (2) the updating phase, where 2–6 of the box values were changed by presenting a simple addition or subtraction (e.g., +4; −1; updates ranged from −7 to +7). Updates appeared randomly and some boxes could have been updated multiple times, or not at all; (3) the recall phase, where subjects were tasked with recalling the final updated value for each box (cued in a random order). Set sizes of 3–5 boxes were crossed with the number of updates (2–6) yielding a total of 15 trials. Credit was granted for correct answers and the score was proportion correct (out of 60). Higher scores reflected better recall.

Positive Schizotypy. Positive schizotypy was assessed using two of the Wisconsin Schizotypy Scales (WSS; Chapman & Chapman, 1983)—the Perceptual Aberration scale and Magical Ideation scale—and the Referential Thinking subscale of the Schizotypal Personality Questionnaire (Raine, 1991). Subjects saw each item on-screen individually and responded via mouse-click if the item was true for them (scored as 1) or false (scored as 0). After appropriate reverse-scoring, items were summed for each scale where higher scores indicated more endorsement of the schizotypic belief or experience. (Note that in Kane et al. [2016], Social Anhedonia item parcels were also included [as cross-loadings] in the positive schizotypy factor; however, their factor loadings were weak [<.30] and, as expected, they loaded more strongly on the negative schizotypy factor, and so they will not be included for the current analyses.)

Results

We again report our preregistered analyses and results and note where we deviated from the preregistered plan. Data and Rmarkdown files for all analyses are available on the Open Science Framework (https://osf.io/xeu63/).

Data Analysis Exclusions

Prior to calculating any of the primary DVs for the current study, we calculated "go" trial accuracy for the SART and identified 6 subjects who had "go" trial accuracy < 70%. Consistent with Study 1, *we deviated from our preregistration* and excluded SART data from these subjects, as such low accuracy might indicate a failure to understand or comply with task instructions, rather than failures of sustained attention.

RT Cleaning Procedures

We implemented the same preregistered multi-step procedure for trial-level cleaning as Study 1. First, we identified and removed RTs for error and post-error trials (and post-probe trials in tasks that included thought probes). Next, we removed RTs for trials that were likely anticipations (i.e., RTs < 200 ms). From the remaining trials, we next calculated for each subject, in each task, a value equal to their median RT + 3*IQR. Any trials outside of this value were replaced with this value. Supplemental Table S15 reports the relevant descriptive information for the trial-level cleaning (e.g., mean number of trials outlying trials replaced per task). Finally, we calculated the number of usable trials each subject had following our RT cleaning protocol. As preregistered, we dropped task data from subjects who did not have at least 40 trials and thus could not reliably contribute to our primary measures of interest. This resulted in dropping data from 2 subjects in the Spatial Stroop task and 6 in the Arrow flanker task.

Selection of Objective Indicators

We followed the same procedures for selecting objective attention consistency indicators as in Study 1. Supplemental Table S15 presents the descriptive statistics for each possible DV, for each task, as well as the number of subjects whose data were censored for each potential measure.

No variables were removed from consideration for problematic distributions (skew > 3.0 or kurtosis > 10.0). We next examined split-half reliability to remove any unreliable indicators (i.e., < .50). No variables were removed for poor reliability. As in Study 1, we next examined the within-task bivariate correlations to see whether any combination of measures provided non-redundant information with the *a priori* measure for each task. As seen in Supplemental Table S16, many of the proposed measures indicated a high level of redundancy (*r*s > .70) within each task, for the Number Stroop, Spatial Stroop, Arrow flanker, Letter flanker, and Circle flanker tasks. Because of this, we selected only the *a priori* measure for each of these tasks as the performance indicator of sustained attention. However, as in Study 1, we found evidence in the SART that some of the potential indicators were not redundant with one another. Specifically, RTsd (the *a priori* measure) and Omissions were non-redundantly correlated (*r* = .51), so we retained both.

Multivariate Outliers

As preregistered, once we established primary indicators, we again checked for multivariate outliers in the final dataset using the *Routliers* package (Leys et al., 2019) to calculate Mahalanobis distance for each observation in the dataset. This analysis indicated there were 10 multivariate outliers in the dataset (~2% of the subjects). These subjects were removed case-wise before any structural modelling was conducted.

Descriptive Statistics and Correlations for Final Dataset

Table 12 provides the descriptive statistics for the final dataset of Study 2. Supplemental Table 16 provides the bivariate correlations of all measures of interest. Consistent with Kane et al. (2016), measures from the same proposed construct (e.g., WMC, Positive Schizotypy, TUT rates) all correlated more strongly with each other than with measures of other constructs. Importantly, and consistent with Study 1, our newly selected objective attention consistency indicators also correlated moderately with each other (median |*r*| = .30) suggesting that subjects who showed variable responding in one task also tended to do so in other tasks.

Table 12. Descriptive statistics for Study 2 measures

Construct/Measure	Mean	SD	Min	Max	Skew	Kurtosis	N
Objective Sustained Attention							
SART RTsd	159.30	58.75	36.64	361.34	1.20	1.73	510
SART Omissions	22.96	24.86	0.00	94.00	1.43	1.14	510
Number Stroop τ	91.14	39.74	14.96	220.92	1.31	1.65	458
Spatial Stroop Bin 5	974.20	289.51	511.50	1857.01	1.34	1.55	446
Arrow Flanker Bin 5	630.31	111.51	426.90	1011.21	0.92	0.67	464
Letter Flanker RTsd	121.79	54.88	40.56	307.44	1.21	1.37	452
Circle Flanker τ	113.31	58.01	0.00	284.81	1.16	1.37	458
Subjective Sustained Attention							
SART TUTs	0.51	0.24	0.00	1.00	-0.04	-0.81	510
Number Stroop TUTs	0.43	0.29	0.00	1.00	0.38	-0.90	458
Arrow Flanker TUTs	0.49	0.30	0.00	1.00	0.11	-1.07	464
Letter Flanker TUTs	0.58	0.26	0.00	1.00	-0.47	-0.54	452
N-Back TUTs	0.42	0.31	0.00	1.00	0.31	-1.09	451
Working Memory Capacity							
OPERSPAN	0.00	1.00	-3.54	1.70	-0.75	0.31	465
READSPAN	0.00	1.00	-2.77	2.27	-0.23	-0.44	413
SYMSPAN	0.01	0.99	-3.22	2.01	-0.37	-0.17	457
ROTSPAN	0.02	0.97	-3.19	2.10	-0.48	-0.09	377
RUNNSPAN	0.00	0.99	-2.72	2.84	0.22	-0.10	452
COUNTERS	-0.01	0.99	-2.04	3.24	0.55	0.17	470
Positive Schizotypy							
PERCABER	6.37	4.96	0.00	31.00	1.55	3.43	523
MAGIDEA	11.43	5.57	0.00	28.00	0.24	-0.52	523
REFTHINK	3.35	2.06	0.00	7.00	0.09	-1.03	469

Note. WMC scores are z-scores. SART = Sustained Attention to Response Task. TUTs = Rate of Task-Unrelated Thoughts in specified task. OPERSPAN = Operation Span. READSPAN = Reading Span. SYMSPAN = Symmetry Span. ROTSPAN = Rotation Span. RUNNSPAN = Running Span. COUNTERS = Updating Counters task. PERCABER = Perceptual Aberration Total score. MAGIDEA = Magical Ideation Total Score. REFTHINKING = Referential Thinking score. Bin 5 = Mean RT of Slowest 20% of correct trials. RTsd = intra-individual RT variability. Bin 1 = Mean RT of Fastest 20% of correct trials.

Measurement Models of Sustained Attention

As preregistered, our first set of analyses attempted to conceptually replicate the latent variable correlation between performance and self-report indicators of attention consistency from Study 1. We first tested a 2-factor model with latent variables for objective (i.e., RT variability) and subjective (i.e., TUT reports) measures; these latent variables were allowed to correlate. Consistent with Study 1 (*but not preregistered for either study*), we included within-task residual correlations between the TUT rate and performance indicator from the same task (e.g., Number Stroop τ with Number Stroop TUTs). We retained these residual correlations for all subsequent models. As seen in Table 13, the model fit the data adequately. Moreover, the latent variables for objective and subjective attention consistency measures again correlated moderately, as in Study 1 (r = .38, here, and r = .32 in Study 1; see Supplemental Table S17 for factor loadings).[11] Again, this *moderate* correlation suggests that while these two types of sustained attention measures share some variance, they are not redundant. Modeling the shared variance among the indicators should provide a more construct valid measure of sustained attention, free from measurement error specific to either indicator type.

[11] *Although not preregistered*, we tested whether a single factor Sustained Attention model fit the data, as we did in Study 1. To do so, we specified a model where all the objective and subjective indicators loaded onto a single latent variable. This model did not adequately fit the data, χ^2 (46) = 317.381, CFI = .825, TLI = .749, RMSEA [90% CI] = .106 [.095-.117], SRMR = .101. A chi-square differences test also indicated that the two-factor model was a significantly better fitting model ($\Delta\ \chi^2$ (1) = 208.17, p < .001).

Table 13. Fit statistics for latent variable models for Study 2

Model	χ^2 (df)	χ^2 /df	CFI	TLI	RMSEA [90% CI]	SRMR
Measurement Models						
2-Factor	109.213 (45)	2.43	.959	.939	.052 [.040-.065]	.042
Bifactor	66.444 (34)	1.95	.979	.959	.043 [.027-.058]	.028
Hierarchical	109.213 (45)	2.43	.959	.939	.052 [.040-.065]	.042
Confirmatory Factor Analysis						
2-Factor	453.284 (253)	1.79	.951	.941	.039 [.033-.044]	.053
Bifactor	403.998 (240)	1.68	.959	.949	.036 [.030-.042]	.049
Hierarchical	460.799 (255)	1.81	.949	.940	.039 [.033-.045]	.056
Limited AC Bifactor	417.746 (242)	1.73	.955	.945	.037 [.031-.042]	.052
Limited AC Hierarchical	454.008 (256)	1.77	.949	.941	.038 [.032-.044]	.056

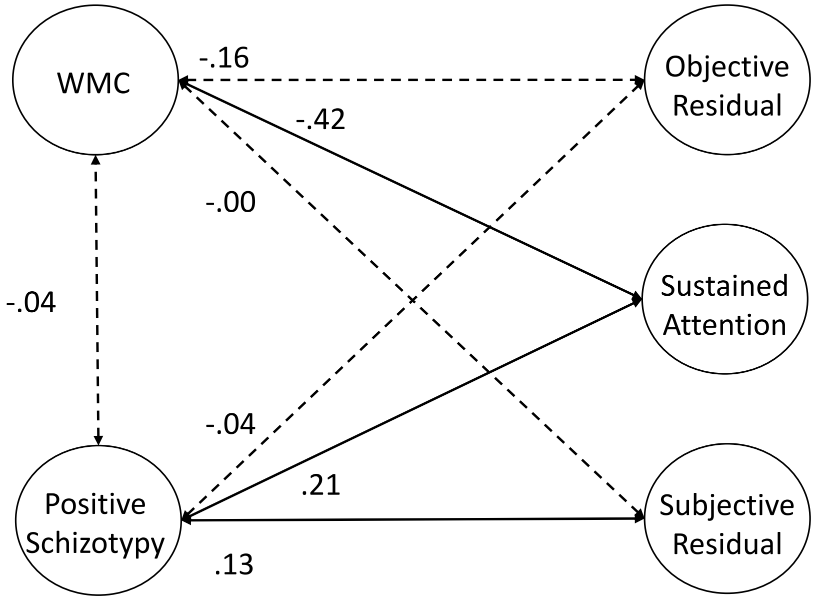
Note. AC = Attention Control

Our next preregistered measurement model was a bifactor model, which attempted to model common sustained-attention variance across all the objective and subjective indicators, as well as residual variance unique to each indicator type. Unlike Study 1, this full bifactor model provided an adequate fit to the data (see Table 13). All indicators loaded significantly onto the general sustained attention factor (although the loadings were mostly stronger for the objective than subjective indicators) and there was also enough shared variance among the measures to model both an objective and subjective residual factor (see Supplemental Table S17 for factor loadings).

Preregistered Confirmatory Factor Analyses of Individual Differences in Sustained Attention

Our next set of preregistered analyses assessed the correlations between our nomological-network constructs with our different sustained attention models. Although our focus was the bifactor model, we first present the correlations between our predictors and the two-factor sustained attention (failures) model. A model with latent variables for WMC and positive schizotypy adequately fit the data (Table 13) and all indicators loaded onto their respective factors (Supplemental Table S17). WMC correlated negatively with the objective (r = −.42) and subjective factors (r = −.19): Subjects with higher WMC exhibited fewer performance lapses and TUT reports. In contrast, positive schizotypy correlated positively with both the objective (r = .16) and subjective factors (r = .21): Individuals who endorsed more positive schizotypy experiences exhibited more performance lapses and more TUT reports. WMC did not correlate with positive schizotypy (r = −.04).

Figure 10. Confirmatory factor analysis of the bifactor model of sustained attention (failures) for Study 2



Note. WMC = Working Memory Capacity. Standardized path estimates are presented. For clarity, factor loadings are not presented here; see Table 17 for factor loadings for all models included in the primary analyses.

Table 14. Standardized factor loadings (and standard errors) for Limited Attention Control latent variable models for Study 2

Construct and Measure	Model Names	
	Bifactor CFA	Hierarchical CFA
Working Memory Capacity		
OPERSPAN	.62 (.04)	.62 (.04)
READSPAN	.50 (.05)	.50 (.05)
SYMSPAN	.62 (.05)	.62 (.05)
ROTSPAN	.52 (.06)	.52 (.06)
RUNSPAN	.58 (.04)	.58 (.04)
COUNTERS	.63 (.04)	.62 (.04)
Positive Schizotypy		
PERCABER1	.60 (.03)	.60 (.03)
PERCABER2	.58 (.03)	.58 (.03)
PERCABER3	.64 (.03)	.64 (.03)
MAGIDEA1	.84 (.03)	.85 (.03)
MAGIDEA2	.83 (.03)	.83 (.03)
MAGIDEA3	.72 (.04)	.72 (.04)
REFTHINK	.66 (.03)	.66 (.03)
Attention Control		
SART d’	-.45 (.04)	-.46 (.04)
Antisaccade Letters	.77 (.03)	.77 (.03)
Antisaccade Arrows	.76 (.04)	.76 (.04)
General Sustained Attention		
Number Stroop τ	.69 (.10)	
Spatial Stroop Bin 5	.25 (.12)	
Arrow Flanker Bin 5	.37 (.12)	
Letter Flanker RTSD	.40 (.11)	
Circle Flanker τ	.63 (.09)	
Number Stroop TUTs	.41 (.08)	
Arrow Flanker TUTs	.30 (.07)	
Letter Flanker TUTs	.22 (.07)	
N-Back TUTs	.27 (.06)	
Objective/Objective^{resid}		
Number Stroop τ	.20 (.16)	.61 (.04)
Spatial Stroop Bin 5	.50 (.09)	.51 (.04)
Arrow Flanker Bin 5	.64 (.09)	.67 (.04)
Letter Flanker RTSD	.47 (.11)	.62 (.04)
Circle Flanker τ	.32 (.13)	.68 (.04)
Subjective/Subjective^{resid}		
Number Stroop TUTS	.62 (.07)	.74 (.05)
Arrow Flanker TUTS	.61 (.07)	.68 (.05)
Letter Flanker TUTS	.45 (.06)	.50 (.05)
N-Back TUTS	.56 (.06)	.63 (.05)

Note. OPERSPAN = operation span; READSPAN = reading span; SYMMSPAN = symmetry span; ROTASPAN = rotation span; RUNNSPAN = running span; COUNTERS = updating counters; SART RTSD = intrasubject standard deviation in RT from SART; Letter Flanker RTSD = intrasubject standard deviation in RT from Letter Flanker; PERCABER1 = perceptual aberration scale (parcel 1); PERCABER2 = perceptual aberration scale (parcel 2); PERCABER3 = perceptual aberration scale (parcel 3); MAGIDEA1 = magical ideation scale (parcel 1); MAGIDEA2 = magical ideation scale (parcel 2); MAGIDEA3 = magical ideation scale (parcel 3); REFTHINK = referential thinking subscale from the Schizotypal Personality Questionnaire (SPQ), SART = Sustained Attention to Response Task. TUTs = TUT rate from task.

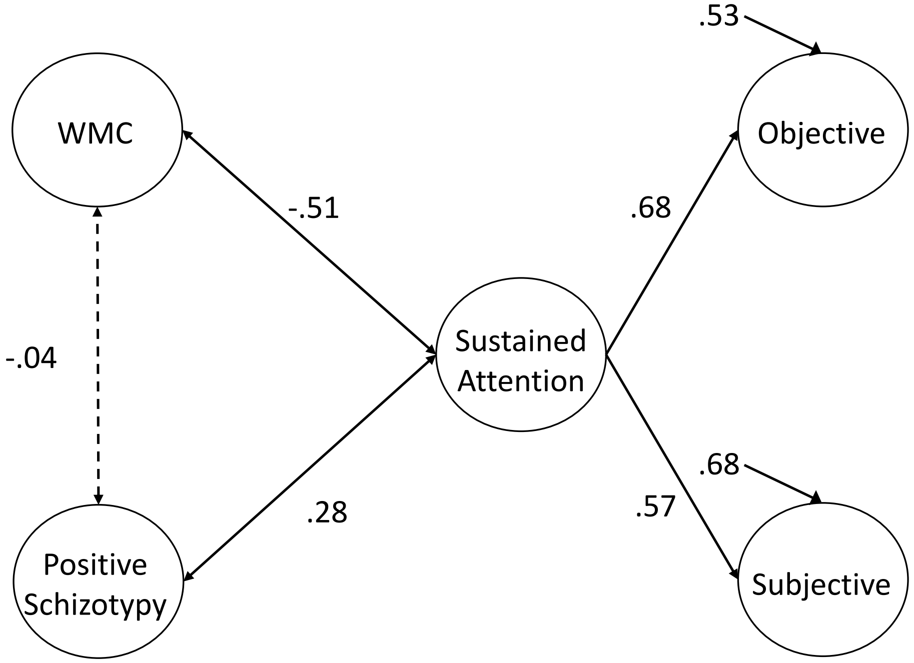
To assess whether our predictor constructs correlated with a general factor of sustained attention failures, we next ran a CFA with the bifactor model of sustained attention. As seen in Figure 10, both WMC and positive schizotypy significantly correlated with the general sustained attention factor in predicted directions: Subjects with higher WMC and those who reported lower positive schizotypy ratings had fewer sustained attention failures. Neither WMC nor positive schizotypy were correlated with the objective-residual factor. However, there was a weak positive association between positive schizotypy and the subjective-residual factor (WMC did not correlate with this factor).

Exploratory Hierarchical Model of Sustained Attention

As a final exploratory model, following from Study 1, we modeled sustained attention (failures) as a second-order factor above the first-order objective- and subjective-indicator factors. We again set the unstandardized paths of the objective and subjective factors to 1 to yield an identified model. The measurement model showed acceptable fit (see Table 15), with both the objective ($\beta = .68$) and subjective ($\beta = .57$) factors loading significantly onto the second-order sustained attention factor. Again, we note that the variances on the first-order factors were large, suggesting there was still unexplained variance in the model (Objective $\zeta = .57$, Subjective $\zeta = .67$), as expected given the moderate correlation between objective and subjective factors.

We next ran a CFA including WMC and positive schizotypy and the model adequately fit the data (Table 15). As seen in Figure 11, both WMC and positive schizotypy significantly correlated with the second-order sustained attention factor, with similar magnitudes to those with the general factor from the bifactor model, although they were a bit larger here (the WMC path was also of nearly identical magnitude here to that from Study 1 [−.47]): Higher WMC was again related to fewer sustained attention failures whereas higher positive schizotypy scores were again related to more sustained attention failures.

Figure 11. Confirmatory factor analysis of the hierarchical model of sustained attention (failures) for Study 2



Note. WMC = Working Memory Capacity. Standardized path estimates are presented. For clarity, factor loadings are not presented here; see Table 17 for factor loadings for all models included in the primary analyses.

Exploratory CFAs Including A Narrow "Attention Control" Factor

A limitation of our preregistered structural models is that they left us unable to address questions about the potential associations between general sustained attention and other factors of attention control. Recall that Study 1 found that in the hierarchical model, attention control correlated > 1.0 with the sustained attention factor, which caused issues with the model overall led us to drop that factor (along with the motivation and alertness factors) in our reduced hierarchical model. Although the original study on which Study 2 is based (Kane et al., 2016) included separate latent factors for "attention restraint" (response-inhibition-type tasks) and "attention constraint" (flanker distractor-control tasks), here we included indicators from most of these tasks (using their congruent and or neutral-baseline conditions) to model our attention consistency performance factor. *As an exploratory, non-preregistered approach* to the question, however, we modeled an attention control factor using the two antisaccade tasks and the SART d′ measure (reflecting part of the "restraint" factor from Kane et al., 2016), and we removed all SART indicators (RTsd, omissions, TUT rate) from the sustained attention factors. Our measurements of attention control and sustained attention constructs were thus independent. Otherwise, the models matched those represented in Figures 10 and 11, including WMC, positive schizotypy, and either the bifactor or hierarchical model of sustained attention.

Both the bifactor and hierarchical models again fit the data well (see Table 15 for fit statistics and Table 17 for factor loadings). In the bifactor model, attention control (failures) was moderately correlated with the general sustained attention factor (*r* = .31), the objective-residual factor (*r* = .38), and, in contrast to WMC, also with the subjective-residual factor (*r* = .22). In the hierarchical model, attention control (failures) was strongly, but non-redundantly, correlated with the second-order sustained attention factor (*r* = .71). Thus, in both cases, individuals with poorer attention control also showed worse sustained attention ability. Unlike Study 1, then, here we were provisionally able to dissociate sustained attention from attention control, which suggests these may be distinct forms of general executive attentional ability.

Mini-Multiverse Analyses

The Study 2 mini-multiverse analyses focused on the preregistered bifactor model and the exploratory hierarchical model of sustained attention (without including the exploratory attention control factor, as modeling this factor required removing all SART indicators from the sustained attention models). Our multiverse decisions on outlying RTs and outlying subjects were identical to those for Study 1. Details of the results are presented in Supplemental Figure S4, and we summarize them here.

We conducted a series of CFAs on the bifactor model that also included WMC and positive schizotypy constructs. These iterations resulted in only five of the nine models converging. Although discouraging, these results might not be too surprising given the general instability of bifactor models (e.g., Eid et al., 2017, 2018). Of the models that converged, the resulting correlations were generally consistent with the primary model estimates: WMC was negatively associated with the general sustained attention factor and positive schizotypy was positively associated with the general factor. Further, WMC was not associated with the objective residual (aside from one iteration) and was not associated with the subjective residual in any iteration. Positive schizotypy was not associated with the objective residual in any iteration and was positively associated with the subjective residual in all but one iteration. Thus, despite some of the iterations failing to converge, those that did converge presented a reasonably consistent pattern of results. We suggest that the bifactor model is still a promising way to measure individual differences in sustained attention, as the individual-differences overlap in objective (performance) and subjective (self-report) attention consistency measures. At the same time, both Study 1 and Study 2 indicated that bifactor models including both theoretically desirable residual factors (objective and subjective) do not always fit the data adequately and they are not as robust as other models to variation in outlier definitions and treatments.

For the hierarchical model, associations between WMC and positive schizotypy with the second-order sustained attention factor were remarkably consistent across all multiverse iterations. Estimates of the correlation between the second-order factor with WMC were within .03 of the primary correlation and estimates of the correlation with positive schizotypy were within .02 of the primary correlation. As in Study 1, the hierarchical model appears to be a more robust model of general sustained attention than the bifactor model. Thus, if one is simply interested in capturing general sustained attention ability, and less so about the residual or separate first-order factors, then the hierarchical model provides a suitable assessment. We do note again, however, that *we did not preregister our exploration of the hierarchical model* and so we suggest further independent replication of its fit to sustained attention data and its correlations with other constructs. It is, of course, encouraging that the hierarchical model results were similar across both Studies 1 and 2.

Exploratory Latent Profile Analyses of Sustained Attention Subgroup Variation for Studies 1 and 2

The latent variable analyses from each study indicated that some of the correlations between the general sustained attention factor and nomological network constructs were different from the correlations with either objective-only or subjective-only factors in the 2-factor model. We interpret these differences as indicating that the shared variance between objective and subjective measures is a most construct valid measure of sustained attention, free from indicator-type-specific measurement error. However, the bifactor and hierarchical CFAs present some limitations. Namely, the reduced bifactor models from Study 1 produced biased factor loadings on the general sustained attention factor and the hierarchal models in both studies had large error variances on the first-order factors, indicating considerable unexplained variance. Both limitations likely influenced the correlations with the nomological network constructs.

To address these limitations, and to conceptually replicate the correlational findings, we asked whether there were different groups of subjects who varied on the objective and subjective sustained attention factors. And if so, we asked whether these different groups also differed at the mean level on other nomological network constructs. To examine these questions, *in a non-preregistered set of analyses*, we submitted the subjects' factor scores for the objective and subjective latent factors, derived from the 2-factor measurement model from each study, to a series of latent profile analyses (LPAs). LPAs identify latent classes or mixtures of subject profiles from continuous input variables (Gibson, 1959; Oberski, 2016).

The first step of each LPA was to identify which number of profiles best fit the data from the objective and subjective measures. A priori, we expected that 4 profiles might provide the most theoretically interesting solution: (a) subjects who were high on both objective and subjective measures (i.e., poor general sustained attention); (b) subjects who were high on only objective measures; (c) subjects who were high on only subjective measures, and; (d) subjects who were low on objective and subjective measures. We conducted the LPAs using the *tidyLPA* package (Rosenberg et al., 2018). For each study, we compared fit indices for models between 3 and 5 profiles. We compared models using multiple fit indices, including Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and Sample Size Adjusted BIC (SABIC), with lower values indicating better fitting models. We also considered the *p*-values of the Bootstrapped Likelihood Ratio Test (BLRT) which compared a model with *k* classes to a model with *k*-1 classes. Significant *p*-values of the BLRT indicated the model with *k* classes better fit the data. Model fits for both studies are presented in Table 15.

Table 15. Fit statistics for each LPA in each study

Model	AIC	BIC	SABIC	BLRT *p*
Study 1				
3 Profiles	954.86	993.33	961.60	0.010
4 Profiles	958.51	1008.52	967.28	0.495
5 Profiles	938.59	1000.13	949.38	0.010
Study 2				
3 Profiles	1287.71	1330.46	1298.71	0.010
4 Profiles	1289.88	1345.45	1304.19	0.317
5 Profiles	1284.50	1352.90	1302.11	0.030

The LPAs suggested a 5-profile model for Study 1 and a 3-profile model for Study 2. The 5-profile model of Study 1 indicated three subgroups with extremely small sample sizes (*n*s = 16–20) and two larger subgroups, which makes drawing conclusions and interpreting the profiles difficult (Lubke & Neale, 2006; Spurk et al., 2020). We therefore opted to use the 3-profile model, as it was the preferred model for Study 2 (which had a larger sample size) and was more parsimonious (note also that the 3-profile model was the next-best fitting model for Study 1). For both Studies 1 and 2, the 3 sub-groups had similar profiles. Group 1 had higher objective and subjective factor scores (and so poor general sustained attention). Group 2 had moderate sustained attention ability across objective and subjective measures (with Group 2 in Study 2 having slightly higher subjective scores compared to their objective scores), and Group 3 had lower objective and subjective scores (and so generally better sustained attention).

We next compared the three profile groups on the nomological-network constructs included in each study (see Table 16). The comparisons largely follow our latent variable analyses, particularly those involving the general sustained attention factors: Individuals with poorer sustained attention abilities (i.e., Group 1) had lower scores on many of the nomological-network constructs compared to those with moderate and good sustained attention, particularly WMC, attention control, processing speed, motivation, and alertness.

Table 16. Means (standard deviations) and omnibus ANOVA results for each group defined by LPA for Study 1 and Study 2.

Study 1 Measures	Group 1 (N = 18)	Group 2 (N = 120)	Group 3 (N = 206)	F	η_p^2
Objective	1.03 (0.22)	0.27 (0.18)	-0.24 (0.19)	578.36***	0.772
Subjective	0.64 (0.75)	0.31 (0.67)	-0.24 (0.55)	41.35***	0.195
WMC	-0.32 (0.65)	-0.21 (0.58)	0.15 (0.58)	17.34***	0.092
Attention Control	-0.53 (0.42)	-0.19 (0.36)	0.16 (0.39)	49.08***	0.222
Processing Speed	0.68 (0.71)	0.13 (0.60)	-0.14 (0.52)	22.99***	0.118
Motivation	-0.65 (0.82)	-0.33 (0.72)	0.25 (0.66)	35.99***	0.173
Alertness	-0.57 (0.61)	-0.31 (0.60)	0.23 (0.67)	35.24***	0.170
Openness	0.02 (0.38)	-0.33 (0.72)	0.02 (0.55)	0.49	0.003
Conscientiousness	-0.04 (0.38)	-0.04 (0.48)	0.00 (0.62)	0.06	0.000
Extraversion	0.18 (0.52)	0.00 (0.71)	-0.02 (0.83)	0.51	0.003
Agreeableness	0.06 (0.50)	-0.05 (0.63)	0.02 (0.55)	0.77	0.004
Neuroticism	0.04 (0.73)	0.16 (0.76)	-0.10 (0.76)	4.53*	0.026
Cognitive Failures	0.11 (0.94)	0.20 (0.85)	-0.12 (0.89)	5.21**	0.029
Study 2 Measures	Group 1 (N = 45)	Group 2 (N = 205)	Group 3 (N = 277)		
Objective	1.03 (0.28)	0.15 (0.29)	-0.27 (0.25)	508.03***	0.660
Subjective	0.46 (0.41)	0.38 (0.35)	-0.35 (0.33)	311.07***	0.543
WMC	-0.27 (0.48)	-0.08 (0.56)	0.10 (0.58)	11.70***	0.043
Positive Schizotypy	0.26 (1.13)	0.17 (1.12)	-0.16 (1.07)	6.68**	0.025

Note. WMC = Working Memory Capacity. ^ $p < .10$; * $p < .05$; ** $p < .01$; *** $p < .001$

Discussion

Study 2 provided additional evidence for the construct validity of general sustained attention factors that reflect the shared individual-differences variance in performance variability and self-reported TUTs. Using an independent dataset from that in Study 1 (Kane et al., 2016), we captured the proposed full bifactor structure of sustained attention failures, as well as the hierarchical structure explored in Study 1. We also assessed the associations of the general sustained attention factors (from the bifactor and hierarchical models) with WMC and positive schizotypy. Consistent with Study 1, WMC correlated negatively with the general sustained attention (failures) factor with a moderate effect size (≈ −.40 to −.50), despite correlating only weakly with the subjective sustained attention factor from the 2-factor model (−.19): Higher-WMC subjects exhibited better sustained attention ability than did lower-WMC subjects. WMC was not associated with the objective or subjective residual factors from the bifactor model.

The preregistered Study 2 analyses included positive schizotypy, which we predicted to correlate weakly positively with the general (failures) factor and the subjective residual factor (but not with the objective residual). These predictions were largely confirmed. Positive schizotypy was weakly to moderately related to general sustained attention (in)ability in the bifactor and hierarchical models (≈ .20–.30). Positive schizotypy was also (weakly) positively related to the subjective residual factor in the bifactor model (.13), suggesting these subjects might also have general reporting biases or self-beliefs that guide their answers to self-report questions about their thoughts or behaviors.

Our *non-preregistered* analyses featuring a reduced attention control factor also provided additional validity evidence that we were unable to examine in Study 1. In the bifactor model, the attention control factor (comprised of antisaccade-arrows errors, antisaccade-letters errors, and SART d′) was moderately correlated with the sustained attention factor (.31) indicating that poor attention control was related to worse sustained attention. This was also true in the hierarchal model, where the correlation between attention control and sustained attention was stronger (.71). Note that in the 2-factor model of Study 1, attention control was strongly correlated with the objective sustained attention factor ($r = -.86$), which at first glance, would suggest that these two constructs may be isomorphic. However, the results from Study 2 suggest that sustained attention and attention control are not redundant.[12]

We conducted a multiverse analysis like that of Study 1 to assess the robustness of our primary CFA results. The results provided some confirmatory evidence for the robustness of the results but also raised some concerns. Specifically, only half of the multiverse iterations for the bifactor model converged. This suggests that the full bifactor model is not robust to different outlier treatments. These multiverse results, combined with the nonconvergence of the full bifactor model in Study 1, suggest that the bifactor model may not be robust enough to be a broadly useful approach to assessing general sustained attention ability.

Again, the exploratory (*non-preregistered*) hierarchical model of sustained attention appeared more robust to different outlier decisions, as the model converged in every iteration. All estimates of the correlations between WMC and positive schizotypy with the second-order sustained attention factor were consistent across the iterations. Despite not being preregistered, the consistency across the multiverse, and across Studies 1 and 2, suggests that the hierarchical model of sustained attention may be the most useful approach for researchers interested in

[12] We are unsure why the attention control × sustained attention correlation differed so much between the bifactor and hierarchical models in Study 2. A closer look at the confidence intervals (CI) suggests that the hierarchical model may have provided a better estimate, as the CI was smaller than in the bifactor model (Hierarchical CI = [.56, .85]; Bifactor CI = [.09, .53]).

examining general sustained attention ability as the individual-differences covariation between objective and subjective indicators. We argue that it is important to measure sustained attention using both subjective and objective indicators, given the possibility of their tapping different degrees of disengagement, as well their independent sources of measurement error, and that this hierarchal approach can provide some insight on the nomological network of general sustained attention (in)ability.

Finally, in exploratory *(non-preregistered)* latent profile analyses*,* we found additional convergent evidence for the validity of our covariation measurement approach and the general sustained attention factor. Specifically, we found that a group with generally poor sustained attention profile (i.e., lower scores on both objective and subjective factors) had the lowest levels on all the nomological networks. A group with moderate scores on the objective and subjective factor appeared to have scores in the middle on all the nomological network constructs. Finally, the group with the best objective and subjective scores (and so better sustained attention) had generally better scores on the nomological network constructs as well. These results then parallel our correlational analyses and support this idea of the covariation between objective and subjective indicators being a viable—if not optimal—approach to assessing sustained attention.

General Discussion

The present studies examined the construct validity of sustained attention measures in two independent datasets (Kane et al., 2016; Unsworth et al., 2021). The primary goals of each study were to: (1) test whether the individual-differences covariation between objective and subjective measures of attention consistency provided a construct valid assessment of general sustained attention ability and; (2) examine how several cognitive, contextual, and dispositional nomological network constructs were associated with sustained attention ability to assess convergent and discriminant validity of the general sustained attention factor. As a secondary goal, for each study we also conducted a “mini-multiverse” analysis on each dataset to assess the robustness of our findings against different trial-level and subject-level outlier decisions.

Regarding our first goal, the results suggested that objective and subjective sustained attention indicators share variance and thus load onto a common sustained attention factor. Although we could not successfully model a full bifactor structure in Study 1, we were able to fit reduced bifactor models that separately captured unique variance to each indicator type. In Study 2, moreover, we were able to fit a full bifactor model to the sustained attention data. These bifactor models presented some problems, however. In Study 1, the reduced bifactor models yielded general factors that were each dominated by the non-residual-factor indicators, which biased the measurement of the general factor toward either the objective or the subjective measures; in Study 2, the full bifactor model was not robust to different outlier definitions and decisions in the mini-multiverse, as some models did not converge across multiverse iterations. The bifactor approach may not be broadly viable, then, for assessing general sustained attention ability.

In both studies, however, we fit an exploratory (i.e., non-preregistered) hierarchical structure, which modeled general sustained attention as a higher-order factor over the objective-indicator and subjective-indicator latent variables (with only two first-order factors, however, these models required constraining the unstandardized loadings of the objective and subjective factors onto the second order sustained attention factor). This hierarchical approach allowed us to model the individual-differences overlap in objective and subjective indicators as a general sustained attention ability. Unlike the bifactor models, the hierarchical model adequately fit the data in both Studies 1 and 2 and it was robust across multiverse iterations.

Regarding our second goal, the results suggested an interesting pattern of convergent and discriminant validity of the common sustained attention factor. First, individual differences in cognitive ability (i.e., WMC and processing speed) correlated moderately to strongly with the common factor in hypothesized ways: Subjects with better cognitive abilities showed better sustained attention. Second, individual difference in contextual variables such as self-reported motivation and alertness also correlated strongly with the common sustained attention factor: Individuals who reported being more motivated and alert during the cognitive tasks also showed better sustained attention. Finally, dispositional characteristics provided evidence of both convergent and discriminant validity: Self-reported everyday cognitive failures and positive schizotypy symptoms consistently (if moderately) correlated with the common sustained attention factor, indicating that individuals who report or exhibit more of these behaviors and experiences also demonstrate poorer sustained attention. Although big-five personality traits like neuroticism, conscientiousness, and agreeableness correlated with the common factor in some models, but not in others, extraversion and openness did not significantly correlate with the common factor in any model.

In exploratory (*non-preregistered*) analyses, we were able to also assess how a (reduced) attention control factor, derived from response-conflict tasks, correlated with sustained attention ability. Individual differences in attention control correlated moderately to strongly with the general sustained attention factor, providing more evidence for convergent validity. Further, the attention control correlations provided additional evidence for discriminant validity. Specifically, the general sustained attention factor correlated more strongly with attention control (in both bifactor and hierarchical models) than with WMC. WMC tasks involve processes like memory retrieval and strategy choices that aren't necessary in attention tasks, which may contribute to the weaker WMC correlation. Given the exploratory nature of this analysis, we suggest that future research attempt to replicate the findings and further explore the dissociation between attention control (as captured by response- or distractor-conflict tasks) and sustained attention.

General Sustained Attention Ability as the Covariation between Objective and Subjective Indicators

The results of the current study suggest that objective and subjective indicators of attention consistency are moderately correlated with each other, replicating prior work (e.g., Kane et al., 2016; Unsworth, 2015; Unsworth et al. 2021; Welhaf et al., 2020). Moreover, the present study argues that this covariation indicates the presence of a common underlying factor of sustained attention ability that is psychologically meaningful. That is, individual differences in a general sustained attention ability can partly explain RT variability and mind-wandering propensity during simple cognitive tasks. The current results extend prior work that has investigated these measures (e.g., RT variability/performance and TUTs) as separate but related constructs (or as objective sustained attention measures providing validation for subjective measures). This covariation approach is important because objective and subjective measures of sustained attention use two very different methods to assess the same proposed ability. Because each of these measurement types may capture different degrees of disengagement (à la Cheyne et al., 2009), and each is influenced by different non-sustained attention processes, relying on either type of measurement as the sole reflection of sustained attention may lead to improper conclusions about how the ability to sustain attention relates to other psychological constructs.

At the task level, many potential objective indicators of sustained attention are redundant with one another. In both studies, correlations among the individual measurement types (e.g., RTsd, τ, slowest RTs) were high in nearly all tasks. This suggests that for many of the sustained attention tasks used in the literature, researchers may use any of several indicators to objectively measure sustained attention. However, in the SART, different indicators might be picking up on different degrees of sustained attention failures, as we found that multiple indicators shared some variance, but were not redundant. Specifically, we found that RTsd, a commonly used measure from the SART, showed moderate bivariate correlations (*r*s = .30–.69) with omission errors (both in Study 1 and 2), and with τ (in Study 1). Thus, these different SART indicators may capture sustained-attention failures ranging from subtle fluctuations across the task and occasional long RTs, to instances of more complete attentional disengagement (Cheyne et al., 2009; Unsworth et al., 2021).

The Sustained Attention Factor: Construct Validity and Measurement Recommendations

Our proposed model of sustained attention was a bifactor structure in which common variance across the objective and subjective measurement factors could be captured by a general factor, and residual variance unique to each indicator type could be modeled as orthogonal measurement-specific factors. While this was our preregistered and theoretical starting point, the bifactor approach turned out to be inappropriate in Study 1 and not robust to varied outlier treatments in Study 2. We therefore provisionally recommend against taking a bifactor approach for measuring general sustained attention ability.

Instead, when researchers are primarily interested in the general factor of sustained attention, a worthy alternative appears to be a hierarchical model. The hierarchical models from Studies 1 and 2 suggest that general sustained attention ability can be robustly modeled as a higher-order factor representing the variance shared between objective-indicator and subjective-indicator factors. In both studies, this model provided adequate fit and stood up well to various outlier treatment decisions in our mini-multiverse analyses. Thus, *although not preregistered*, we argue that this approach can allow researchers to assess individual differences in general sustained attention ability.

With caveats about both bifactor and hierarchical models in mind (see additional discussion of limitations below), the present results speak to the construct validity of the general sustained attention factor. We focus this discussion on the hierarchical models of Study 1 and Study 2 and the full bifactor model in Study 2, as they provided the most unbiased estimates of the general factor.

First, many correlations with the general factor were slightly stronger than those with the separate objective-measure or subjective-measure factor. Specifically, WMC correlated strongly with the objective factor and weakly with subjective factor in both studies, as is common in the literature (Kane et al., 2016; Unsworth, 2015; Unsworth et al., 2021). WMC correlated substantially, however, with the general factor of sustained attention across models (*r*s = .40–.50). By typically focusing on only objective or subjective indicators, then, prior work may have misestimated the associations between these cognitive ability measures and the ability to sustain attention. If TUT rates reflect a study's only measure of sustained attention, researchers may interpret the association between WMC and sustained attention abilities to be weak. Using the overlap in objective and subjective indicators, in contrast, provides evidence that the link between WMC and sustained attention is rather strong, and perhaps even stronger than associations with either indicator on their own.

This was also the case for processing speed, with much stronger correlations with the objective than subjective measures. The correlation with the common factor, however, was strong (*r* = −.59), and stronger than with either method-specific correlation. Thus, processing speed may be more tied to general sustained attention ability than previous work has shown, especially given the previously mixed results (Unsworth et al., 2021; Welhaf et al., 2020). At the same time, when sustained attention is defined as a general factor derived from both TUT rates and performance variability, it cannot be simply reduced to a processing speed construct. When sustained attention is instead only defined by RT variability measures, however, the close link between *M* RT and RT variability makes it difficult to differentiate sustained attention ability from processing speed.

Sustained attention correlations with dispositional measures provided the clearest evidence of discriminant validity. In Study 1, conscientiousness and agreeableness both correlated with the subjective factor (they did not correlate with the objective factor). However, both variables had weak-to-null, correlations with the general factor in the hierarchical model, suggesting that these traits are not related to general sustained attention ability. Again, if a study used only TUT rates to measure sustained attention, it might erroneously infer a robust association. By using the individual-differences covariation in objective and subjective measures, however, the present evidence suggests a *lack* of a relationship. Further, neuroticism correlated modestly with the subjective factor and nonsignificantly with the objective factor, but it correlated with the general sustained attention factor as strongly as it had with the subjective factor, suggesting a relationship with general sustained ability that goes beyond potential self-report biases. Finally, self-reported cognitive failures and positive schizotypy both correlated more strongly with the common factor than with either the individual objective or subjective factors, again suggesting associations that reflect more than shared method variance. Using the shared variance between objective and subjective indicators as a measure of sustained attention, then, correlations with some trait factors appear to be reliable. Future research should aim to replicate sustained attention × personality relationships and consider other dispositional factors that may inform the nomological network of sustained attention measures, such as daily stress, rumination-proneness, trait anxiety, and ADHD-related symptoms.

Our correlational evidence was conceptually replicated in the LPAs conducted for each study. Here, we found that a 3-profile model could identify specific subgroups in each dataset. When comparing these groups on the nomological network constructs, those with poor sustained attention ability also had lower cognitive and contextual scores compared to those with moderate and good sustained attention ability. In general, modeling sustained attention as the individual-differences covariation between objective and subjective indicators provides us with a more construct valid assessment of sustained attention.

Limitations and Constraints on Generalizability

While the current study has several strengths—such as preregistered analyses of two independent, large-scale, latent-variable studies with different samples of subjects and of tasks—there are limitations worth noting. First, as previously mentioned, a full bifactor model did not adequately fit the Study 1 data, and reduced bifactor models that separately modeled residual objective- and subjective-indicator factors yielded biased general factors. In Study 2, the full bifactor model converged, but did not hold up well across different outlier treatment decisions. Our initially preferred model, then, may not be the most appropriate assessment of general sustained attention ability.

As well, the CFA models in Study 1 to assess the nomological network did not all meet the minimum recommendation for adequate fit, with TLI < .90. We therefore suggest some caution in interpreting these models as there may be some misfit. At the same time, the model fits were consistent with a model using FIML presented in the Appendix of Unsworth et al. (2021) and they were consistent across different outlier decision criteria in our mini-multiverse analysis, which gives us some confidence in their robustness (they were also conceptually replicated in our exploratory latent profile analyses). Given these potential concerns with the Study 1 bifactor models, we emphasize again that our exploratory (*non-preregistered*) hierarchical model fit the data well in both studies and withstood different outlier treatments.

However, we should note that our hierarchical models, themselves, have limitations. Because we were only able to use two first-order latent variables as indicators of the second-order general factor, we had to constrain the unstandardized loadings of the objective and subjective first-order factors to be equal for the model to be identified (Kline, 2011). As well, the variances of the first-order factors were large, indicating considerable unexplained variance in the first-order measurement factors not accounted for by the general factor. It may therefore be useful to include other first-order factors in the hierarchical model to better identify the general factor. One possibility is to design future studies to model two correlated objective-measure factors (i.e., RT-based and accuracy-based factors) and two correlated subjective-measure factors (i.e., TUT-rate factors derived from two different probe types). Or, as we discuss below, study designs could include another sustained attention indicator types in the model (e.g., pupillary responses). Broadening the first-order factors may improve measurement of the general sustained attention factor.

Because the present study reanalyzed existing data, we had no control over task selection. Although some of the most common tasks were used in one or both studies (e.g., the PVT, SART, and CRT), there are other tasks that might be more suitable for measuring sustained attention in future latent-variable studies, because they more purely tap into sustained attention processes than those that might be heavily influenced by additional cognitive processes (e.g., conflict tasks like the Stroop and flanker tasks). For example, the metronome response task (Seli, Cheyne et al., 2013), continuous temporal expectancy task (O'Connell et al., 2009), and gradual onset continuous performance task (Rosenberg et al., 2013) have all been used in studies of sustained attention (mostly for their performance measures, but some have also included thought probes to measure TUTs).

Although the present work focuses on overt behavioral measures (i.e., performance measures and subjective self-reports), recent research has also identified pupil diameter, and its fluctuations, as potential physiological indicators of sustained attention processes and abilities (Unsworth & Robison, 2017a, 2017b). Pupil dilations may be indirectly related to locus coeruleus-norepinephrine system functioning (Aston-Jones & Cohen, 2005; Rajkowski et al., 1994; but see Megemont et al., 2022), which is linked to overall physiological arousal and attention. Most relevant to our current operationalization of sustained attention, fluctuations in pupillary responses may also (imperfectly) reflect moment-to-moment consistency of sustained attention (Hutchinson et al., 2020; Unsworth & Robison, 2017a). Indeed, fluctuations in pupil measures correlate moderately with both objective and subjective measures of sustained attention (Murphy et al., 2011; Unsworth et al., 2020; Unsworth & Robison, 2017b; 2017c; van den Brink et al., 2016). Pupillary responses may therefore serve as another indicator of sustained attention in bifactor or hierarchical models. Our fundamental argument is that future research should consider a *methodological triangulation* approach (Denzin, 1970) to assessing sustained attention. Here, the individual difference covariation between objective performance measures, subjective self-reports, and physiological indicators of attention consistency may be used to best capture the general ability to sustain attention.

Finally, the two datasets analyzed here relied on student samples. It is possible that this factor structure differs, or is even inadequate, with clinical or older adult samples. Older adults shower greater RT variability (consistent with theories of age-related declines in executive attentional control), but often show lower TUT rates, compared to younger adults (see Bunce et al., 2004; Hultsch et al., 2002; Jackson & Balota, 2012; Jordano & Touron, 2017). One possibility is that there are processes, independent of sustained attention, that selectively influence one of these sustained attention indicators in older adults. For example, older adults traditionally have slower processing speed compared to younger adults (Salthouse, 1996), and processing speed is strongly correlated with RT variability. This might contribute to why older adults show poorer sustained attention as indicated by objective measures, but it does not explain why older adults show reduced TUT rates. To overcome this ambiguity, it might be necessary to examine the covariation between objective and subjective measures to better understand how attention consistency changes with age. Future work should consider the implications of aging on the current factor structure of sustained attention.

Conclusions

Sustained attention is an understudied individual-differences construct, given its important contributions to successful performance of many laboratory tasks and everyday activities. The results of the current reanalyses suggest that individual differences in sustained attention, as measured by the shared variance across objective (performance) and subjective (self-report) indicators of attention consistency, can provide a more construct valid measurement of sustained attention than either of these methods separately. Individuals with higher WMC, better attention control, and faster speed of processing showed better sustained attention. Further, contextual factors were strong correlates of sustained attention: Subjects who reported being more alert and motivated also had better sustained attention in the context of challenging lab tasks. Finally, sustained attention failures were selectively, and more weakly, related to some dispositional factors: Subjects who reported higher levels of neuroticism, more frequent everyday cognitive failures, and higher positive schizotypy scores had poorer sustained attention in challenging lab tasks.

In general, hierarchical models of sustained attention may be a suitable approach for researchers interested in estimating general sustained attention ability, given concerns about measurement and robustness of bifactor models. These results expand on previous literature by suggesting an improved way for measuring sustained attention in two ways. First, objective and subjective measures of sustained attention may capture different attentional states along a continuum of disengagement (Cheyne et al., 2009) and each measurement type has its own unique sources of measurement error. So, relying solely on one of these types of measurement approaches can lead to biased, improper conclusions about their relationship with nomological network constructs. Second, some constructs' correlations might be *stronger* with a general sustained attention factor than with either objective or subjective factor alone, which suggests a possible underestimation of the link between sustained attention and other factors in its nomological network in prior research.

CHAPTER IV: A COMBINED EXPERIMENTAL–CORRELATIONAL APPROACH TO THE CONSTRUCT VALIDITY OF PERFORMANCE-BASED AND SELF-REPORT-BASED MEASURES OF SUSTAINED ATTENTION

Abstract

The ability to sustain attention is often measured with either objective performance indicators, like within-person RT variability, or subjective self-reports, like mind wandering propensity. A more construct valid approach, however, may be to assess the covariation in these performance and self-report measures, given that each of these is influenced by different sources of measurement error. If the correlation between performance-variability and self-report measures reflects the sustained attention construct, then task manipulations aimed at reducing the sustained attention demands of tasks should reduce the correlation between them (in addition to reducing mean levels of variability and mind wandering). The current study investigated this claim with a combined experimental-correlation approach. In two experiments (*N*s ~ 1500 each), participants completed tasks that either maximized or minimized the demand for sustained attention. Our demand manipulations successfully reduced the mean levels of sustained attention failures in both the objective and subjective measures, in both experiments. In neither experiment, however, did the covariation between these measures change as a function of the sustained attention demands of the tasks. We can therefore claim only minimal support for the construct validity of our measurement approach to sustained attention.

Introduction

Sustained attention, from an attention consistency perspective (Unsworth & Miller, 2021), reflects “the purposeful act of maintaining optimal task focus to successfully, and consistently, perform goal-relevant actions” (Welhaf & Kane, 2022), and is a crucial ability for many everyday behaviors and tasks. Laboratory investigations into failures of sustained attention indicate that such lapses can manifest in multiple ways. The two most common methods for assessing them are intra-individual variability in task performance and probed self-reports of task-unrelated thoughts (TUTs). Each of these measures are multi-determined and are influenced by separate factors unrelated to sustained attention, however, and so neither alone should be relied on to measure sustained attention ability (Welhaf & Kane, 2022). Instead, we argue that the individual-differences covariation in these measures should be the most construct valid reflection of sustained attention (in)ability.

The goal of the current study is to investigate and evaluate the construct validity of sustained attention measurement using a combined experimental–correlational approach. We aim to harness the strengths of both the construct representation approach (i.e., experimental) and nomothetic span approach (i.e., individual-differences) to construct validation (e.g., Borsboom et al., 2004; Cronbach & Meehl, 1955; Embretson, 1983). More specifically, we will investigate whether individual differences in sustained attention failures, as indicated by the covariation in objective (i.e., task performance) and subjective (i.e., self-reported mind wandering) measures, are reduced through theoretically driven experimental manipulations of sustained attention demands. That is, when tasks make significant sustained attention demands, objective and subjective attention consistency measures should be more strongly correlated because variation in each outcome is caused more by sustained attention processes than by other nuisance variables. When tasks make little demand on sustained attention, however, these correlations should weaken because individual variation in the measures is caused more by *other* (non-sustained-attention) processes, which are unique to either RT variability or TUT rate.

Sustained Attention as the Covariation between Objective and Subjective Measures

The cognitive psychology literature has primarily measured sustained attention abilities in one of two ways. The first has relied on ("objective") performance measures, like intraindividual reaction time (RT) variability, or particular error types committed, during simple lab tasks. The second approach has relied on ("subjective") probed self-reports of task-unrelated thought (TUT) during ongoing tasks or activities.

Performance measures, like RT variability, capture subtle fluctuations in participants' response readiness and their consequent frequency of relatively long RTs. Errors of commission (i.e., responding when a non-response is required) and errors of omission (i.e., not responding when a response is required) may also capture sustained attention failures that reflect mindless responding or more complete disengagement from the task. When participants maintain optimal sustained attention, they should exhibit greater consistency in their responding, fewer instances of relatively long RTs, and fewer performance errors. Likewise, reports of TUTs reflect sustained attention failures that participants subjectively experience and can verbalize, or at least can report on when asked. TUT reports are often correlated with poor task performance in the moment, and TUT rates are often correlated with overall task performance, suggesting that TUTs reflect, at least in part, momentary lapses of sustained attention.

Nomothetic Span Evidence for Construct Validity

Between-subject analyses from latent-variable studies have found that performance-based and self-report-based measures of attention consistency correlate moderately positively (*r* ~ .30–.40; Kane et al., 2016; Unsworth 2015; Unsworth et al., 2021; Welhaf & Kane, 2022; Welhaf et al., 2020): Participants who show more variable responding also report more TUTs. Further, within-subject analyses indicate more variable RTs on task trials leading up to TUT reports compared to on-task reports (e.g., Bastian & Sackur, 2013; Kane, Smeekens et al., 2021; Seli, Cheyne, et al., 2013). These findings suggest that there may be a common ability, or collection of cognitive processes, that contributes to each kind of measure.

We have previously examined this individual-differences overlap between performance and self-report measures as a construct-valid way to assess sustained attention ability (Welhaf & Kane, 2022). Taking a nomothetic span (i.e., individual-differences) approach to construct validation (Cronbach & Meehl, 1959; Embretson, 1983), we reanalyzed two large-*N* data sets (Kane et al., 2016; Unsworth et al. 2021) that included multiple tasks from which we could derive performance indicators of sustained attention, and multiple tasks including thought probes from which we could calculate TUT rates. Each of these studies also included multiple nomological-network constructs, such as working memory capacity (WMC), attention control, processing speed, self-reported motivation and alertness, and multiple dispositional factors like the Big-5 personality traits and positive schizotypy.

In each dataset, we found that latent variables of objective and subjective sustained attention correlated moderately (*r*s = .32 and .38). This shared variance in objective and subjective indicators could be modeled as a general sustained attention factor (with both bifactor and hierarchical structures). Moreover, we argued that the *moderate*—rather than strong—correlation between objective and subjective factors indicates that each is significantly affected by distinct, non-sustained-attention processes, and so using their overlap should be especially important to validly capturing the sustained attention construct.

To evaluate the construct validity of this general sustained attention factor, we compared its pattern of correlations with the nomological network constructs to those with the individual objective and subjective factors from the two-factor sustained attention model. These comparisons provided evidence for both convergent and discriminant validity of the general factor. In terms of convergent validity, WMC, processing speed, self-reported neuroticism, everyday cognitive failures, and positive schizotypy experiences were all correlated with the general sustained attention factor. Critically, these constructs correlated as strongly, if not more strongly, with the general factor than they did with either the individual objective-measure or subjective-measure factor. As for discriminant validity, self-reported agreeableness and conscientiousness were both weakly and non-significantly correlated with the general factor, even though they were significantly correlated with the subjective-measure factor in the two-factor model. These findings suggested that the individual-differences covariation in objective and subjective sustained attention measures may be the most construct-valid method to assess sustained attention ability.

Our initial evidence for the construct validity of sustained attention measures was correlational, as our investigation took a purely nomothetic-span (individual-differences) approach. The construct representation approach to construct validity, in contrast, calls for researchers to examine the psychological processes that cause response variation in sustained attention measures, often via experiment (Borsboom et al., 2004; Embretson, 1983). From this perspective, the measurement question becomes focused on the task parameters that might be manipulated to cause changes in both RT variability and TUT rates. The present study marries the construct-representation (experimental) and nomothetic span (correlational) approaches to further ask whether experimental manipulations of tasks' sustained attention demands can also change the *covariation* in RT variability and TUT rates, and thus our measurement of general sustained attention ability.

Construct Representation Evidence for Construct Validity

Our pursuit of construct representation evidence for the validity of attention consistency measures is not entirely new. Previous experimental research has examined how task manipulations, such as changes in task parameters (e.g., stimulus pacing), or situational contexts (e.g., providing performance incentives), alter RT variability or TUT rates during simple laboratory attention tasks.

Manipulations of Stimulus Pacing and Expectancy

Successfully sustaining attention requires being optimally ready to appropriately respond, but not all actions can be performed at predictable moments. One way to alter the task demand for sustained attention, then, is to manipulate stimulus expectancy via inter-stimulus intervals (ISIs). Faster-paced tasks (i.e., shorter ISIs) may minimize demands on sustained attention and improve performance because they exogenously scaffold participants' attention to the task (De Jong et al., 1999; Langner & Eickhoff, 2013; Shaw et al., 2012). Indeed, more fast-paced attention tasks, with shorter ISIs, elicit faster and less variable responding (Massar et al., 2020; Unsworth et al., 2018). Most studies of task pacing also show similar effects on TUT rates, with faster-paced tasks producing lower TUT rates compared to slower-paced tasks (Antrobus, 1968; Giambra, 1995; Grodsky & Giambra, 1991; Smallwood et al., 2008; Teasdale et al., 1993; Unsworth & Robison, 2020). Faster-paced tasks may therefore reduce sustained attention demands enough to change the individual-differences overlap between objective and subjective measures.

Expectancies can also be manipulated by varying or fixing the interstimulus interval (ISI) across a task. Predictable tasks (i.e., those with a fixed ISI) make less demand on sustained attention because participants know when each trial will begin, or when each response should be initiated. Indeed, RT variability on predictable tasks is typically reduced compared to that on tasks that vary the ISI from trial-to-trial (Langner & Eickhoff, 2013; Massar et al., 2020; Shaw et al., 2012; Unsworth et al., 2018). For example, Unsworth and Robison (2020; Exp. 2) compared performance on the psychomotor vigilance task (PVT) with a fixed 2 s ISI versus the standard PVT with ISIs that varied across trials from 1–10 s. The fixed condition produced not only faster RTs, but also fewer "lapses" (i.e., RTs > 500ms) and a smaller tail of the RT distribution, both of which may reflect momentary failures of sustained attention.

The effects of trial expectancy on TUTs are less clear than those on performance indicators. Massar et al. (2020, Exp. 2) varied whether participants received a *mostly short*-ISI PVT, or a *mostly long*-ISI PVT, both of which presented thought probes. Participants, here, had some expectancy of the trial onset, but with some uncertainty. TUT rates were similar across conditions, whereas performance measures of sustained attention differed as expected. A similar pattern of null TUT-report results was reported by Hawkins et al. (2019; Experiment 2), who fixed or randomly varied the ISI during a sustained attention to response task (SART; a go/no-go task) that also presented periodic thought probes. Mean TUT ratings were nearly identical (M_{fixed} = 2.70 vs. M_{random} = 2.58 on a 1 – 5 scale). Thus, whereas stimulus pacing appears to reliably affect both objective and subjective sustained attention indicators, trial expectancy might only impact objective measures. To better understand how these task parameters relate to the sustained attention construct, rather than to the process-impure measures of either RT variability or TUT reports, it is necessary to investigate how they influence the covariation between measures of RT variability and TUT rates.

Manipulations of Trial Type Frequency

Go/no-go tasks that use high go-trial frequency, like the SART and gradual-onset continuous performance task (gradCPT; Rosenberg et al., 2013), require sustained attention for consistent and accurate responding on repetitive go trials (in addition to appropriately withholding responses to no-go trials). Failing to sustain attention will result in habitual and sometimes mindless (and thus inconsistent) responding to go trials. Changing the frequency of specific trial types can thus reduce the demand on sustained attention by minimizing the time participants engage in repetitive responding. For example, in go/no-go tasks that had go-trial frequencies ranging from 20–80%, go-trial mean RTs were longer, and no-go accuracy was higher, as go-trial frequency decreased (Bedi et al., 2022; Jones et al., 2002; Nieuwenhuis et al., 2003; Young et al., 2018; but see Wessel, 2016). Objective indicators of attention consistency improved, then, as participants were given less opportunity to engage in long periods of mindless responding.

However, in versions of go/no-go tasks that alter trial-type frequency, performance changes might not always be due to changes in sustained attention demands, but rather to changes in response strategy (e.g., speed-accuracy trade-off; Dang et al., 2018; Head & Helton, 2014; Helton, 2009; Helton et al., 2010; Mensen et al., 2022; or response biases, Bedi et al., 2022). Wilson et al. (2016) replicated prior findings that, as no-go-trial frequency decreased, commission errors on no-go trials increased. They argued, however, that if this trend were due to failures of sustained attention, it should be reflected also in increased TUT reports (which was assessed in a post-task questionnaire). In fact, TUT frequency ratings *decreased* rather than increased as no-go frequency decreased. Wilson et al. (2016) therefore argued that these manipulations of SART performance reflected speed-accuracy trade-offs rather than sustained attention changes (Peebles & Bothell, 2004; Wilson et al., 2015). Future work on this question should use thought probes to measure TUTs in the moment, rather than relying on potentially problematic retrospective reports of mind wandering (Kane, Smeekens et al. 2021).

Only a few other studies have investigated the effect of trial-type manipulations on TUT rates. In a vigilance task that altered response frequency (10% or 30% of trials required responses), participants reported fewer TUTs in the 30% condition (i.e., in the condition that had more “no-go” trials; Giambra, 1995), in line with the hypothesis that requiring participants to engage in moderately frequent responding keeps them engaged in the task and gives them less opportunity to mind-wander. Similarly, in a go/no-go task that manipulated the presentation of no-go frequency to 20% or 40%, participants exhibited fewer TUTs in the 40% no-go condition (Smallwood et al., 2007). Manipulations of trial-type frequency appear to provide generally positive evidence for the construct validity of sustained attention measures, although they may also affect non-sustained-attention processes and we have concerns about the retrospective ratings of Wilson et al. (2016) that may muddy the overall conclusion.

Manipulations of Motivational State

Motivation manipulations should change how individuals use or engage their sustained attention abilities during a task. Participants’ ability (and willingness) to sustain attention should improve, at least momentarily, by experimentally increasing their motivation levels or engaging their interest. In fact, manipulations of motivational state significantly improve objective performance measures of attention consistency (e.g., Chiew & Braver, 2013; Esterman et al., 2014; Locke & Braver, 2008; Tomporowski & Tinsley, 1996). Seli et al. (2019), for example, had participants complete a metronome response task (MRT). During the MRT, participants try to press a key in sync with the presentation of an auditory or visual stimulus over an extended period. Sustained attention is required for participants to maintain maximally consistent responding over the course of many minutes. Seli et al. (2019) found that motivating participants, by telling them they could leave the experimental session early for good performance, resulted in less variable and more accurate MRT performance compared to a control condition.

Robison et al. (2021) manipulated a variety of motivation-related variables (providing specific goals or feedback, or telling participants they could leave early) while participants completed the psychomotor vigilance task (PVT), which requires participants to press a key to stop a counter, much like a stopwatch. The PVT taxes sustained attention by presenting a variable—and often long—waiting period for the onset of the counter. If attention momentarily wanes during the waiting period, participants will miss the start of the counter and produce a longer-than-normal RT. Robison et al. (2021; see also Unsworth et al., 2022) found that many of the motivation conditions significantly improved the behavioral indicators of attention consistency (e.g., reducing RT for the slowest 20% of trials) versus control conditions.

TUT rates also change with experimental manipulations of motivation. Specifically, compared to control conditions, monetary rewards or time-based incentives elicit lower TUT rates in a variety of tasks, including in the Seli et al. (2019) and Robison et al. (2021) studies described above (see also Antrobus et al., 1966; Smallwood et al., 2007; Unsworth et al., 2022). Collectively, then, inducing motivation or providing incentives are promising manipulations of both objective and subjective attention consistency measures.

Evidence Summary

Many of the reviewed experimental studies demonstrate that manipulating the demands on sustained attention produces predicted changes in both objective and subjective indicators of attention consistency. Specifically, manipulations of stimulus pacing and participants' motivational state provide strong evidence that changes in the need for sustained attention can alter RT variability and TUT rates. Manipulations of trial-type frequency provide more mixed evidence. We note, however, that many of the reviewed studies focused mainly on one aspect of measuring sustained attention, either with objective (performance variability) or subjective (TUT self-report) indicators. A possible—and potentially promising—way forward is to assess how experimental manipulations impact the individual-differences covariation in subjective and objective measures of sustained attention.

Goals and Hypotheses

The present studies used a combined experimental–correlational approach to examine whether manipulations of theoretically relevant task parameters would reduce the association between objective and subjective measures of attention consistency. Whereas the prior literature has relied on changes in either objective or subjective measures as indication of construct validity, we further asked whether experimental manipulations also alter the covariation of these measures, which is arguably a more construct valid way to measure sustained attention ability than is either measurement approach alone (Welhaf & Kane, 2022). This combined approach draws on the strengths of both the nomothetic span (i.e., individual-differences) and construct representation (i.e., experimental) approaches to construct validation.

Our primary, preregistered hypotheses were as follows: (a) For both Studies 1 and 2, if minimizing sustained-attention demands reduces the individual-differences overlap between objective and subjective attention consistency indicators—because variation in each is now more influenced by extraneous processes and abilities—then it should weaken the correlations between objective (RT variability) and subjective (TUT rate) indicators, compared to maximizing sustained-attention demands; (b) For Study 1 only, minimizing the need for sustained attention should also weaken the association between RT variability measures across tasks. That is, the between-task correlation between RT variability measures should be significantly stronger between two maximized-demand conditions than between a maximized- and minimized-sustained attention conditions. If sustained attention demands are effectively reduced, then any remaining correlation between RT variability measures must reflect individual differences in non-sustained-attention processes that also contribute to RT variability, like speed of processing or response strategies (e.g., speed-accuracy trade-offs); (c) For Study 1 only, minimizing the need for sustained attention should also weaken the between-task correlation between TUT rates. That is, the correlation between probed TUT rates should be significantly stronger between the two maximized-demand conditions than between a maximized- and minimized-sustained attention conditions. If sustained attention demands are effectively reduced, any remaining correlation between TUT rates must reflect individual differences in non-sustained attention processes that also contribute to TUT reports (i.e., reporting biases or demand characteristics).

Study 1

Study 1 examined whether experimentally reducing the sustained attention demands in three prototypical sustained attention tasks also reduces the correlations between objective (i.e., RT variability) and subjective (i.e., TUT rates) measures from these tasks versus their standard versions that maximize demand ("Standard" and "Maximized-SA" tasks). Reducing sustained attention demands should similarly affect the correlations among the performance measures from the tasks and among the TUT rates from the tasks. That is, minimizing the need for sustained attention in those tasks ("Minimized-SA" tasks) should reduce the proportion of individual-differences variance attributed to sustained attention ability—thus weakening any correlations driven by sustained attention processes and abilities—and increase the contributions of non-sustained-attention-related processes to the variance.

To test this hypothesis, we had each subject complete three different prototypical sustained attention tasks (for more details see the *Tasks* section below) that varied in sustained attention demand. As the COVID-19 pandemic forced data collection to occur online, we prioritized using within-subject manipulations so that any between-subject variation in environmental events or other contextual effects would not contribute to the comparisons of interest. Further, using multiple tasks (rather than one repeated task, e.g., three versions of the PVT) allowed for greater generalizability of our conclusions; it also minimized concerns about vigilance decrements across successive versions of the same task or any development of strategies that might occur due to extended practice of the same task.

Methods

We report our sample size justification and data exclusion criteria, as well as all measures and manipulations included in the study (Simmons et al., 2012).

Participants

As stated in our preregistration, we aimed to collect data from 1500 participants via the recruitment site Prolific Academic (https://www.prolific.co). We based this sample size on several calculations. First, we conducted a power analysis in G*power for differences between dependent correlations with a common index. For a one-tailed test, alpha of .05, and the correlations of interest being .30 (Maximized-SA × Standard) versus .20 (Minimized-SA × Standard), with the remaining correlation (Maximized-SA × Minimized-SA) of .20, we would have just under 95% power to detect the difference between correlations with N = 1500. Second, to achieve a 95% CI around a *r* = .30 correlation with a lower bound above .25 (for Maximized-SA × Standard), and around a *r* = .20 correlation with an upper bound below .25 (for Minimized-SA × Standard), requires 1420 participants. Finally, with a sample of 1500, we would be able to interpret correlations of .30 (Maximized-SA × Standard) and .25 (Minimized-SA × Standard) as statistically equivalent via a Two One-Sided Tests equivalence test for dependent overlapping correlations (per formulas in Counsell & Cribble, 2014).

Some participants (age restriction 18–40 years) were recruited via Prolific Academic; eligibility required living in the U.S., U.K., Ireland, Canada, New Zealand, or Australia, reporting English as a first language, earning at least a high school degree or its equivalent, and having a ≥ 95% study approval rating. Participants were paid $7.13 for the 45-min study.

Other participants were recruited from UNCG (age restriction 18–35 years), via the introductory psychology research pool. Participants had to indicate English was their first language. They were awarded partial credit towards an introductory psychology research-participation requirement.

Apparatus and Materials

All tasks and questionnaires were programmed in Gorilla (https://gorilla.sc). The experiment can be accessed via Gorilla's Open Materials (https://app.gorilla.sc/openmaterials/388985). Participants were required to complete the study on a laptop or desktop computer to ensure accurate recording of response times (Anwyl-Irvine et al., 2019, 2020).

Tasks

Below we describe the two groups of tasks used in the current study (See Table 17 for task order in each counterbalancing condition; for details, see *Procedure*), which varied in sustained-attention demand.

Table 17. Task order by counterbalancing condition for Studies 1 and 2.

Study 1					
Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
PVT$_{\text{min}}$	SART$_{\text{min}}$	vMRT$_{\text{min}}$	*vMRT*$_{max}$	*PVT*$_{max}$	*SART*$_{max}$
SART$_{stand}$	*vMRT*$_{stand}$	*PVT*$_{stand}$	*SART*$_{stand}$	*vMRT*$_{stand}$	*PVT*$_{stand}$
vMRT$_{max}$	*PVT*$_{max}$	*SART*$_{max}$	PVT$_{\text{min}}$	SART$_{\text{min}}$	vMRT$_{\text{min}}$
Study 2					
Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
PVT$_{\text{min}}$	SART$_{\text{min}}$	vMRT$_{\text{min}}$	vMRT$_{\text{max}}$	PVT$_{\text{max}}$	SART$_{\text{max}}$
vMRT$_{\text{max}}$	PVT$_{\text{max}}$	SART$_{\text{max}}$	PVT$_{\text{min}}$	SART$_{\text{min}}$	vMRT$_{\text{min}}$

Note. For Study 1, the top row indicates participants’ first task, the second row their second task, and the bottom row their third task. For Study 2, the top row indicates the participants’ first task, and the bottom row is their second task min = minimized task; stand = standard task; max = maximized task; PVT = Psychomotor Vigilance Task; SART = Sustained Attention to Response Task; vMRT = visual Metronome Response Task. In Study 1, italicized tasks were followed by DSSQ items.

Standard/Maximized Sustained Attention Tasks. The standard and maximized versions were identical to each other. They were designed to induce significant sustained attention demands by requiring participants to maintain focus over extended, and sometimes unpredictable, periods of time to the presentation of repetitive stimuli that, in some cases, promoted habitual responding. Failure to sustain attention in these task versions would result in more erroneous and more variable performance. Although our standard and maximized tasks were formally the same, for purposes of all analyses we defined the standard task as the second task in the sequence, and the maximized task as either the first or third task in the sequence, depending on the order condition (where the maximized task is presented first, the minimized task is presented last, and vice versa).

Sustained Attention to Response Task (SART). In this go/no-go task, participants were instructed to press the space bar for words from a target category (*animals*; 90% of total trials) while withholding responses to another (*crops,* i.e., fruits/vegetables; 10% of total trials). Participants first completed 10 practice trials by responding to boy's names and withholding responses to girls' names. The real task began with 16 unanalyzed buffer trials. Each stimulus word was presented for 272 ms, followed by a mask (XXXXXXXXXX) which was presented for 1224 ms. Participants were instructed to press the spacebar during either the word or the mask.

After being presented with task instructions, participants answered a quiz question about the SART: "*When performing this task, which words should you NOT press any key for?*" with the following response options, 1) *animal names*, 2) *girl names*, 3) *crop names*, 4) *boy names*. Participants pressed the number key corresponding to the correct answer (Option 3). If participants answered the questions incorrectly, they were reminded of the task instructions and then were able to answer the question again to continue. Participants completed this instruction quiz for whichever version of the SART they completed (SA-Maximized, Standard, or SA-Minimized).

Participants completed 480 trials divided into four seamless blocks of 120. In each block, stimuli were comprised of 36 animal names (i.e., “go” trials) and 4 fruit/vegetable names (i.e., “no-go” trials). Each stimulus was pseudorandomly presented 3 times, thus each block contained a total of 108 “go” animal stimuli and 12 “no-go” crop stimuli. Each block contained a new set of stimulus words. Each block of 120 response trials and 6 probe trials was randomized and presented to all participants in the same order. The dependent measure for the SART was the within-subject standard deviation (SD) of RTs to correct “go” (animal) trials.

Psychomotor Vigilance Task (PVT). On each trial of this task, participants saw a set of zeros (oo.ooo) in the center of a white screen. Unpredictably (at SOAs from 1000–10000 ms, in 1000-ms increments), the zeros began counting upward in milliseconds. Participants were required to press the spacebar as soon as they noticed the numbers were counting upward to stop them. The numbers then stopped and were displayed for RT feedback.

Participants completed 5 practice trials; the real task began with 6 unanalyzed buffer trials (SOAs = 1000, 2000, 5000, 6000, 9000, 10000 ms). Participants completed two seamless blocks of 45 trials for a total of 90 trials (9 at each SOA). Eighty of these trials were performance trials in which to zeros began counting-up after the SOA. The other 10 “yoked” trials served as thought probe trials, where the probe appeared at the completion of the SOA, rather than the digits counting up (see *Thought Probes* section below). Each block of 40 response trials and 5 probe trials was separately randomized and presented to all participants in the same order (the only randomization constraint was that probe trials could not appear within 3 trials of another probe). SOAs were allowed to repeat on consecutive trials.

If participants pressed the spacebar before the numbers started counting upward, they were presented with an error message ("Do not press the spacebar before the numbers start!"). These anticipation error trials were repeated at the end of the last block of the task. The primary DV for all PVTs was the mean RT of the slowest 20% of trials.[13]

Visual Metronome Response Task (vMRT). In this task, participants were presented with the regular onset and offset of a black square in the center of a white screen. The goal for this task was to respond in synchrony with the onset of the black square by pressing the spacebar. Each trial began with a blank screen presented for 650 ms, followed by the black square presented for 150 ms, then another blank screen presented for 500 ms (Laflamme et al., 2018). Thus, from the subject's perspective, a single trial lasted 1300 ms (with 1150 ms intervals between squares).

Participants first completed a practice block of 20 trials. For the real trials, participants first completed 6 unanalyzed buffer trials followed by 420 trials divided into 12 seamless blocks. Each block of 35 response trials and one probe trial was individually randomized and presented to all participants in the same order. Probe trials appeared immediately following the 500 ms

[13] We preregistered also calculating a second commonly used DV in the PVT, the number of lapses (i.e., RTs > 500ms; Lim & Dinges, 2008) and using the variable with the better measurement characteristics. Both DVs showed similar split-half reliabilities (Slowest 20% Cronbach α = .98; Lapses Cronbach α = .96). Note this split-half method reflects a deviation from the preregistered reliability assessment, but it is consistent with how we and others have investigated reliability in the PVT (e.g., Unsworth et al., 2021; Welhaf & Kane, 2022). We therefore focus our primary analyses with the mean RT of Slowest 20% of trials as the primary DV (as we have done in previous research with the PVT; Welhaf & Kane, 2022). We report parallel results with lapses as the DV in Supplemental Tables S19 and S20.

blank screen of the previous trial. The dependent measure for this task is variability in the rhythmic response time (RRT). RRTs are calculated as the difference between response and stimulus onset (Laflamme et al., 2018; Seli, Cheyne, et al., 2013) and can be positive (responding after the stimulus appears) or negative (responding before the stimulus appears); consistent with prior studies, we calculated overall mean RRT variability using a moving window of the current and previous four trials across all trials of the task and then took the log of that value (see also Seli, Jonker et al., 2015).

Minimized ("min") Sustained Attention Tasks. In the demand-minimized versions of the tasks, we aimed to decrease sustained attention demands by including task breaks occupied by a separate task and altering stimulus presentation rates or target frequency (or both).

Continuous Temporal Expectancy Task (CTET; O'Connell et al., 2009). The CTET was used as a brief break activity between trial blocks in the minimized sustained attention tasks, to reduce the vigilance demands of those tasks. Participants viewed a stream of abstract, geometric images. Most stimuli (non-targets) appeared onscreen for a brief duration (600 ms); occasionally, target stimuli were presented for a longer duration (1200 ms). Participants reported these infrequent targets by pressing the spacebar.

Participants first saw a brief example of what the short- and long-duration images looked like and then completed 20 practice trials (5 of each image, with each image once as a target). Participants completed one block (60 trials; 6 targets) of the CTET at the following times in the minimized tasks: PVT—after the first 30 and 60 trials (2 blocks of CTET); vMRT—after every 144 trials, aside from the last (2 blocks of CTET); SART—after every 120 trials, aside from the last (3 blocks of CTET). Stimuli were randomized and presented to all participants in the same order, with the only constraints being the same image could not repeat immediately and targets had to occur at least 4 trials apart. Each image appeared as a target three times across all CTET blocks in the PVT_{min} and vMRT_{min} tasks and three or four times in the SART_{min}.

***$\textbf{SART}_{\textbf{min}}$*.** Participants responded to animal names while withholding responses to crop (fruit and vegetable) names. The ratio of go to no-go trials was shifted to 65/35% (from 90/10%), thus increasing no-go target probability to support participants' focus on the task (e.g., Smallwood et al., 2007; Wilson et al., 2016). Participants completed 4 blocks of 120 trials, with 26 unique animals and 14 unique crops per block. Within each block, each stimulus appeared 3 times. Each block of 120 response trials (including 6 probed trials) was separately randomized and presented to all participants in the same order.

As a second method to reduce active sustained-attention demands, we also increased stimulus presentation rate. Each stimulus word was again presented for 272 ms but the mask was reduced from 1224 to 935 ms.

Participants completed 10 practice trials responding to boy's names and withholding responses to girls' names. The real trials began with 10 unanalyzed buffer "go" trials. To account for potential post-restart RT costs after each CTET block (e.g., Gopher et al., 2000), we included 2 unanalyzed buffer "go" trials at the start of every post-CTET block (6 total).

***$\textbf{PVT}_{\textbf{min}}$*.** The critical manipulation to reduce sustained-attention demands (in addition to interpolating blocks with CTET) was reducing the SOAs and their range; the distribution of SOAs was shifted to 2100–3000 ms (in 100-ms intervals; vs. 1000–10000 ms in maximized/standard tasks). As in the maximized and standard tasks, participants completed 90 trials (9 at each SOA); 80 were performance trials and 10 were probe trials; SOAs could repeat on consecutive trials. Each block of 40 response trials and 5 probe trials was individually randomized and presented to all participants in the same order.

If participants pressed the spacebar before the numbers started counting upward, they were presented with an error message ("Do not press the spacebar before the numbers start!"). These anticipation error trials were repeated at the end of the last block of the task.

To match the standard PVT and account for restart RT costs after the CTET breaks, we included six unanalyzed buffer trials, two at the beginning of the task, and two following each CTET break.

***vMRT$_{min}$*.** The critical manipulation to reduce sustained attention demands (in addition to interpolating blocks with the CTET) was speeding the target presentation rate. Each trial in this minimized version was reduced from 1300 ms to 800 ms. Each trial began with a 400 ms blank screen, followed by 150 ms presentation of the black square, and ending with another 250 ms blank screen. Participants again completed 420 trials across 12 blocks. Each block of 35 response trials and one probe trial was separately randomized and presented to all participants in the same order. To match the standard vMRT and account for restart RT costs after the CTET breaks, we included six unanalyzed buffer trials, two at the beginning of the task, and two following each break; CTET blocks followed every 144 critical vMRT trials (every 4 blocks). The dependent measure for this task is the RRT variance, calculated in the same way as in the standard/maximized version of the task.

Thought Probes

In each task, participants occasionally reported on their immediately preceding thoughts by responding to unpredictably presented probes. Before beginning the tasks, participants read the following instructions about responding to each probe:

> During this study's tasks, we will occasionally ask what you were just thinking about. It's normal to sometimes lose focus during computer tasks.

> We want to know what you think about. The task will sometimes stop to ask what you were just thinking about, in the instant before we asked. You will see a thought-choice menu. Please take stock of what your thoughts just were and choose the closest description.

Participants then saw the thought-choice menu, which asked, “What were you just thinking about?” and had participants “Please press a number on the keyboard” that represented their thoughts. Instructions then provided descriptions of the 6 probe response options: *1. the task / task performance*, thinking about the computer task, or about how well (or poorly) you are doing on it; *2. everyday things / personal worries,* thoughts were about normal, routine things you did recently or that you'll be doing later, or about big or small life concerns or worries; *3. current state of being,* thinking about your state of mind or feelings, such as thinking about being sleepy, cheerful, hungry, curious, or bored; *4. daydreams / fantasies,* fantasies or thoughts disconnected from reality, like thoughts about flying, or being at the beach; *5. external environment,* thinking about something in your environment, other than this task, like sights or sounds in the room; *6. Other,* only if thoughts don’t fit the other options. During the tasks, thought-probe screens only included the above italicized labels. The TUT dependent measure was the proportion of probe responses 2 to 6.

SART-TUT. Probes followed six no-go trials in each block, for 24 probes total.

PVT-TUT. Probes appeared once after each SOA delay, for 10 probes total. In the maximized and standard tasks, the first block presented probes after SOAs of 2000, 4000, 6000, 8000, and 10000 ms, whereas the second block presented probes after SOAs of 1000, 3000, 5000, 7000, and 9000 ms. In the minimized task, the first block presented probes after SOAs of

2200, 2400, 2600, 2800, and 30000 ms, whereas the second block presented probes after SOAs of 2100, 2300, 2500, 2700, and 2900 ms.

vMRT-TUT. Within each block, participants were presented with one thought probe, for a total of 12 probes.

Questionnaires

Task Motivation (Dundee State Stress Questionnaire; DSSQ). Following completion of the maximized and standard sustained attention tasks, participants completed a shortened, 7-item version of the DSSQ–Motivation subscale (Matthews et al., 2002). Examples of items included, "I was eager to do well," "I didn't really care about my performance," and "I would have rather spent my time doing something else other than this task." All items were presented on the screen at the same time and each item was rated on a Likert scale with the following options: 1) *Not at all,* 2) *A little bit,* 3) *Somewhat,* 4) *Very much,* 5) *Extremely*. Note that only the italicized response options appeared on-screen during the DSSQ, numeric values were used for analyses. Participants responded via mouse-click to each item. Each presentation of the DSSQ contained the same motivation items. The score on the DSSQ was the sum of the item ratings, after appropriate reverse-scoring.

Within each of the two DSSQ presentations, we added a negatively worded version of one item ("I would have rather spent my time doing something else other than this task," was reworded to, "<u>I would have rather spent my time doing this task than almost any other task.</u>") One of these two contradictory items was presented at the beginning (i.e., item 1) and one at the end (i.e., item 7) of the DSSQ. These items were used both as scored DSSQ items and as an attention check to assess consistency in responding to the questionnaire items.

Post-Study Questionnaires. Following completion of all tasks, participants answered several questions about their experiences during the study. As an open-ended "bot" check item, we asked participants to describe which task in the study they found the most challenging. Next, participants answered Likert-scale questions about how noisy their immediate environment was while completing the study (i.e., *How noisy (people, music, TV) was your immediate environment while completing this study? Not at all noisy, Slightly noisy, Moderately noisy, Extremely noisy)* and how distracted they were by their immediate environment *(i.e., How distracted were you by things in your immediate environment while completing the study? Not at all distracted, slightly distracted, moderately distracted, extremely distracted).* We also asked participants about their media multitasking frequencies during the study (i.e., *During this study, how often did you interact with: phone (calls or texts); email or social media; video games*); each of the three multitasking questions was rated on the same scale: *Never, Some of the time, Most of the time, All of the time*. Finally, participants reported on their sleepiness at the beginning of the study (i.e., *How sleepy were you at the beginning of this study? Not at all sleepy, Slightly sleepy, Moderately sleepy, Extremely sleepy)*.

Procedures

To best sample from the different geographic locations we selected, time slots on Prolific were posted at different times throughout the day. Specifically, we typically posted slots ranging from 7:00 am (EST) to 9:00 pm (EST), with the hope that earlier time slots would capture European participants and later timeslots would capture participants from Australia/New Zealand. Upon accepting the study on Prolific, subjects were directed to Gorilla to complete the experiment.

Participants first provided informed consent and then answered two unrelated "botcha" questions to flag potential bots. If participants failed both questions they were directed out of the study (this did not occur for any participants). Participants then completed a demographics questionnaire, then they read the initial study and thought-probe instructions. At the end of the initial study instructions and learning about the thought probe details, participants were asked a single comprehension question regarding the thought-probe screens to ensure they were paying attention to the instruction. Specifically, the question asked, "*When responding to questions about your thoughts, what time frame should you report your thoughts from?* " with the following response options: *1) Since the very beginning of the task, 2) The moment right before the thought menu appeared, 3) Over the last 30-60 seconds , 4) Since the last time a thought-menu appeared*. Participants pressed the key corresponding to the correct answer (Option 2). If participants failed this question, they were shown a separate screen that told them they were wrong and were reminded to reminded of the appropriate time frame. After reviewing this information, participants were given another chance to answer.

Following these initial instructions, participants were randomly assigned to one of six experimental (counterbalancing) conditions (Table 17). Participants were assigned sequentially to each condition until each condition had a participant and then the assignment started over again (i.e., the first participant was assigned to Condition 1, the second to Condition 2… the seventh participant was then assigned to Condition 1, etc). In each condition, participants completed one "maximized," one "standard," and one "minimized" sustained attention task; for each participant, one of these tasks was the SART, one was the PVT, and one was the vMRT. As noted above, the maximized and standard versions of all tasks were identical, but the task we identified as "standard" was always presented as the second of the three tasks; the maximized and minimized appeared as the first or third tasks, across counterbalancing orders (for 3 orders, the maximized appeared first, and for 3 orders, the maximized appeared last).

As noted in the description of the DSSQ above, participants completed motivation items following completion of each of the two standard/maximized tasks regardless of condition. Finally, participants completed the post-experiment questionnaire.

Preregistered Data Analysis Exclusions

We report the number of participants with data exclusions in the Results section. Participants were dropped case-wise from all analyses if they either: (a) responded "extremely" to two of the three post-experiment questions regarding their subjective state or immediate environment (i.e., noisy, distracted, and/or sleepy); or (b) responded to any of the media-multitasking questions with "Most of the time" or "All of the time".

We dropped DSSQ questionnaire data (but not performance or TUT data) if participants failed to respond appropriately to the consistency items ("I would have rather spent my time doing something else other than this task;" "I would have rather spent my time doing this task than almost any other task"). Both items were reported on the same scale (e.g., not at all, a little bit, somewhat, very much, extremely). After reverse-scoring the negatively worded item, we dropped DSSQ data for participants who did not respond within one response option on both items.

Preregistered RT Cleaning Procedures

Before scoring the main DV in each sustained attention task, we removed trials following thought probes. Additionally, in both versions of the SART, we removed post-error trials and trials that were faster than 200 ms (which reflect anticipatory responses). In the vMRTs, we removed trials following omissions. Also consistent with prior work using an auditory version of the MRT, we removed all data for participants who had >15% omission errors (Seli, Cheyne et al., 2013). For the PVTs, we removed trials that were faster than 200 ms.

Following these exclusions, we calculated, for each subject for each task, their Median RT and a cutoff value equivalent to 3*IQR (for the MRT, this cutoff was created before calculation of the RRT at the individual trial-level). Values exceeding this threshold were replaced with values equivalent to the threshold, as they represented excessively slow responses that were likely not caused by failures of sustained attention (e.g., sneezing, blinking, stretching).

Upon calculating the main DV for each task, participants whose value was outside 3*IQR of the sample were censored to a value equal to the 3*IQR value.

Non-Preregistered Thought-Probe Exclusions

Although we did not preregister data exclusions based on thought-probe RTs, upon preliminary data screening, we found that some participants spent a surprisingly long time on thought probes (much longer than laboratory participants do, in our extensive experience). We were concerned that these participants were likely not continuously participating in the task; instead, they were likely using the probes as opportunities to take breaks. In these instances, participants' sustained attention performance and abilities were likely not adequately measured in one sitting; they may have forgotten task instructions or may have only restarted tasks when they were feeling refreshed. We therefore deviated from the preregistered data exclusion plan by screening for probe RT before calculating any performance measures.

The probe RT screening procedures were as follows: (1) remove all data from any subject who had a probe RT (in any task) longer than 5 min; (2) for remaining participants, calculate the number of probes (in each task) with RTs > 15 s and set those probe responses as missing data; (3) for each subject and each task, calculate the number of remaining valid probes; (4) remove participants' data case-wise if they did not have at least 6 valid probes in each task. The criterion of 6 probes per task was decided based on a reanalysis of prior lab-based studies where we examined how many probes were needed to reliably estimate individual differences in TUT rate across multiple probed tasks (Welhaf et al., 2022); those analyses suggested that in a large-scale study with multiple probed tasks, reliabilities, factor loadings and correlations with common individual difference predictors like working memory capacity stabilized when estimating a TUT latent variable with just 6–8 probes.

Results

Below we report the results of our preregistered analyses and note where we deviated from the preregistered plan. All data aggregation and analyses were performed in R (R core team, 2020) using *tidyverse* (Wickham, 2019). ANOVAs were performed using the *afex* package (Singmann et al., 2020). Data visualizations were created using *ggplot2* (Wickham, 2016). Meta-analyses were conducted using the *meta* package (Balduzzi, Rücker & Scwarzer, 2019). Data and Rmarkdown files for all analyses are available on the Open Science Framework (https://osf.io/rm735/).

Exclusions Based on Preregistered Criteria and Thought Probe RTs

As detailed in the preregistration, we dropped data casewise from 23 participants for failing to provide a reasonable open-ended response to the post-study question about the most challenging task (e.g., "nothing," "idk," "3"). We dropped all data from an additional 24 participants for not passing the media-multitasking or subjective-state checks at the end of the study. Finally, we dropped all data from 35 participants for having > 15% omission errors in the vMRT. *Although not preregistered for Study 1* (preregistered only for the subsequent Study 2), we dropped data from an additional 44 participants for having SART "go" trial accuracy below

70% (as in Welhaf & Kane, 2022), as this might indicate failure to understand or comply with task instructions.

Although not preregistered for Study 1, we dropped all data from 82 participants based on our post-hoc concerns described above regarding invalid thought probe responses, given long probe RTs (we were otherwise blinded to the performance and thought-report responses for these participants): 58 participants had at least one thought probe RT of ≥ 5 min; an additional 24 participants did not have enough valid thought probes in each task after we screened out all probe responses with RTs > 15 seconds. Finally, we excluded DSSQ data (while retaining all performance and TUT report data) from 726 subjects for failing the consistency items. Given this large number of excluded participants, we will also report exploratory analyses that retain these participants.

Final Sample Demographics

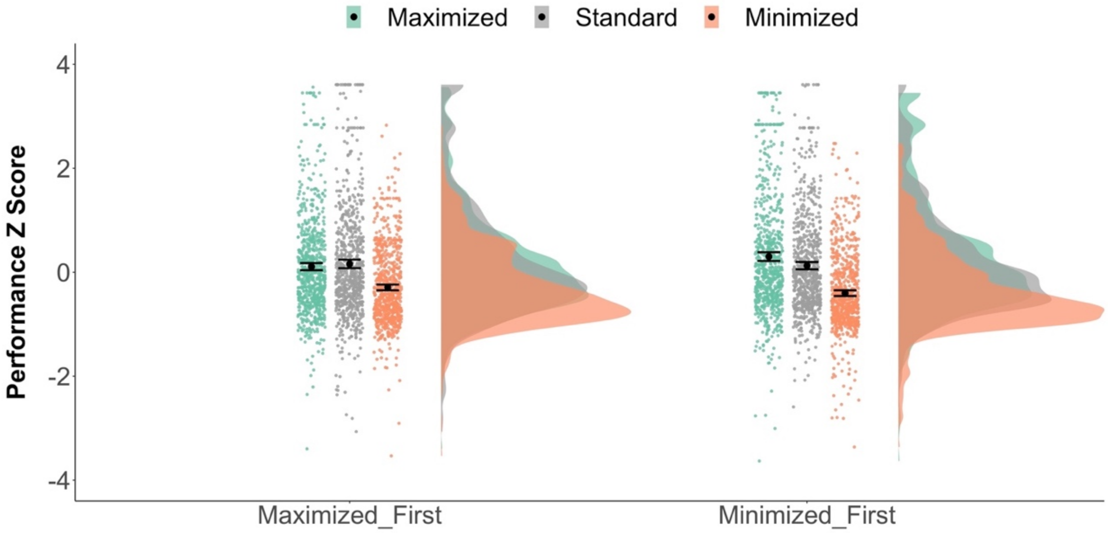
After implementing the above exclusions, our final sample consisted of 1470 participants. Of the final sample, 53.8% self-identified as female, 44.9% self-identified as male, and 1.3% self-identified non-binary or gender-nonconforming. Mean age was 26.7 years (SD = 6.4; 1 not reporting). The self-identified racial breakdown of our final sample was 69% White/European descent, 10% Black/African descent, 6% Multiracial, 6% South Asian descent, 6% East Asian descent, 3% Hispanic or Latino/Latina, 1% Middle-Eastern, Arab, or North African, and <1% Native Hawaiian or Pacific Islander, or Native American or Alaskan Native. Finally, the Prolific participants (*n* = 1278) showed a range of educational attainment: 28% had a high school diploma or A-level certificate, 12% earned a technical or community college degree, 42% earned an undergraduate degree, 17% earned a graduate degree, and 2% earned a doctorate (all 192 participants recruited from UNCG were undergraduates).

Experimental Effects of Sustained Attention Demand and Task Order

Supplemental Table S18 presents the descriptive statistics (before standardization) for RT variability measures and TUTs for each task, in each condition. All conditions had at least 240 participants with usable performance and TUT-rate data (Condition 1 = 249, Condition 2 = 241, Condition 3 = 246, Condition 4 = 246, Condition 5 = 242, Condition 6 = 246). Before assessing whether our manipulations of sustained attention demand altered the correlations between RT variability and TUTs, we first tested whether RT variability or TUT rates were impacted at an experimental level (*these analyses were not preregistered*). As the RT variability measures from each task were on different scales (e.g., milliseconds vs. log-transformed variance), we first z-scored the RT measures within each task (e.g., SART RTsd was z-scored across all SART task versions and orders). Lower z-scores reflected better performance (i.e., less RT variability and lower TUT rates).

RT Variability. The results of a 3 (Demand: Maximized vs. Standard vs. Minimized) × 2 (Order: Maximized First vs. Minimized First) ANOVA (with Greenhouse-Geisser correction for within-subject comparisons) on RT variability indicated a significant effect of Demand, $F(1.99, 2927.12) = 212.31$, $p < .001$, $\eta_p^2 = .126$, but no significant effect of Order, $F(1, 1468) = 0.19$, $p = .661$, $\eta_p^2 < .001$. However, these main effects were qualified by a significant (but small) Demand × Order interaction, $F(1.99, 2927.12) = 14.39$, $p < .001$, $\eta_p^2 = .010$. These results are visually depicted in Figure 12.

Figure 12. Raincloud plots (Allen et al., 2021) depicting z-scored RT variability between Demand conditions by Order, collapsed across all sustained-attention tasks for Study 1



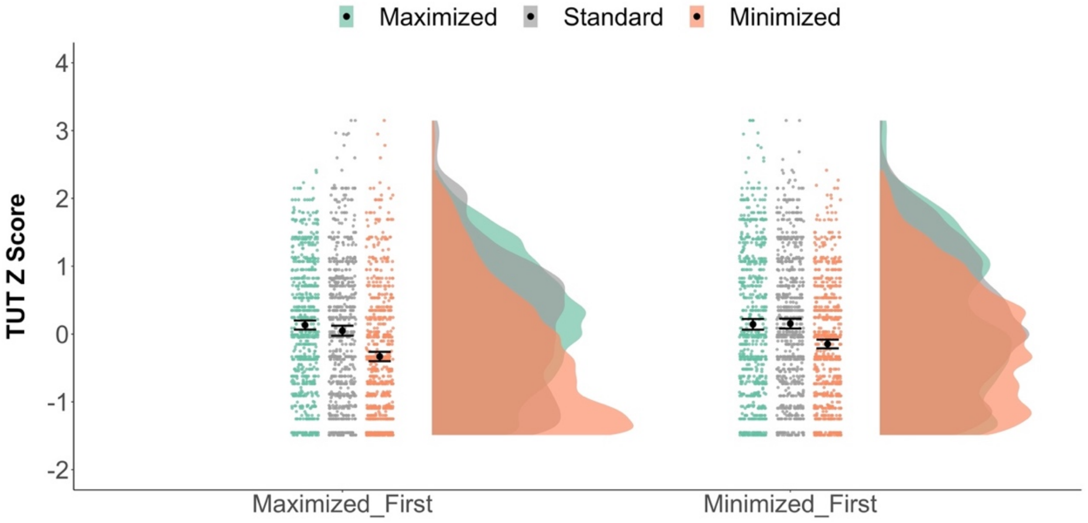
Note. Dots represent individual subject means in each condition. The closed black dots represent group-level mean estimates for each demand level. Error bars are 95% confidence intervals.

We followed up this significant interaction with separate one-way ANOVAs within each Order condition. Of most importance, the effect of Demand was significant in each: Maximized First, $F(1.93, 1424.91) = 69.11$, $p < .001$, $\eta_p^2 = .086$, and Minimized First, $F(1.96, 1438.05) = 160.18$, $p < .001$, $\eta_p^2 = .179$. Pairwise contrasts further specified where performance differences occurred: When the maximized task appeared first, there was no difference in performance between the maximized ($M = 0.11$) and standard task ($M = 0.16$), $t(733) = -1.200$, $p = .231$, Cohen's d [95% CI] = -.04 [-.12, .03]; however, performance was significantly worse in the standard than in the minimized task ($M = -0.29$), $t(733) = 10.238$, $p < .001$, $d = .38$ [.30, .45], and performance was significantly worse in the maximized than in the minimized task, $t(733) = 10.425$, $p < .001$, $d = .38$ [.31, .46].

When the minimized task appeared first, performance on the maximized task ($M = 0.30$) was significantly worse than in the standard task ($M = 0.13$) and the minimized task ($M = -0.40$): Maximized – Standard, $t(735) = 4.172$, $p < .001$, $d = .15$ [.08, .23]; Maximized – Minimized, *t*(735) = 16.354, *p* < .001), *d* = .60 [.52, .68]. The standard task also yielded significantly worse performance than did the minimized task, *t*(735) = 13.977, *p* < .001, *d* = .52 [.44, .59]. In general, then, our manipulation of demand significantly reduced RT variability across the demand levels, with maximized and standard tasks eliciting greater RT variability compared to the minimized task.

TUTs. We next conducted the same *non-preregistered* analyses on TUT rates. A 3 (Demand: Maximized vs. Standard vs. Minimized) × 2 (Order: Maximized First vs. Minimized First) ANOVA (with Greenhouse-Geisser correction for within-subject comparisons) indicated a significant effect of Demand, *F*(1.99, 2916.42) = 139.62, *p* < .001, η_p^2 = .087, and a significant effect of Order, *F*(1.99, 2916.42) = 5.56, *p* = .019, η_p^2 = .004. Again, these main effects were qualified by a significant (but small) Demand × Order interaction, *F*(1.99, 2916.42) = 6.17, *p* = .002, η_p^2 = .004. These results are visually depicted in Figure 13.

Figure 13. Raincloud plots depicting differences in z-scored TUT rates between Demand conditions by Order, collapsed across all sustained attention tasks.



Note. Dots represent individual subject means in each condition. The closed black dots represent group-level mean estimates for each demand level. Error bars are 95% confidence intervals.

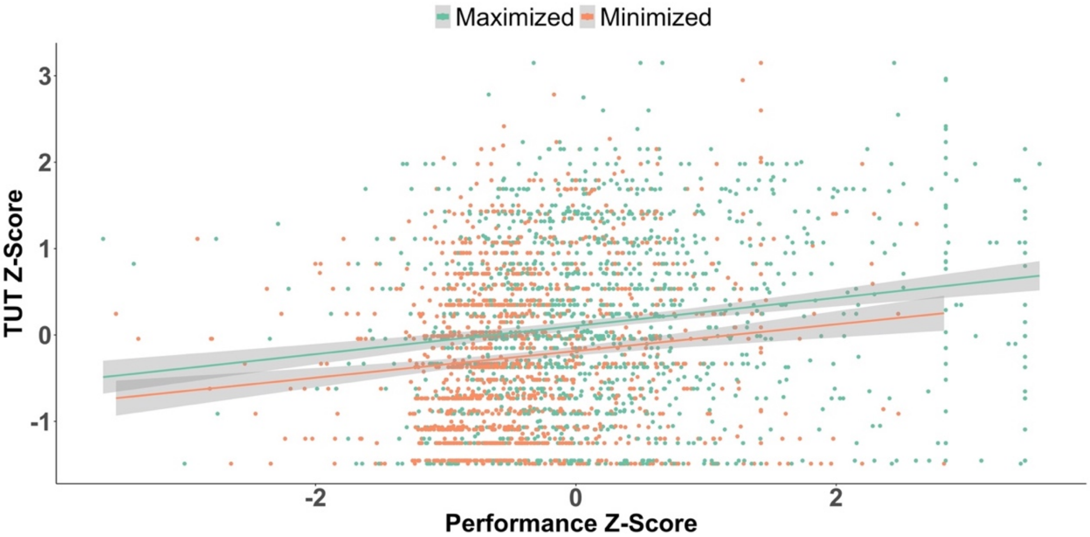
We again followed up this interaction with one-way ANOVAs in each Order condition. There was again a significant effect of Demand in each: Maximized First, *F*(1.99, 1460.46) = 94.92, $p < .001$, $\eta_p{}^2 = .115$, and Minimized First, *F*(1.95, 1436.85) = 48.92, $p < .001$, $\eta_p{}^2 = .062$. Pairwise contrasts further specified where differences in TUTs occurred: When the maximized task appeared first, there was a significant difference in TUT rates between the maximized (*M* = 0.13) and standard task (M = 0.05), *t*(733) = 2.390, *p* = .017, *d* = .09 [.02, .16]. However, TUT rates were higher in the standard than in the minimized task (*M* = -0.33), *t*(733) = 10.600, $p < .001$, *d* = .39 [.32, .47], as well as being higher in the maximized than in the minimized task, *t*(733) = 12.591, $p < .001$, *d* = .46 [.39, .54].

When the minimized task appeared first, there was again no significant difference in TUT rates between the maximized (*M* = 0.14) and standard task (*M* = 0.15), *t*(735) = -0.295, *p* = .768, *d* = -.01 [-.08, .06]. However, TUT rates were significantly higher in the standard than in the minimized tasks (*M* = -0.15), *t*(735) = 9.460, $p < .001$, *d* = .35 [.27, .42] and higher in the maximized than in the minimized task, $t(735) = 8.073$, $p < .001$, $d = .30$ [.22, .37]. In general, then, our manipulation of demand reduced TUT rates, as it did RT variability.

Does manipulating sustained attention demands alter the correlations between RT variability and TUT rates?

Our next set of analyses focused on the study's main question: If the individual-differences covariation in RT variability and TUT rate is a construct-valid measure of sustained attention, then by reducing the sustained attention demands in a task (making performance more reliant on non-sustained attention processes), we should reduce the correlation between RT variability and TUT rates.

Figure 14. Scatterplot depicting the correlation between z-scored RT variability and z-scored TUT rates in the Maximized and Minimized Tasks for Study 1.



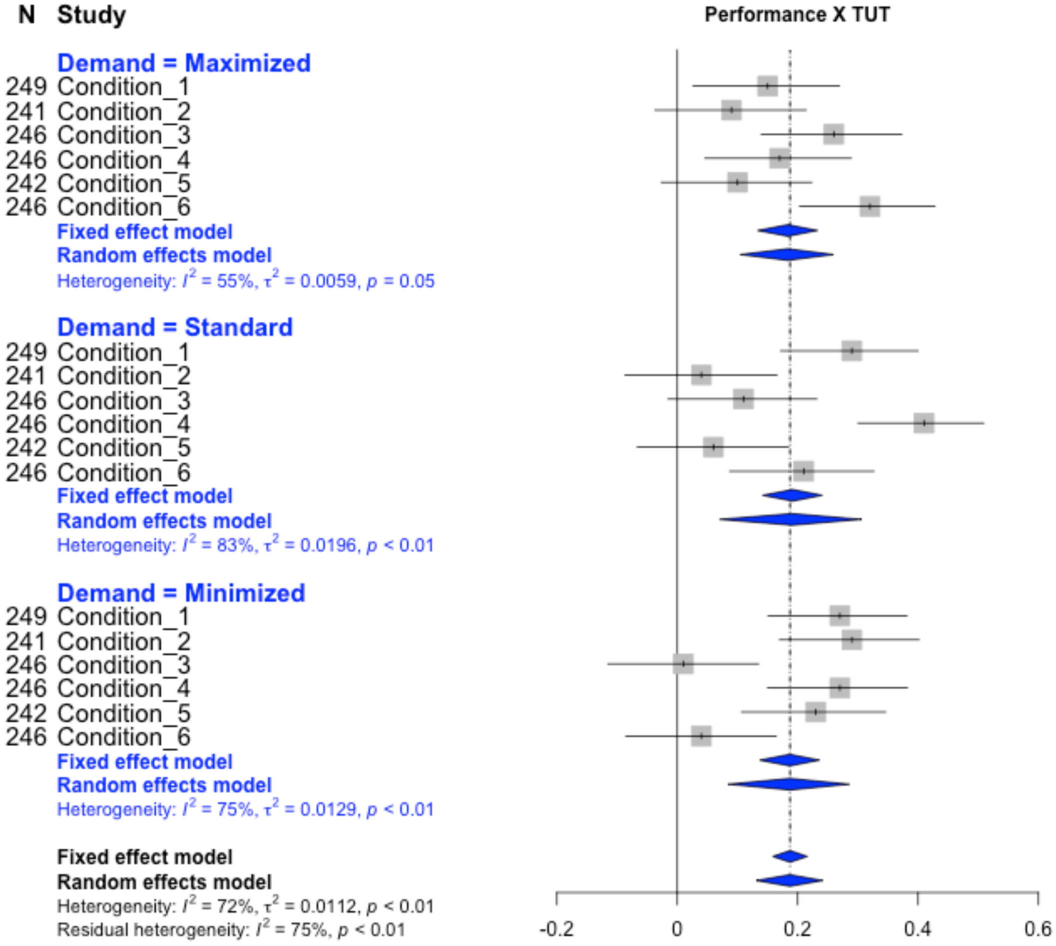
Note. Shaded areas represent the confidence interval around the regression line for each group.

As preregistered, we first correlated RT variability scores with TUT rates collapsed across all the maximized tasks and collapsed across all the minimized tasks. Figure 14 shows the scatter plots of these associations. First, RT variability and TUT rate were significantly, although weakly, correlated in both demand levels: Maximized *r*(1468) = .17 [95% CI: .12, .22], *p* < .001, Minimized *r*(1468) = .13 [.08, .18], *p* < .001.

Next, as preregistered, we compared these correlations using Steiger's test for non-overlapping correlations (one-tailed). The correlation in the maximized condition was not significantly larger than that in the minimized condition, z = 1.191, *p* = .117. Thus, our manipulations aimed at reducing the need for sustained attention in the minimized tasks did not significantly reduce the individual-differences overlap in the two proposed measures of attention consistency (despite eliciting the medium-sized experimental effects described above).

As a second preregistered approach to investigating the covariation of objective and subjective sustained attention measures, we conducted an internal meta-analysis to get an alternative estimate of the performance–TUT rate correlation across the conditions (see Figure 15). Consistent with the bivariate correlations collapsed across conditions (presented above), the meta-analytic correlations indicated no difference in the objective–subjective correlations between the maximized (or standard) and minimized conditions. Specifically, the correlation from a random effects model for the maximized condition was *r* = .18 [.10, .26] and for the minimized condition was *r* = .19 [.08, .29].

Figure 15. Condition-specific correlations and meta-analytic estimates of the Performance × TUT rate correlation for each demand level for Study 1.

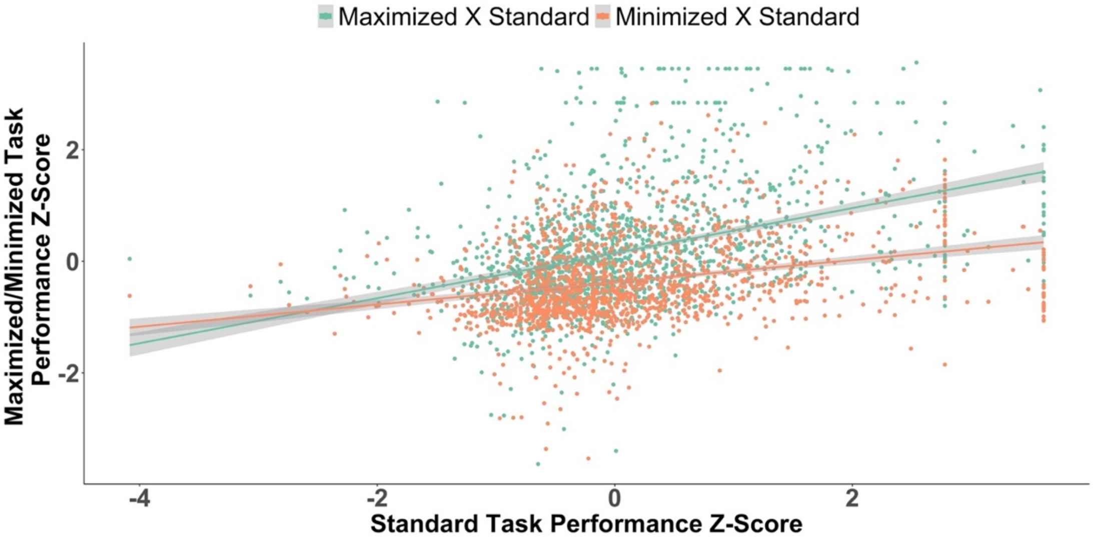


Note. Grey squares represent the correlation estimate for each counterbalancing condition (see Table for descriptions), blue diamonds represent the meta-analytic estimate for each demand level across the conditions and the overall meta-analytic estimate across Demand conditions. Error bars are the 95% CI around the correlation.

Does manipulating sustained attention demands alter the correlations between RT variability measures?

Our next set of analyses focused on whether experimental manipulations of demand affected the associations between RT variability measures across demand levels. In tasks that require sustained attention for optimal performance (i.e., the maximized and standard tasks) the correlation between performance measures should be stronger than when these measures are taken from tasks requiring different levels of sustained attention (i.e., the standard and minimized tasks). As preregistered, we assessed whether the correlation between the maximized and standard tasks was statistically stronger than that between the standard and minimized tasks. Figure 5 displays the scatter plots for these comparisons of interest.

Figure 16. Scatterplot depicting the correlation between z-scored RT variability between the Demand conditions for Study 1



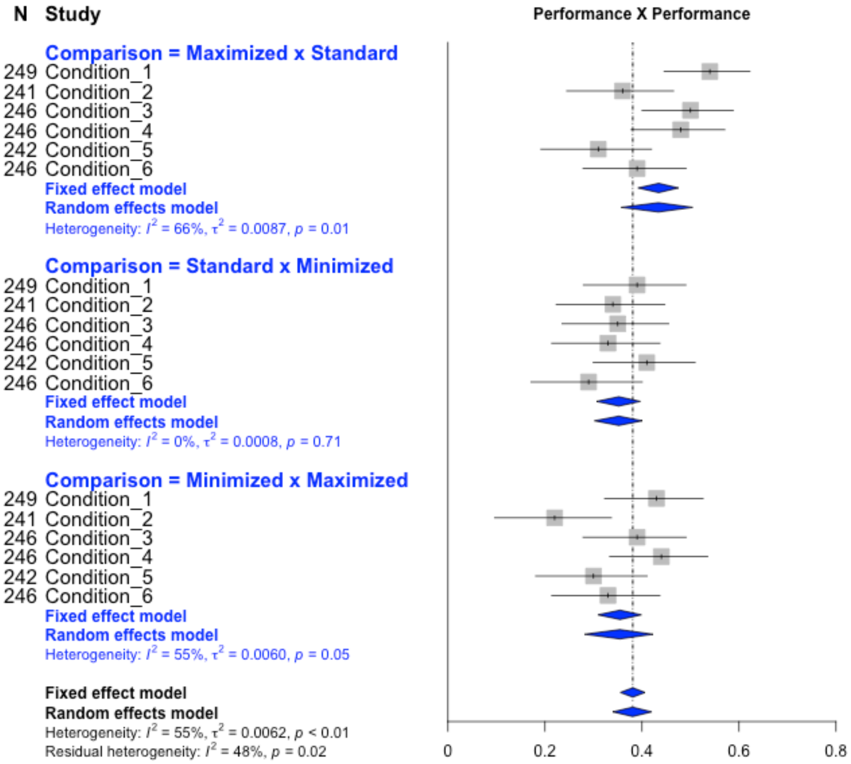
Note. Shaded areas represent the confidence interval around the regression line for each correlation.

As expected, RT variability was strongly correlated between the maximized and standard tasks, *r*(1468) = .40 [.36, .44], *p* < .001. RT variability also correlated moderately and significantly across the standard and minimized tasks, *r*(1468) = .28 [.23, .33], *p* < .001. We next tested whether these two correlations were statistically different from one another. They were. Results of a one-tailed Williams' test of overlapping dependent correlations showed that the correlations between the maximized and standard conditions was significantly stronger than that of the standard and maximized conditions, *t*(1467) = 4.226, *p* < .001. Thus, our experimental manipulation of sustained attention demand did appear to weaken the association between RT variability measures.[14]

As a secondary preregistered approach, we conducted an internal meta-analysis of the performance correlations (see Figure 17). Consistent with the bivariate correlations collapsed across conditions, results from a random effects model indicated that the performance overlap was numerically stronger between the maximized and standard conditions (r = .43 [.36, .50]) compared to the standard and minimized conditions (r = .35 [.30, .40]).

[14] As a secondary analysis we preregistered to examine these performance × performance correlations within each condition and conduct Williams' t tests within each condition. The results of these individual tests are presented in Supplemental Table S21 for completeness, but they are also easily seen in the corresponding meta-analysis.

Figure 17. Condition-specific correlations and meta-analytic estimates of the Performance × Performance correlation for each demand level for Study 1

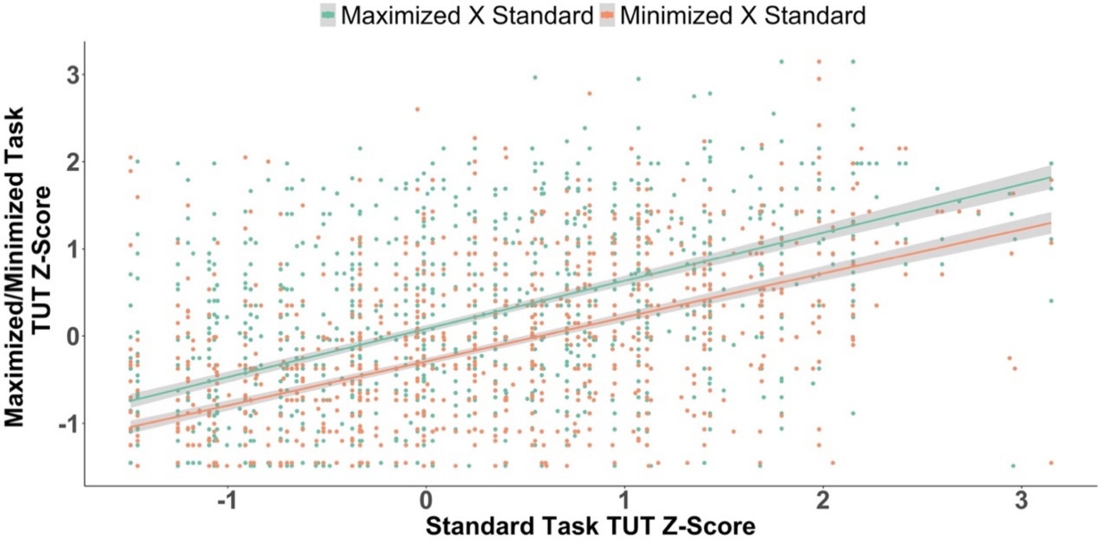


Note. Grey squares represent the correlation estimate for each counterbalancing condition (see Table 17 for descriptions), blue diamonds represent the meta-analytic estimate for each demand level across the conditions and the overall meta-analytic effect across Demand conditions. Error bars are the 95% CI around the correlation.

Does changing sustained attention demands alter the correlations between TUT rates?

We repeated the same analyses reported above, but here testing whether the experimental manipulations of demand altered the relationship between TUT rates across the different demand levels (for scatterplots, see Figure 17). Again, our preregistered hypothesis was that the correlation between TUT rates in the maximized and standard conditions would be stronger than that in the standard and minimized conditions. We employed the same analytic procedure as described above.

Figure 18. Scatterplot depicting the correlation between z-scored TUT rates between the Demand conditions (maximized and minimized each correlated with the standard) for Study 1.



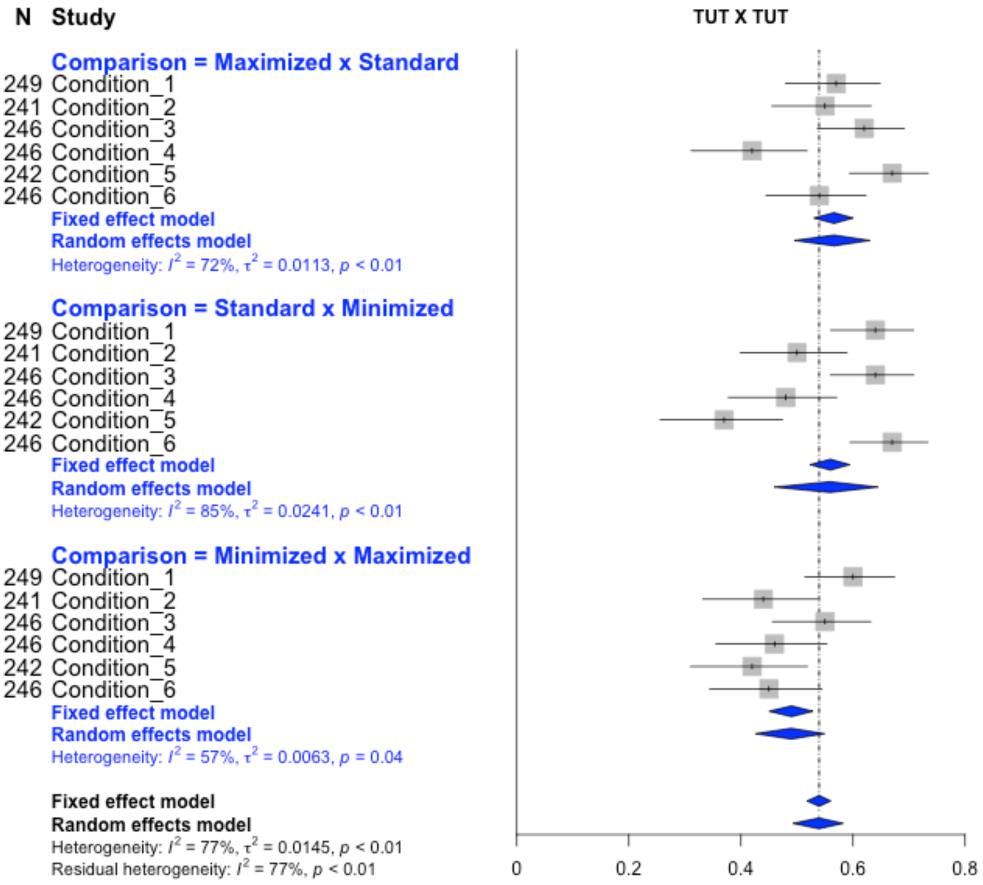
Note. Shaded areas represent the confidence interval around the regression line for each correlation.

TUT rates were strongly correlated between the maximized and standard conditions, *r*(1468) = .55 [.51, .58], *p* < .001, but also between the standard and minimized conditions, *r*(1468) = .55 [.51, .58], *p* < .001. A one-tailed Williams' test indicated that these two correlations were not statistically different, *t*(1467) = 0.00, *p* = .500.[15] As preregistered, we next conducted an equivalence test to see whether these two correlations were statistically equivalent, with equivalence bounds at +0.05 and −0.05 around the maximized × standard correlation (Counsell & Cribbie, 2015; Lakens, 2017). The results of the equivalence test were in line with the above findings indicating the correlations were statistically equivalent, *t*(2937) = 3.378, *p* <

[15] As a secondary analysis we preregistered to examine these TUT × TUT correlations within each individual condition and conduct Williams' *t* tests within each condition. The results of these individual tests are presented in Supplemental Table S22 for completeness, but they are also easily seen in the meta-analysis.

.001. Thus, our experimental manipulation of sustained attention demand had no effect on the correlations between TUT rates. We return to this point in the discussion of Study 1.

Figure 19. Condition-specific correlations and meta-analytic estimates of the TUT rate × TUT rate correlation for each demand level for Study 1.



Note. Grey squares represent the correlation estimate for each counterbalancing condition (see Table 17 for descriptions), blue diamonds represent the meta-analytic estimate for each demand level across the conditions and the overall meta-analytic estimate across Demand conditions. Error bars are the 95% CI around the correlation.

Again, as a second preregistered analysis, we conducted an internal meta-analysis across conditions (see Figure 18). Consistent with the bivariate correlations collapsed across conditions, results from a random effects model suggested that the TUT correlations between maximized and standard conditions, *r* = .57 [.50, .63], was nearly identical to that between standard and minimized conditions, *r* = .56 [.46, .64].

Secondary Analyses of Motivation

Previous research has found modest to strong correlations between self-reported motivation and both objective and subjective measures of attention consistency (e.g., Unsworth & Robison, 2020; Unsworth et al., 2021). In these preregistered secondary analyses, we examined self-reported motivation correlations with RT variability and TUT rates from the maximized and standard tasks (where motivation was reported). Note that we preregistered to assess whether changing the sustained attention demands might change correlations with motivation. However, we ultimately did not include the DSSQ in the design for the minimized tasks because the DSSQ asked about motivation in the previous "task;" we reasoned that, because each minimized task included the CTET as a "break" task between each minimized-task block, this might have confused participants as to what the target "task" was (i.e., the focal sustained attention task or the interpolated CTET). We therefore could only examine correlations between DSSQ scores and the maximized and standard task measures, *deviating from our preregistration*.

Self-reported motivation was weakly correlated with RT variability in the maximized condition, *r*(763) = −.17 [−.23, −.10], and standard condition, *r*(763) = −.17 [−.24, −.10]: Participants who reported higher levels of motivation also showed less RT variability in both tasks. Self-reported motivation was more strongly correlated with TUTs in both the maximized, *r*(763) = −.40 [−.46, −.34], and standard tasks, *r*(763) = −.40 [−.46, −.37]: More motivated participants reported fewer TUTs. *Although not preregistered*, we tested whether the correlations within in each condition were significantly stronger between participants' motivation score and their performance or their TUT rates. To do so, we ran two Williams' tests of dependent correlations. In the maximized condition, the motivation × TUT correlation was significantly stronger than the motivation × RT variability correlation, $t(760) = 5.404$, $p < .001$. This was also true in the standard condition, $t(760) = 5.427$, $p < .001$. Thus, in task contexts that require optimal sustained attention to perform well (i.e., the maximized and standard tasks) self-reported motivation was related to both objective and subjective indicators of attention consistency, but more strongly to the latter (see also Welhaf & Kane, 2022).[16]

Discussion

Study 1 produced several key findings. On one hand, implementing theoretically driven experimental manipulations of task parameters to minimize sustained attention demands significantly reduced mean levels of both RT variability and TUT rates. These significant mean changes, in both types of measures that are thought to assess sustained attention ability (performance and self-report), provide construct-representation evidence for their construct validity. On the other hand, these manipulations of sustained attention demand did not significantly reduce the correlation between objective and subjective measures of attention

[16] Although we preregistered dropping participants who failed the attention check in the DSSQ, we did not anticipate having to drop nearly half of the sample. This may have indicated that subjects failed to understand the reverse-scored item which caused them to answer inappropriately. As a *non-preregistered* exploratory analysis, we reconducted the correlations using DSSQ data from all subjects who were included in the final dataset (N = 1443; note that some participants had missing data on one of the DSSQ measures and so we only looked at participants who had *both* DSSQ scores). The correlations with RT variability and TUT rates were nearly identical as those reported for the reduced sample. Specifically, motivation scores correlated weakly with RT variability in the maximized and standard conditions: Maximized $r(1441) = -.17$ [−.22, −.12], $p < .001$; Standard $r(1441) = -.14$ [−.19, −.09], $p < .001$. Motivation scores also strongly correlated with TUT rates in both conditions: Maximized $r(1441) = -.38$ [−.43, −.34], $p < .001$; Standard $r(1441) = -.38$ [−.42, −.33], $p < .001$. As with the reduced sample, the correlations were stronger with TUT rate than with RT variability in both conditions: Maximized $t(1438) = 6.640$, $p < .001$; Standard $t(1438) = 8.036$, $p < .001$.

consistency in the minimized tasks, which is at odds with our primary predictions and with our previous nomothetic span findings (Welhaf & Kane, 2022).

The lack of reduction in the correlation between objective and subjective indicators of sustained attention, despite the significant (and medium-sized) experimental effects, presents an interesting puzzle. Why might mean rates of our indicators have changed but not the individual-differences overlap between them? One possibility is that our manipulations simply weren't strong enough to impact the correlation between our measures (i.e., the relative standing of individual participants on each measure). For example, our inclusion of periodic "breaks" in the minimized task with the interpolated CTET may have had a counterproductive effect and made the minimized tasks more similar to the maximized/standard tasks. We intended the CTET to break up the monotony and repetitive demands of the focal sustained attention task, with hopes that switching between the tasks would provide an attentional refresh when participants returned to the primary task. However, the CTET is, itself, a sustained attention task, having been used in previous studies of the vigilance decrement over long-duration vigils (Irrmischer et al., 2017; O'Connell et al., 2009). Thus, rather than giving participants a sufficient break in the minimized task, we may have simply switched where their sustained-attention processes were directed. Even though the minimized-task breaks forced the participants to momentarily switch goals (between the focal task and CTET), then, we may have kept them in a context of continuous attentional work. Thus, providing participants with a true break period may be necessary to effectively reduce sustained attention demands in prototypical sustained attention tasks.

Alternatively, our manipulations clearly "worked" to some degree, as both RT variability and TUT rates dropped significantly in the demand-minimized tasks, but perhaps these manipulations affected all participants similarly. That is, we may have simply shifted all participants' RT variability and TUT rates down but roughly preserved their rank order. This might have occurred because some other, non-sustained-attention variable that is correlated with sustained attention ability was unaffected by the demand manipulations. For example, basic processing speed appears to be correlated to general sustained attention ability (Welhaf & Kane, 2022)—that is, to the shared variance between objective and subjective measures—and so, even if the demand-minimized tasks reduced the contribution of sustained-attention ability to the individual-differences variation within each measure, the residual contribution of processing speed may have preserved the correlation between RT variability and TUT rate. We address some of these possibilities in Study 2 and return to this issue in the General Discussion.

It is worth noting briefly that our secondary analyses using PVT lapses as the primary dependent measure for that task (rather than RT for the slowest 20% of trials), reported in Supplemental Table S19, appeared to provide more supportive evidence for construct validity. That is, the simple bivariate correlations between RT variability and TUT rates in the maximized PVT conditions were statistically larger than those in the minimized conditions. However, the meta-analytic approach, which included measures across all tasks, was consistent with the analyses reported in the main text—there was no difference in the correlations between the maximized and minimized conditions. As such, we focus on Slowest 20% as the primary measure in the PVT for Study 2, as we preregistered, and to maintain consistency with other studies investigating sustained attention in the PVT (e.g., Robison & Brewer, 2021; Robison et al., 2021; Unsworth et al., 2021; Welhaf & Kane, 2022).

Study 2

Our Study 1 manipulations of theoretically relevant task parameters yielded experimental reductions in both RT variability and TUTs, but there was no measurable effect on the individual-differences covariation between these measures. Thus, from a construct representation approach, we found only limited support for the construct validity of sustained attention measures. As noted above, however, some of the design features of Study 1 may have prevented a significant enough reduction in sustained attention demands in the minimized tasks.

Inserting the CTET as a break task may not have reduced sustained attention demands in the focal tasks, but instead simply switched the target for participants' sustained attention deployment. We therefore replaced the CTET breaks with true "rest breaks" during the minimized tasks in Study 2. Moreover, to reduce the overall vigilance demand of the entire procedure, we also removed the "standard" demand tasks, leaving participants to complete only a maximized and a minimized task. Finally, we attempted to further reduce the sustained attention demand in the minimized SART and PVT by making them slightly faster (as detailed below).

Methods

We report our sample size justification and data exclusion criteria, as well as all measures and manipulations included in the study (Simmons et al., 2012).

Participants

As preregistered, we again aimed for a sample of 1500 participants. Also as in Study 1, we collected data from both Prolific Academic and the UNCG undergraduate subject pools.

We based this sample size on several calculations. First, we conducted a power analysis in G*power for differences between dependent correlations with no overlap. For a one-tailed test, alpha of .05, and the correlations of interest being .30 (SA-Maximized RTsd × SA-Maximized TUT) versus .20 (SA-Minimized RTsd × SA-Minimized TUTs), and using the correlations from Study 1 for the remaining nonoverlapping correlations, we would have just under 94% power to detect the difference between correlations with N = 1500. Second, to achieve a 95% CI around a *r* = .30 correlation with a lower bound above .25 (for SA-Maximized × SA-Maximized correlations), and around a *r* = .20 correlation with an upper bound below .25 (for SA-Minimized × SA-Minimized correlations), requires at least 1420 participants.

Screening of participants from Prolific and UNCG followed the same protocol as Study 1 (with the same inclusion/exclusion criteria; in addition, participants who participated in Study 1 were not eligible to participate in Study 2). Prolific participants were paid $4.75 for participating in the 30-min study and UNCG students were awarded partial credit toward a research-participation course requirement.

Across both recruitment platforms, we collected data from 1750 participants. Due to an error in a screener on Prolific, we collected data from roughly 175 participants over the age of 40. To retain as much data as possible, we therefore extended the upper limit of our age range from 40 to 45, *deviating from our preregistration*. This resulted in removing data from 143 participants before any analyses were conducted (i.e., we were blinded to the performance and responses from the retained participants between ages 40 and 45). We used our preregistered data screening procedures on the remaining 1607 participants.

Apparatus and Materials

All tasks were programmed as in Study 1, using Gorilla. The experiment for Study 2 can be found on Gorilla's Open Materials site (https://app.gorilla.sc/openmaterials/389050). Again, participants were required to complete the study on a laptop or desktop to ensure accurate recording of RTs.

Tasks

The tasks used for Study 2 were nearly identical to those used in Study 1. Therefore, we only describe the changes made (see Table 17 for task order in each counterbalancing condition).

Maximized ("max") Sustained Attention Tasks. The maximized versions of the SART, PVT, and vMRT were identical to Study 1.

Minimized ("min") Sustained Attention Tasks. The main (and consistent) change across all minimized tasks was that we replaced the CTET "break task" with a true break period, in which participants were allowed to rest briefly and reset their focus. Participants took 15 s breaks at predetermined intervals in each task, described below. After 10 s of each break period, a 5 s countdown timer appeared onscreen and participants were instructed that the task was about to resume.

SART$_{min}$. Beyond rest breaks, the additional change to this task was a reduction of the mask duration to 765 ms (from 935 ms in Study 1). Breaks occurred following probes, as these are natural break points in the task. We aimed to have breaks occur roughly in the middle of each block and at the end of the block, so they occurred after either the third or fourth probe of each block and the sixth probe of each block (which also ended the block). In total, then, the SART presented 7 breaks (vs. 3 CTET breaks in Study 1). To account for potential post-restart RT costs after each break period, we included 2 unanalyzed buffer "go" trials when the task resumed (14 total).

PVT$_{min}$. The main change from Study 1, beyond the rest breaks, was reducing the possible SOA range by 1000 ms compared to Study 1, to 1100–2000 ms (still in 100-ms intervals). Task breaks were inserted following every 15 trials, for 5 breaks total (vs. 2 CTET breaks in Study 1). To match the maximized PVT and account for restart RT costs after breaks, we included 12 unanalyzed buffer trials, two at the beginning of the task, and two following each break.

***vMRT$_{min}$*.** Rest breaks were the only change from Study 1. Break periods followed every two blocks, with 5 breaks total (vs. 2 CTET breaks in Study 1). To match the maximized vMRT and account for restart RT costs after the breaks, we included 12 unanalyzed buffer trials, two at the beginning of the task, and two following each break.

Thought Probes

We used the identical probe wording and response options as Study 1 and TUTs were again scored as the proportion of probes selecting options 2-6. In contrast to Study 1, participants had 10 s select their response to the thought probe, after which the probe disappeared, and the response was counted as missing (Study 2 instructions warned participants of this time limit). As preregistered for Study 2, participants with more than 4 missing probes in any one task were dropped from analyses.

Post-Study Questionnaires

We used the same post-study questionnaire as in Study 1 to screen participants who were distracted by their surroundings, in a sub-optimal subjective state, or a potential bot.

Procedures

The procedures for Study 2 were nearly identical to those of Study 1. Participants provided informed consent and answered the two unrelated questions from Study 1 to screen out potential bots. Next, participants completed the demographics survey and read through the general study and thought-probe instructions (in Study 2, the demographics survey added a forced-choice question for the participants' country of residence, from the list of eligible countries via Prolific screening); following these instructions, they answered the same quiz question as in Study 1 to proceed. Participants were then randomly assigned to one of the six experimental (counterbalancing) conditions as in Study 1. In each condition, participants completed one “maximized” and one “minimized” sustained attention task and, as in Study 1, for three condition orders the maximized task appeared first and for the other three the maximized task appeared second. Finally, participants completed the post-experiment questionnaire.

Data Analysis Exclusions

As preregistered, we screened participants for multiple indicators of inattention or misunderstanding task instructions. First, as previously noted, we dropped data from 143 participants who reported being older than 45. Next, we dropped data from 65 participants who failed to respond to enough task trials (using the same criteria from Study 1: < 70% “go” trial accuracy in the SART or > 15% omission rate in the vMRT), or missed > 4 thought probes within a task. We next dropped data from 18 participants because they either: (a) responded “extremely” to two of the three post-experiment questions regarding their subjective state or immediate environment (i.e., noisy, distracted, and/or sleepy); or (b) responded to any of the media-multitasking questions with “Most of the time” or “All of the time.” Finally, we dropped data from 21 participants who provided an inappropriate written response to the open-ended question about the “most challenging” task in the study. We retained data from 1502 participants.

Results

Below we report the results of our preregistered analyses and note where we deviated from the preregistered plan. All data aggregation and analyses used the same packages listed for Study 1. Data and Rmarkdown files for all analyses are available on the Open Science Framework (https://osf.io/kmqbs/).

Final Sample Demographics

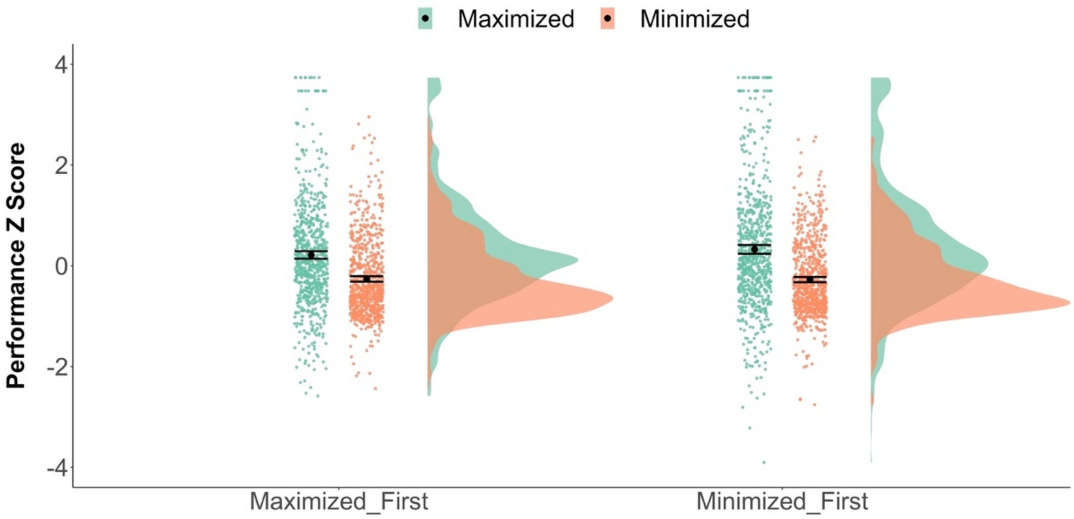
Of the retained 1502 participants, 45.1% self-identified as female, 52.1% self-identified as male, and 2.7% self-identified non-binary or gender-nonconforming. The mean age of the full sample was 27.4 years (SD = 7.4). The self-identified racial breakdown of our final sample was 68% White/European descent, 11% Black/African descent, 7% Multiracial, 4% South Asian descent, 4% East Asian descent, 4% Hispanic or Latino/Latina, 1% Middle-Eastern, Arab, or North African descent, and <1% Native Hawaiian or Pacific Islander or Native American or Alaskan Native descent. Finally, Prolific participants (*n* = 1207) showed a range of educational attainment: 31% had a high school diploma or A-level certificate, 11% earned a technical or community college degree, 41% earned an undergraduate degree, 15% earned a graduate degree, and 1% earned a doctorate; the 295 UNCG participants were all undergraduates. Prolific participants indicated the following countries of residence: 49.1% UK, 41.3% US, 7.1% Canada, 1.4% Ireland, 0.9% Australia, 0.2% New Zealand.

Experimental Effects of Sustained Attention Demand and Task Order

As preregistered, we first examined whether our experimental manipulations had any effect on measures of RT variability or TUT rates. Again, to make performance measures comparable across the tasks, we z-scored performance collapsed across all levels (maximized/minimized; first/last) within each task. Supplemental Table S23 presents the raw descriptive statistics for RT variability and TUT rates in each task, for each condition.

RT Variability. The results of a 2 (Demand: Maximized vs. Minimized) × 2 (Order: Maximized First vs. Minimized First) ANOVA (with Greenhouse-Geisser correction for within-subject comparisons) on RT variability indicated a significant effect of Demand, $F(1, 1500) = 359.60$, $p < .001$, $\eta_p^2 = .193$, but no significant effect of Order, $F(1, 1500) = 1.42$, $p = .234$, $\eta_p^2 < .001$, and a significant (but small) Demand × Order interaction, $F(1, 1500) = 4.69$, $p = .031$, $\eta_p^2 = .003$. Figure 19 visually presents these results.

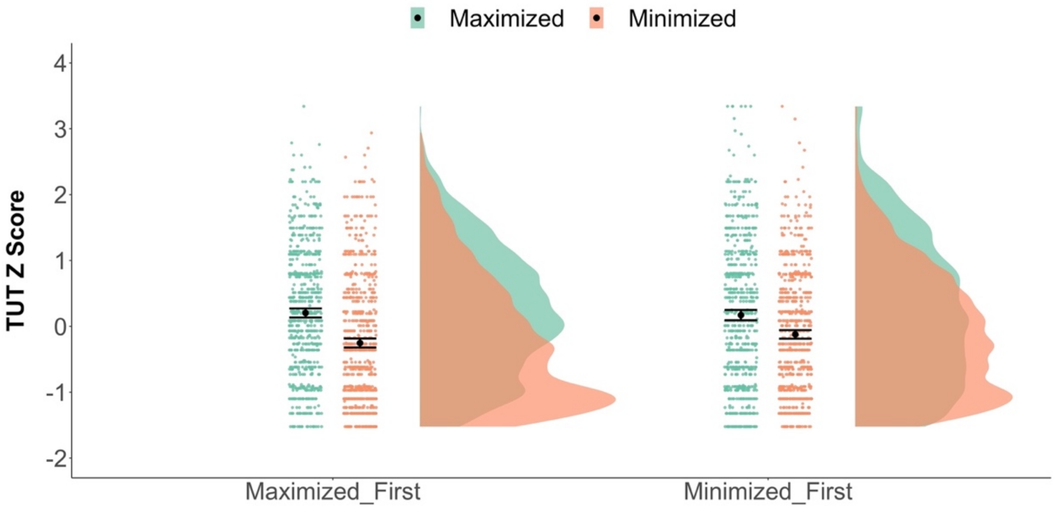
Figure 20. Raincloud plots depicting Demand-condition differences in z-scored RT variability by task Order, collapsed across all sustained attention tasks for Study 2



Note. Dots represent individual subject means in each condition. The closed black dots represent group-level mean estimates for each demand level. Error bars are 95% confidence intervals.

To follow up the significant interaction, we conducted paired t-tests separately for each Order condition. When the maximized task appeared first, RT variability was significantly lower for the minimized versus the maximized task, $t(749) = 12.354$, $p < .001$, $d = 0.45$ [0.38, 0.53]. The same was true when the minimized task appeared first, $t(751) = 14.41$, $p < .001$, $d = 0.53$ [0.45, 0.60]. Although significant in both orders, the effect of Demand was larger when the minimized task appeared first.

Figure 21. Raincloud plots depicting Demand-condition differences in z-scored TUT rates by task Order, collapsed across all sustained attention tasks for Study 2



Note. Dots represent individual subject means in each condition. The closed black dots represent group-level mean estimates for each demand level. Error bars are 95% confidence intervals.

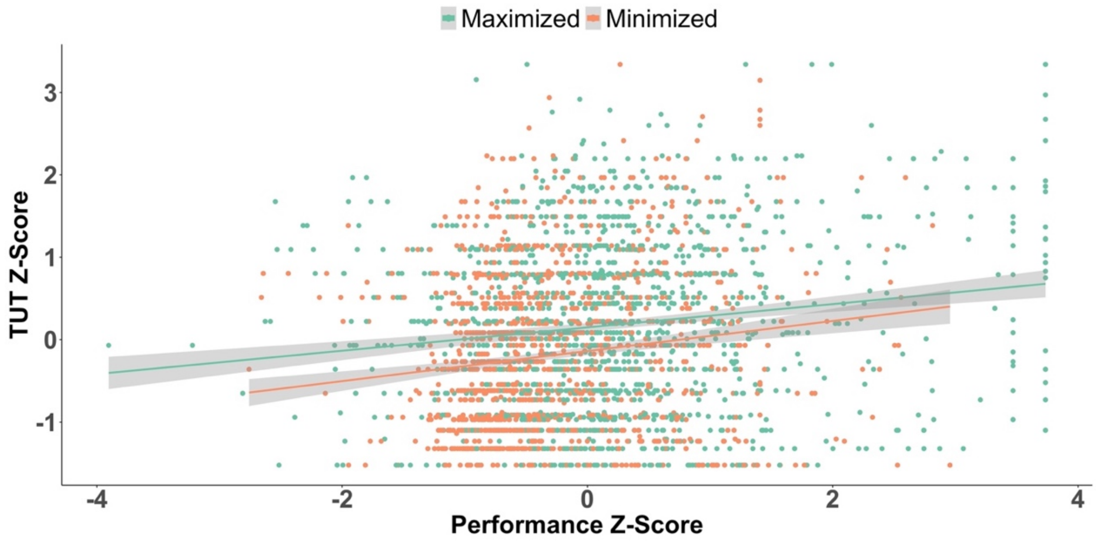
TUT Rates. We conducted the same 2 (Demand: Maximized vs. Minimized) × 2 (Order: Maximized First vs. Minimized First) ANOVA on TUT rates (see Figure 20). The results indicated a significant effect of Demand, $F(1, 1500) = 256.00$, $p < .001$, $\eta_p^2 = .146$, no significant effect of Order, $F(1, 1500) = 1.20$, $p = .273$, $\eta_p^2 < .001$, but a significant (but small) Demand × Order interaction, $F(1, 1500) = 12.16$, $p < .001$, $\eta_p^2 = .008$.

We then conducted paired *t*-tests separately for each order condition. When the maximized task came first, TUT rates were significantly lower for the minimized than maximized task, $t(749) = 14.057$, $p < .001$, $d = 0.51$ [0.44, 0.59]. When the minimized task appeared first, TUT rates were again significantly lower for the minimized task, $t(751) = 8.681$, $p < .001$, $d = 0.32$ [0.24, 0.39]. Although significant in both orders, Demand effects were larger when the maximized task appeared first.

Does Manipulating Sustained Attention Demands Alter the Correlation Between RT Variability and TUT Rates?

If the individual-differences covariation between RT variability and TUT rates provides a construct valid assessment of sustained attention ability, then reducing the sustained attention demands should reduce the amount of overlap in these indicators compared to a condition where sustained attention is necessary to perform optimally. As in Study 1, we again tested whether this overlap (i.e., the correlation between RT variability and TUTs) was significantly weaker in the minimized than in the maximized condition.

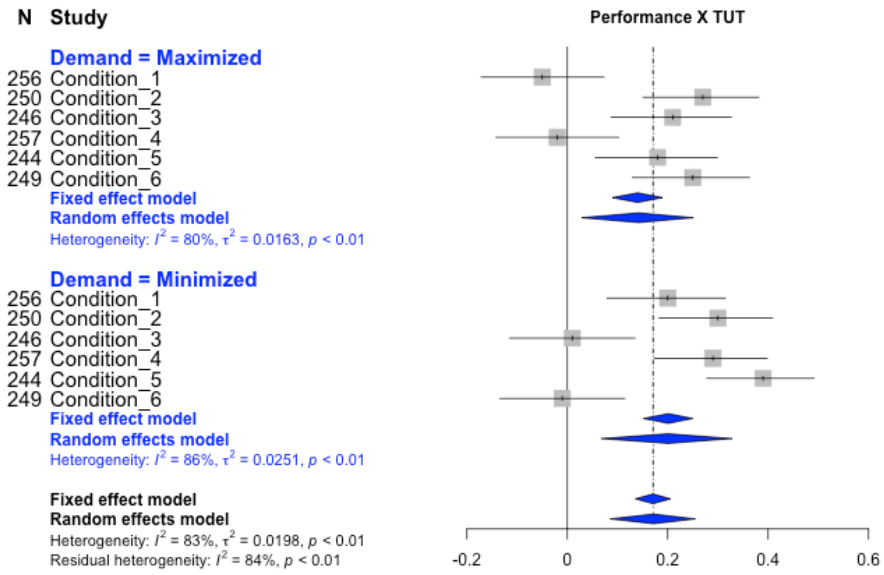
Figure 22. Scatterplot depicting the correlation between z-scored RT variability performance and z-scored TUT rates in the Maximized and Minimized Tasks for Study 2.



Note. Shaded areas represent the confidence interval around the regression line for each group.

As preregistered, we first correlated RT variability scores with TUT rates for the maximized tasks and for the minimized tasks (see Figure 21). RT variability and TUT rate were significantly, although weakly, correlated in both demand levels: Maximized *r*(1500) = .16 [95% CI .11, .21], *p* < .001, Minimized *r*(1500) = .15 [.10, .20], *p* < .001. Participants who were more variable in their responding during the task also tended to report more TUTs. Next, as preregistered, we compared these correlations using Steiger's test for non-overlapping correlations (one-tailed). The correlation in the maximized condition was not significantly larger than that in the minimized condition, z = 0.316, *p* = .376. Thus, our manipulations aimed at reducing the need for sustained attention in the minimized tasks did not reduce the shared variance between the performance-based and self-report measures of sustained attention.

Figure 23. Condition-specific correlations and meta-analytic estimates of the Performance × TUT correlation for each demand level for Study 2.



Note. Grey squares represent the correlation estimate for each condition, blue diamonds represent the meta-analytic estimate for each demand level across the conditions and the overall meta-analytic estimate across the Demand conditions. Error bars are the 95% CI around the correlation.

As a second preregistered approach to investigating this overlap, we conducted an internal meta-analysis to get an alternative estimate of the correlations across the conditions (see Figure 22). The results of a random effects meta-analysis indicated that the meta-analytic correlation in the maximized conditions was again weak, *r* = .14 [.03, .25], with a slightly *stronger* correlation in the minimized conditions, *r* = .20 [.07, .33] (although their confidence intervals overlapped). These results are largely consistent with the bivariate correlations reported above, as well as with the results of Study 1. Our manipulations in the minimized tasks did not reduce the sustained attention demands enough to reduce the correlation between RT variability and TUTs.

Combined-Study Exploratory Analyses of Experimental Effects

Although there were clear experimental effects of our demand manipulations in both studies, it is unclear whether the additional changes we made in Study 2 to the minimized tasks (i.e., replacing the CTET with break periods, increasing the task pacing in the SART and PVT) further reduced sustained attention demands, as intended. We investigated this via a series of exploratory (*non-preregistered*) analyses combining data from both Study 1 and Study 2. Specifically, we pooled task data from each condition where the maximized or minimized task was presented *first* in each study (to avoid order effects).

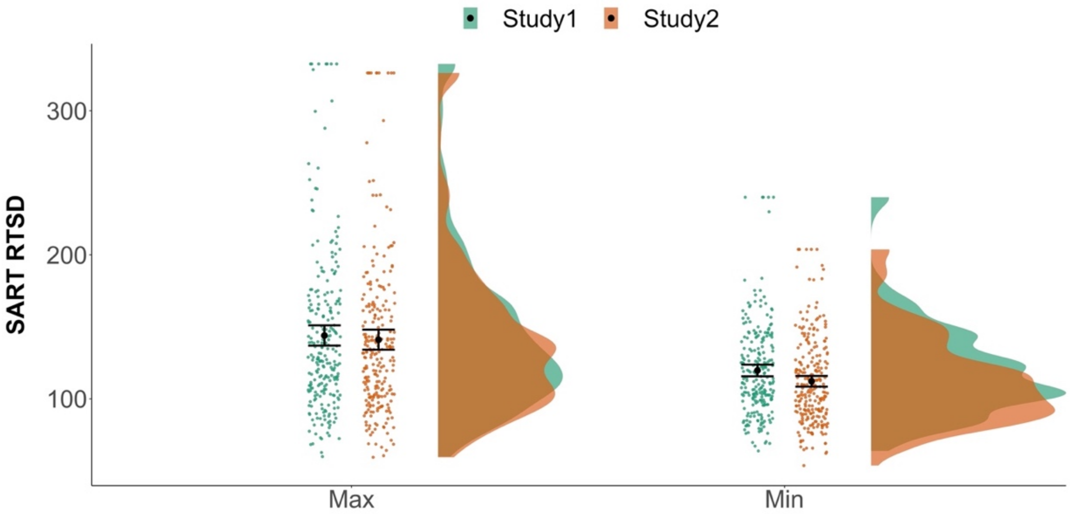
Performance measures. We present the results for raw performance variability measures for each task below.

SART. A 2 (Study: Study 1 vs. Study 2) × 2 (Demand: Maximized vs. Minimized) ANOVA on SART RTsd did not indicate a significant main effect of Study, $F(1, 982) = 3.27$, $p = .071$, $\eta_p^2 = .003$, or an interaction, $F(1, 982) = 0.64$, $p = .424$, $\eta_p^2 < .001$. However, there was a significant main effect of Demand, $F(1, 982) = 85.14$, $p < .001$, $\eta_p^2 = .080$.

As displayed in Figure 23, across the studies, SART RTsd was higher in the Maximized ($M = 143$) compared to the minimized task ($M = 116$). As expected, SART RTsd was nearly identical in the Maximized tasks across studies (Study 1 $M = 144$; Study 2 $M = 141$, $t(982) = 0.715$, $p = .475$). SART RTsd in the Minimized tasks was numerically, but not significantly, lower in Study 2 ($M = 112$) than in Study 1 ($M = 120$), $t(982) = 1.841$, $p = .066$. Thus, in the

SART, our changes to the Study 2 did not significantly reduce the sustained attention demands beyond those initially made in Study 1.

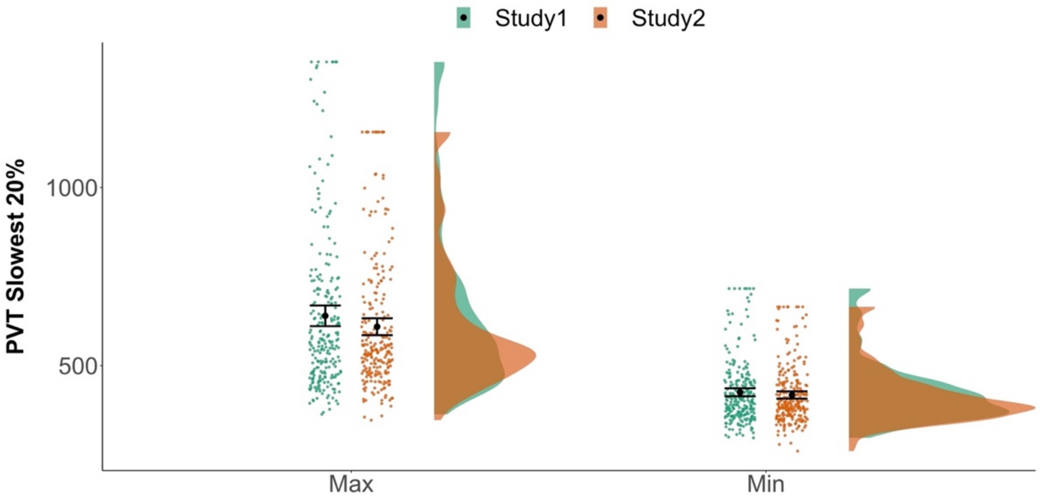
Figure 24. Raincloud plots depicting Study differences in SART intrasubject RTsd means × Demand



Note. Dots represent individual subject means in each Study and condition. The closed black dots represent group-level mean estimates for each Study. Error bars are 95% confidence intervals. Max = maximized sustained attention demand; Min = minimized sustained attention demand.

PVT. We next conducted the same 2 (Study) × 2 (Demand) ANOVA in the PVT Slowest 20% outcome. The ANOVA indicated a nonsignificant effect of Study, $F(1, 987) = 3.54$, $p = .060$, $\eta_p^2 = .004$ a significant effect of Demand, $F(1, 987) = 402.43$, $p < .001$, $\eta_p^2 = .290$, and no significant interaction, $F(1, 987) = 1.29$, $p = .255$, $\eta_p^2 = .001$.

Figure 25. Raincloud plots depicting Study differences in PVT Slowest 20% RT means × Demand

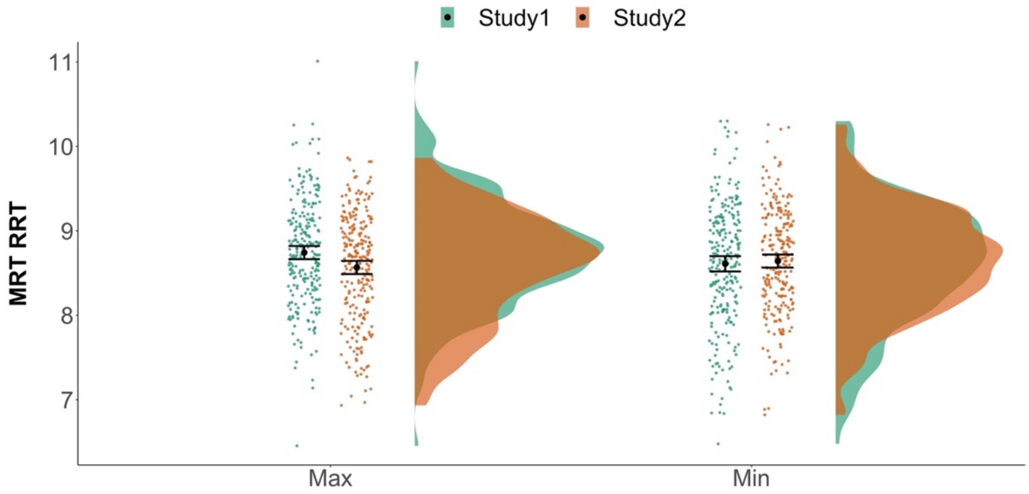


Note. Dots represent individual subject means in each Study and condition. The closed black dots represent group-level mean estimates for each Study. Error bars are 95% confidence intervals. Max = maximized sustained attention demand; Min = minimized sustained attention demand.

As displayed in Figure 24, participants, on average, had longer "long" RTs in the PVT in the maximized condition across studies (*M* = 625) compared to the minimized condition (*M* = 421), *t*(987) = 20.061, *p* < .001. Surprisingly, performance in the two maximized tasks differed across the studies (Study 1 *M* = 640 vs Study 2 *M* = 609; *t*(987) = 2.115, *p* = .035), but the minimized tasks did not differ (Study 1 *M* = 425 vs. Study 2 *M* = 418; *t*(987) = 0.531, *p* = .595). Thus, as in the SART, our Study 2 PVT manipulations did not successfully reduce the sustained attention beyond those in Study 1.

MRT. Finally, we ran the same 2 (Study) × 2 (Demand) ANOVA on MRT RRT scores. The ANOVA did not indicate an effect of Experiment, *F*(1, 991) = 3.05, *p* = .081, η_p^2 = .003 or Demand, *F*(1, 991) = 0.48, *p* = .487, η_p^2 = .487. However, there was a significant Study × Demand interaction, *F*(1, 991) = 6.43, *p* = .011, η_p^2 = .006.

Figure 26. Raincloud plots depicting Study differences in MRT RRT means × Demand



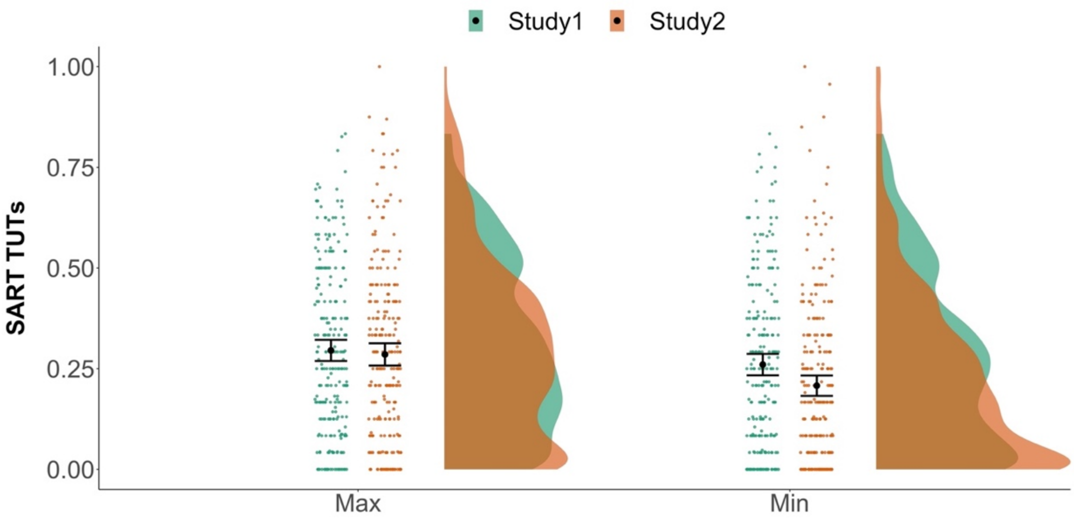
Note. The closed black dots represent group-level mean estimates for each Study. Error bars are 95% confidence intervals. RRT = rhythmic response time; Max = maximized sustained attention demand; Min = minimized sustained attention demand.

As shown in Figure 25, RRT in the *maximized* tasks decreased significantly from Study 1 (*M* = 8.74) to Study 2 (*M* = 8.57), *t*(991) = 3.043, *p* = .002. However, contrary to expectation and intention, there was not a significant decrease in the minimized tasks between Study 1 (*M* = 8.61) and Study 2 (*M* = 8.64), *t*(991) = -0.555, *p* = .5789. Thus, again, we appear to have not reduced the sustained attention demands in Study 2 beyond those of Study 1 in the minimized MRT.

TUT rates. In parallel to the performance analyses, we next investigated whether our additional changes to the minimized tasks in Study 2 further reduced TUT rates beyond the manipulations implemented in the minimized tasks in Study 1. Here we again pooled raw TUT rate data from the maximized and minimized tasks when they appeared first in the condition order. We present results for each task separately below.

SART TUTs. A 2 (Study) × 2 (Demand) ANOVA indicated a significant effect of Study, $F(1, 982) = 5.37$, $p = .021$, $\eta_p^2 = .005$, and a significant effect of Demand, $F(1, 982) = 17.64$, $p < .001$, $\eta_p^2 = .018$, with no significant interaction, $F(1, 982) = 2.49$, $p = .115$, $\eta_p^2 = .003$.

Figure 27. Raincloud plots depicting Study differences in SART TUT rates × Demand

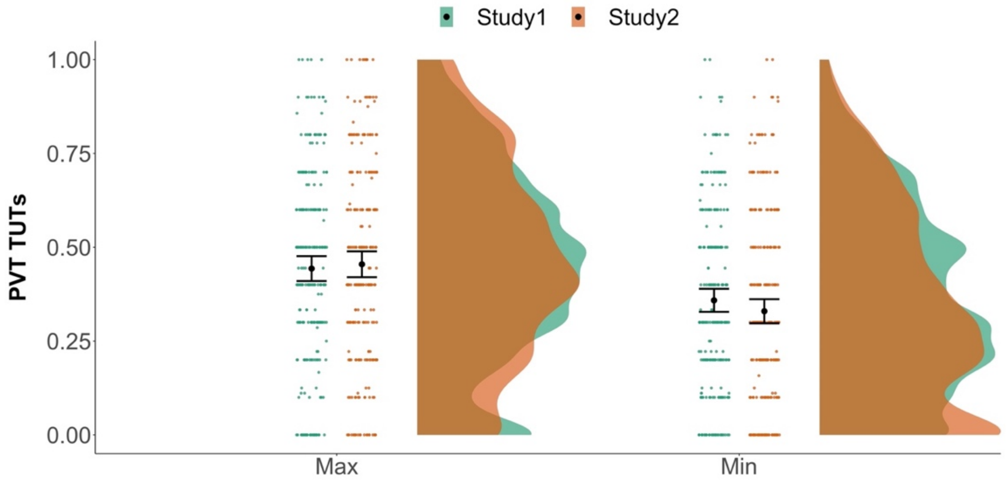


Note. Dots represent individual subject means in each Study and condition. The closed black dots represent group-level mean estimates for each Study. Error bars are 95% confidence intervals. Max = maximized sustained attention demand; Min = minimized sustained attention demand.

As shown in Figure 26, TUT rates in the maximized SARTs were similar across the studies (Study 1 *M* = .30, Study 2 *M* = .30; *t*(982) = 0.523, *p* = .601). However, in the minimized tasks, TUTs significantly decreased from Study 1 (*M* = .26) to Study 2 (*M* = .21), *t*(982) = 2.750, *p* = .006. Thus, in the self-report dependent measure (but not the objective performance measure) the changes we made to the minimized task in Study 2 did appear to further reduce the sustained attention demands from the minimized task of Study 1.

PVT TUTs. A 2 (Study) × 2 (Demand) ANOVA on PVT TUT rate indicated no significant effect of Study, *F*(1, 987) = 0.28, *p* = .597, $\eta_p^2 < .001$, but a significant effect of Demand, *F*(1, 987) = 40.17, *p* < .001, $\eta_p^2 = .039$. There was no significant interaction, *F*(1, 987) = 1.50, *p* = .222, $\eta_p^2 = .002$.

Figure 28. Raincloud plots depicting Study differences in PVT TUT rates × Demand

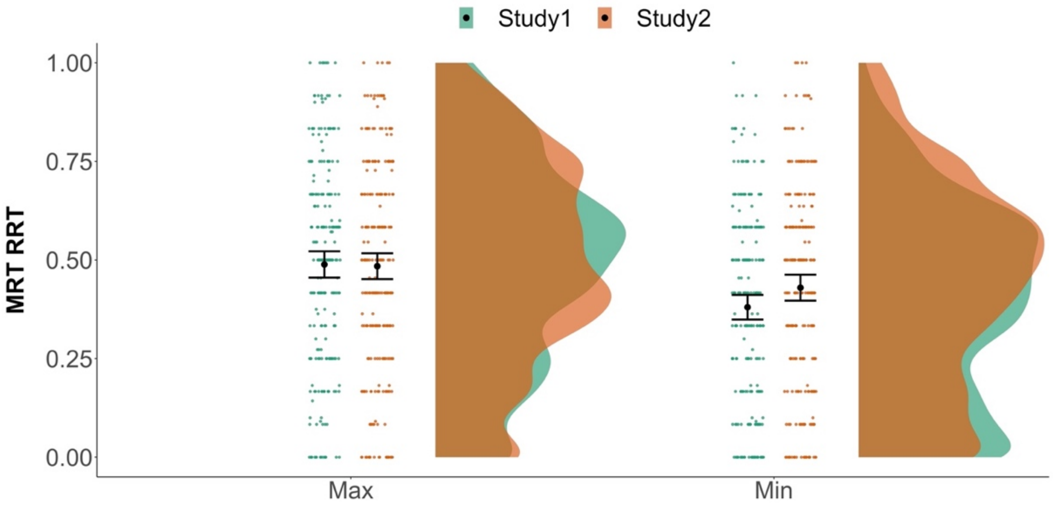


Note. Dots represent individual subject means in each Study and condition. The closed black dots represent group-level mean estimates for each Study. Error bars are 95% confidence intervals. Max = maximized sustained attention demand; Min = minimized sustained attention demand.

As shown in Figure 27, TUT rates in the PVT were higher, but similar, in the maximized condition (Study 1 *M* = .44, Study 2 *M* = .46, *t*(987) = -0.486, *p* = .627), compared to those in the minimized tasks, which also did not differ from each other (Study 1 *M* = .36, Study 2 *M* = .33, *t*(987) = 1.251, *p* = .211). Thus, our additional changes to the minimized PVT in Study 2 did not yield further reduction in TUT rates beyond those implemented in Study 1.

MRT TUTs. A 2 (Study) × 2 (Demand) ANOVA on MRT TUT rates indicated no effect of Experiment, *F*(1, 991) = 1.87, *p* = .172, η_p^2 = .002, but a significant effect of Demand, *F*(1, 991) = 24.30, *p* < .001, η_p^2 = .024, and no significant interaction, *F*(1, 991) = 2.66, *p* = .103, η_p^2 = .003.

Figure 29. Raincloud plots depicting Study differences in MRT TUT rates × Demand



Note. Dots represent individual subject means in each Study and condition. The closed black dots represent group-level mean estimates for each Study. Error bars are 95% confidence intervals. Max = maximized sustained attention demand; Min = minimized sustained attention demand.

As seen in Figure 28, TUTs in the maximized MRT were not statistically different from each other (Study 1 *M* = .49, Study 2 *M* = .48, *t*(991) = 0.188, *p* = .851). TUT rates in the minimized version were lower (Study 1 *M* = .38, Study 2 *M* = .43), but they were higher in Study 2 than in Study 1, *t*(991) = -2.108, *p* = .035. Thus, our additional manipulations to the minimized MRT in Study 2 appear to have had the inverse effect on TUT rates, compared to the minimized MRT in Study 1.

In sum, none of the individual sustained attention tasks in Study 2 produced lower performance variability than did those in Study 1, and only one of three tasks produced lower TUT rates in Study 2 than in Study 1. Our additional manipulations of sustained attention demand in Study 2 were thus unsuccessful.

Discussion

Study 2 attempted to conceptually replicate Study 1 after further reducing the sustained attention demands of the minimized tasks, to assess the evidence for the construct validity of performance and self-report sustained attention measures. The results largely replicated Study 1. Our experimental manipulations again had their intended effects: Participants showed less RT variability and reported fewer TUTs in the minimized than in the maximized conditions. However, our "enhanced" manipulations of demand again failed to reduce the sustained attention contributions to the minimized tasks enough to weaken the correlation between RT variability and TUT rates (they also did not produce larger experimental effects than did our original Study 1 manipulations). The results of Study 2 again provide only limited construct-representation evidence that the individual-differences overlap between objective and subjective measures provides a construct valid way to measure general sustained attention ability.

One possible reason that the additional manipulations in Study 2 did not reduce the covariation between objective and subjective attention consistency measures is that, as noted above, they did not actually affect the outcome measures any further than did those in Study 1. Comparing outcomes for the individual maximized and minimized tasks (presented first in each participant's task order) indicated that our additional manipulations did not generally change RT variability or TUT rates from Study 1 to Study 2. Thus, while both Study 1 and Study 2 demonstrated that it is possible to reduce mean levels of sustained attention indicators somewhat, it may be extremely difficult to reduce the sustained attention demands substantially, and substantially enough to diminish the contributions of sustained attention processes to *between-person* variation. We return to this point in the General Discussion.

General Discussion

The current studies tested whether theoretically driven, experimental manipulations of task parameters linked to sustained attention demands could reduce the correlation between objective and subjective indicators of attention consistency, thereby providing support for their covariation as a construct valid measure (Welhaf & Kane, 2022). The results from these two large-N studies provide only limited construct validation evidence. On one hand, both measures were similarly impacted by our manipulations at the mean level; that is, in exploratory analyses, RT variability and TUT rates both decreased significantly in versions of the tasks that were designed to minimize their sustained attention demands. On the other hand, these experimental manipulations failed to reduce the correlation between these measures, which we have argued is a more construct-valid way to assess sustained attention ability than is either performance-based or self-report-based measurement alone (Welhaf & Kane, 2022).

Prior research has generally found that experimental manipulations aimed at reducing sustained attention demands lead to lower RT variability or TUT rates (e.g., Giambra, 1995; Langner & Eickhoff, 2013; Seli et al., 2019; Unsworth & Robison, 2020). Nonetheless, both performance- and self-report-based indicators of attention consistency are independently affected by psychological processes and cognitive abilities beyond sustained attention (e.g., processing speed, meta-awareness, self-report biases). The results of previous studies that have examined these measurement types separately may therefore have landed on incomplete conclusions about how these manipulations relate to general sustained attention processes or ability. The present study addressed this by examining whether the individual-differences overlap between RT variability and TUT rates changed as a function of the sustained attention demands of the task. It did not: In neither study did our manipulations significantly reduce the correlation between RT variability and TUT rates.

Were We Wrong About Sustained Attention Measurement?

Why, despite conventionally medium-sized experimental effects, were we unable to reduce the correlation between RT variability and TUT rates? It is possible that we were simply wrong about the covariation between RT variability and TUT rates as a construct valid way to measure sustained attention ability: Perhaps either measurement type alone is as valid a sustained attention measure as is their covariation. We don't yet favor this possibility.

First, there are numerous construct-specific nuisance variables that contribute to RT variability and TUT rates that are not related to sustained attention ability. For example, changes in response strategies (i.e., speed-accuracy trade-offs) influence RT variability, but they should not influence TUT reports. Likewise, self-report biases that influence participants when answering questions about their conscious experiences will contribute to TUT rates to some degree, but they should not influence RT variability. Using the individual-differences overlap in these two indicator types should therefore provide a measure of sustained attention that is less contaminated by these sources of measurement error. Second, the construct validity evidence from the nomothetic span approach (Welhaf & Kane, 2022) has indicated that the shared variance in these measures can be modelled as a higher-order factor and this general factor differentially correlated with nomological-network constructs when compared to either RT-variability-specific or TUT-rate-specific latent variables. Third, the experimental effects on mean RT variability and TUT rate found in the present study's exploratory analyses also suggest that these measures are similarly impacted by theoretically derived manipulations that should impact sustained attention measurement. Taken together, these findings are suggestive that both measure types, but especially their covariation, may be valid indicators of sustained attention ability.

Were Our Manipulations of Sustained Attention Demand Ineffective?

If we take a "Lakatosian-defense" posture (Meehl, 1990) and *for now* act as though this covariation approach has merit, what do we make of the current findings? Which of our auxiliary hypotheses or assumptions that drove the design of our study might need reconsideration or revision? One possibility is that sustained attention, perhaps even more than other executive or attention-control processes (Engle, 2002; Engle & Kane, 2004), is fundamental to nearly any task that requires more than a few seconds of active engagement. If so, it will be extremely difficult to reduce the sustained attention demands enough in any task to substantially reduce the between-person variation in attention consistency measures.

Potential Problems

Perhaps, then, we can only create significant reductions in the RT variability–TUT rate association by using tasks that feature our demand-reducing manipulations but are also exceedingly brief. The potential cost of using brief tasks to minimize sustained attention demands, however, is that it may also reduce the reliability of the tasks for use in individual-differences research: Shortening tasks reduces the overall number of trials available for accurate assessment of RT variability and TUT rates.

A second possibility is that measuring and manipulating sustained attention in an online setting cannot produce the necessary effect sizes to reduce correlations between performance variation and TUT rates. Although the COVID-19 pandemic forced us to conduct this study online, we attempted to minimize threats to internal validity by focusing our analyses on comparing correlations across *within-subjects* conditions, and by asking participants about their immediate environments and dropping data from those who acknowledged significant distraction during the study. With that said, our Prolific and UNCG participants may have completed this study in non-ideal settings that were moderately distracting (or, at least, more distracting than a typical laboratory).

We dropped data, as preregistered, from participants indicating an "extreme" level of noise or distraction, but 59% of retained participants across both studies indicated that their surrounding environment was at least slightly noisy or distracting. As well, 12% of retained participants self-reported occasionally multi-tasking on their phone or email during the study, but again we screened out only those participants who self-reported media multi-tasking "most of the time," or "all of the time." Of course, by having participants self-report on these environmental factors, we also had to trust that their ratings were accurate and truthful; we may have *underestimated* environmental distraction and multitasking in our sample if participants were concerned that admitting to distraction might put their compensation in jeopardy.

Potential Solutions

It is possible that the correlation between RT variability and TUTs *can be* substantially reduced, but our manipulations simply weren't strong enough to do so. In Study 2, for example, we implemented periodic rest breaks, rather than having participants switch to an alternative task, as in Study 1. Perhaps these task switches or rest breaks were too short to have their intended effect. Previous research has found that task rest breaks (e.g., Helton & Russell, 2015, 2017), or switching between primary and secondary tasks (e.g., Ariga & Lleras, 2011; Ralph et al., 2017), can reduce the vigilance decrement that occurs over long tasks, at least briefly. These studies have typically provided participants with a single break or task switch lasting from just under 2 min (Helton & Russell, 2015, 2017) up to roughly 8 min (Ralph et al., 2017). In Study 1, our task switches lasted about 1 min; in Study 2, we fixed the more frequent rest breaks to 15 s. Although we used more frequent switches and breaks than in prior work, the total switch and break time may not have been enough across the minimized tasks to effectively reduce their sustained attention demands.

The literature suggests two additional manipulations that we did not use in the current study that might further reduce the sustained attention demands of a task: the frequency of thought probes and heightened motivational states. We have previously argued that using fewer thought probes can reduce the number of task interruptions and the possibility of reactivity to probes (Welhaf et al., 2021). More frequent probing may therefore reduce sustained attention demands in simple tasks. First, participants will only be engaged in the primary task for short periods before they are given a probe, and so probes may serve as additional task breaks. Second, probes can serve as reminders for participants to keep their thoughts focused on the task and so may help scaffold participants' sustained attention processes (see, e.g., Robison et al., 2019; Schubert et al., 2020; Seli, Carriere, et al., 2013). Increased probe frequency might therefore reduce the sustained attention demands enough to show measurable changes in mean levels of RT variability or TUT rates, and importantly in their covariation.

Participants' self-reported motivation is often related to both objective and subjective indicators of sustained attention, with higher motivation associated with lower RT variability and TUT rates (e.g., Unsworth et al., 2021); self-reported motivation also correlates with the shared variance between objective and subjective indicators of sustained attention (Welhaf & Kane, 2022). Additionally, as noted earlier, experimental manipulations of motivation appear to reduce RT variability and TUT rates compared to control conditions (e.g., Esterman et al., 2014; Robison et al., 2021; Seli et al., 2019; Unsworth et al., 2022). These findings are suggestive, then, that manipulations of motivation, either through monetary incentives or achievable performance goals, may affect the covariation between objective and subjective measures of sustained attention.

Finally, while we focused on implementing manipulations to minimize the sustained attention demands of some of the tasks, we did not consider the possibility of boosting the demands in our “maximized” tasks. While these tasks did challenge sustained attention, it may be necessary to both minimize and maximize the demands of tasks to see any substantial difference in the correlations. For example, we could have made our maximized tasks more difficult by requiring less frequent responding (in the SART and MRT) or increasing the length, or variability, of the SOAs in the PVT, which should make these tasks more demanding on sustained attention.

We therefore encourage future experimental work on sustained attention measurement to consider the covariation of performance variability and TUT rates as a construct valid assessment. Moreover, we recommend that researchers test such claims in controlled laboratory settings and by manipulating a number of variables simultaneously to reduce, and also increase, the sustained attention demands of prototypical tasks.

A Potential Constraint on Generalizability

It is important to note that our perspective on sustained attention differs from some traditional operationalizations. Historically, the measurement of sustained attention has focused on the need to maintain focus over *many* trials (and over *many* minutes) of a single task. Failures of sustained attention from this approach are reflected in worsening performance over time, the so-called “vigilance decrement” (e.g., Lim & Dinges, 2008; Mackworth, 1950; Parasuraman, 1986). Our view and measurement approach to “attentional consistency” (see Esterman &

Rothlein, 2019; Unsworth & Miller, 2021) has instead focused on the moment-to-moment ability to maintain attention focus and consistency within tasks of short-to-moderate duration. We leave it to future work, then, to consider the extent to which our claims and findings apply to the vigilance decrement as an indicator of sustained attention.

Conclusions

We have previously argued that the individual-differences overlap in objective and subjective measures is a more construct-valid way to measure sustained attention ability than is using either indicator in isolation (Welhaf & Kane, 2022). The results of the current study suggest that each of these indicator types is separately, and similarly, affected by theoretically derived manipulations to reduce sustained attention demands. Contrary to predictions, however, the covariation between these measures was not. Thus, we found only limited construct-representation evidence for the construct validity of measuring sustained attention as the covariation between performance-variability and self-report measures. We speculate that sustained attention processes and abilities may be so fundamental to any given task that it may be exceedingly difficult to find experimental manipulations that substantially reduce their correlation, especially in an online setting.

CHAPTER V: INTEGRATIVE DISCUSSION

The goal of this integrated dissertation was to present a program of research aimed at evaluating the construct validity of sustained attention measures, and critically, their individual differences covariation as a construct-valid approach to measuring sustained attention ability. The literature has traditionally used two approaches to assessing moment-to-moment fluctuations in sustained attention, or "attention consistency" (Unsworth & Miller, 2021): objective performance measures and subjective self-reports of task-unrelated thoughts (TUTs). Studies typically investigate how these two forms of sustained attention correlate with each other (e.g., RT variability–TUT-rate correlations) or with other theoretically relevant variables (e.g., WMC–TUT rate). Some studies have even used one approach to validate the other, that is, objective performance measures to validate subjective self-reports of mind wandering (e.g., Bastian & Sackur, 2013; Kane, Smeekens, et al., 2021; McVay & Kane, 2009, 2012a). Each of these measures, however, has their own unique sources of error which reduce our ability to accurately capture variation in the sustained attention construct.

This dissertation presents a set of studies that first provide evidence that objective and subjective measures of attention consistency may be influenced by a common underlying ability, and then tests this measurement approach in the contexts of two construct validation strategies: nomothetic span (correlational) and construct representation (experimental). Below, I discuss the implications that this work has for our measurement of sustained attention ability and next steps on how to further improve its measurement. Throughout, I discuss some of the limitations of the included empirical papers.

Implications for Sustained Attention Measurement

The current dissertation defines sustained attention as *the purposeful act of maintaining optimal task focus to successfully, and consistently, perform goal-relevant actions*. In this view, sustained attention ability has traditionally been measured using either objective performance measures (like RT variability or task accuracy), or subjective self-reports (like rates of mind wandering) during simple lab tasks over the course of a seconds to a few minutes. This attention consistency approach holds that trial-by-trial fluctuations in RT, performance accuracy, and conscious focus are indicative of sustained attention ability. This attention consistency approach differs from the vigilance approach to sustained attention, which requires subjects to respond to rare, unpredictable, targets, over many tens of minutes, and is primarily concerned with performance decrements (in accuracy or RT) over the entirety of a task.

While the objective and subjective measures described above likely reflect variation in sustained attention ability to some degree—and perhaps to different degrees (Cheyne et al., 2009)—they also reflect non-sustained attention processes. For example, objective indicators are also impacted by processes like general processing speed and speed-accuracy trade-offs. Subjective indicators are impacted by things like response biases and reactivity to task performance. Thus, studies using either indicator type on its own are not assessing sustained attention in a process-pure manner. To overcome this, the current empirical papers argue that looking at the individual-differences covariation in these indicators as a more construct-valid way to measure attention consistency than is either type on its own. The results of the papers provided general support for this claim.

In the first empirical paper (Welhaf et al., 2020b), evidence for the worst performance rule appeared when TUT rates were used a measure of cognitive (i.e., sustained attention) ability: The correlation between RTs and TUT rates increased from subjects' fastest to slowest RTs. This was not the case when WMC was used as the ability measure, with correlations being of similar magnitude for both average and worst performance (and weakest with best performance). From an attention control perspective (Larson & Alderton, 1990; Unsworth et al., 2010), subjects' worst performing (or slowest) trials and their TUT reports are both partial reflections of momentary failures of sustained attention. If this perspective is accurate, then both indicators should be explained by a common underlying ability.

The second empirical paper (Welhaf & Kane, 2022a) explicitly tested this measurement approach by reanalyzing two large-*N* datasets (Kane et al., 2016; Unsworth et al., 2021) both of which included multiple tasks to derive objective performance measures and probed mind wandering. Here, we found that the individual-differences covariation in objective and subjective could be modeled with a hierarchical structure. Critically, this higher-order factor showed a unique pattern of correlations with nomological network constructs. In some cases, the nomological network constructs (e.g., WMC, processing speed, positive schizotypy) correlated numerically more strongly with the higher-order factor than they did with *both* the individual objective and subjective factors. In other cases, the nomological network constructs correlated more strongly with the higher-order factor than they did with *either* the individual objective or subjective factor (e.g., neuroticism and self-reported cognitive failures were similarly correlated with TUT rate factor and the higher-order factor, and less strongly with the objective factor). These correlations provided evidence for convergent validity of the higher-order factor and suggest that previous correlations may have been underestimated when only relying on one of the indicator types as a measure of attention consistency. The higher-order factor did not correlate with measures of agreeableness or conscientiousness, however, even though both measures correlated with the individual TUT rate factor. This discriminant validity evidence also suggests that by only using TUT rates as an indicator of sustained attention, previous findings may not extend to a general sustained attention ability, but rather they may have been driven by processes shared only between subjective indicators and personality measures (e.g., self-report biases).

Empirical paper 3 (Welhaf & Kane, 2022b) expanded on the proposed measurement approach in the context of two experimental studies, in which we tested whether the individual-differences covariation between objective and subjective indicators of attention consistency was reduced by implementing a series of theoretically derived manipulations to minimize the sustained attention demands of prototypical tasks. While these manipulations did reduce mean levels of RT variability and TUTs rates in the demand-minimized tasks (in exploratory analyses), the correlation between these measures remained unchanged (in our primary analyses). Thus, we only found partial support for the construct validity of our sustained attention measurement approach.

Across the presented program of dissertation research, the results suggest general support for the covariation approach in measuring sustained attention. Empirical Papers 1 and 2 provide support for the hypothesis that there is an underlying ability explaining variation in objective and subjective indicators of attention consistency. Empirical Paper 2 further showed that this general ability could be modeled through the covariation of objective and subjective measures. Finally, Empirical Paper 3 showed that both indicator types could be impacted by theoretically derived manipulations. As discussed below, however, there are still some outstanding questions and concerns that future research needs to address.

Challenges for Measuring Sustained Attention

The current studies focused on two main types of indicators of sustained attention: objective performance measures and subjective self-reports of mind wandering. While these two indicator types have traditionally been the main ways of assessing attention consistency, the results of the current dissertation studies indicate some challenges when using only these indicators. Below I discuss these issues and propose some solutions that future measurement work could adopt.

Strengthening the Correlation between Objective and Subjective Indicators

The current studies found that objective and subjective indicators of attention consistency were only modestly correlated with each other (*r*s = .20–.40). We have argued that, despite a common sustained attention ability partially explaining variation in both measures, measurement error drives these correlations down. Are there alternative ways of assessing objective or subjective measures that might strengthen this association that might be useful for between-subject analyses?

First, our objective factor was derived from a mixture of indicators that captured not only performance errors, but measures of variable RTs (e.g., intra-individual RTsd) and long RTs (e.g., *tau* and Slowest 20%). We have argued that these dependent measures reflect attention consistency to some degree, but it is possible that variable RT measures are somewhat different than long RT measures, and by using this mixture we reduced the covariation among measures of the objective factor. Indeed, recent research has found that variable RTs, captured by the *sigma* component of the ex-Gaussian model (rather than long RTs captured by the *tau* component) were better at distinguishing subjects between "optimal" and "suboptimal" brain states associated with sustained attention (Yamashita et al., 2021). This variance component also correlated strongly with self-reported mind wandering, while the long RT component correlated less strongly (*rho* = .56 vs. .37, *p* = .0502; but note *n* = 29). Although other work has found TUT rates to correlate more strongly with *tau* than with *sigma* (McVay & Kane, 2012a; Welhaf et al., 2020), future research should explore using measures that more closely reflect variable RTs (i.e., RTsd and *sigma*) as objective indicators of sustained attention rather than, or in addition to, measures that capture especially long RTs (e.g., *tau* or Slowest 20%).

The present studies relied on RTs taken across entire tasks, but doing so might reduce the correlation between RT variability and TUTs because it combines periods of optimal (ceiling-level) and suboptimal sustained attention. An alternative approach may be to only use trials that occur before TUTs (e.g., 10 trials preceding TUT reports). As previously discussed, trials preceding TUT reports show more variable RTs and more errors compared to on-task reports (e.g., Bastian & Sackur, 2013; Kane, Smeekens et al., 2021). By analogy to the worst-performance rule, measuring RT variability at its most extreme, and when supported by thought off-task thought reports, may be a more construct-valid way to assess attention consistency via performance, because it isolates those moments that are most reflective of sustained attention ability. That is, it's possible that these pre-TUT trials are more reflective of a subjects' sustained attention ability than is RT variability across an entire task (which includes RT variability preceding on-task reports).

The methods used by Esterman and colleagues (e.g., Esterman et al., 2013; Fortenbaugh et al., 2018; Rosenberg et al., 2013) might provide a still more sophisticated approach. Here, trial-to-trial variation in RT is modeled within-subject using a variance time course, in which RTs are identified as being above or below a subjects mean RT. The variance time course is then smoothed by integrating information from some number of surrounding trials (e.g., 20 trials); this temporal choice in smoothing is based on previous work showing that attentional fluctuations occur on the order of 16–20 s (De Martino et al., 2008). After all RTs for each run are smoothed, low-versus high-variability periods are defined using a median split, yielding periods of "in the zone" versus "out of the zone" performance. Previous work using this approach has found that “out of the zone” periods are associated with poorer accuracy and greater RT variability (i.e., not simply stable periods of extremely fast or slow RTs; Rosenberg et al., 2013) compared to periods of “in the zone” performance. Thus, the variance time course analysis is useful for identifying two potentially distinct attention states.

Future studies, then, could use the variance time course approach and pull performance measures (i.e., errors and RTs) from “out of the zone” periods as an individual differences measure. Much like using RTs on the trials preceding TUTs (rather than on-task or complete task data), the data from these “out of the one” periods may be more reflective of sustained attention ability as they eliminate the influence of measurement error from “in the zone” periods. Thus, this approach may be a more construct-valid way for capturing attention consistency in continuous performance tasks like the SART or gradCPT.[17]

Improving the Hierarchical Model of Sustained Attention

While the hierarchical model fit the data well in both studies in Empirical Paper 2, it does present some limitations in its current form. Specifically, by having only two first-order factors, we had to set these paths to be equal to appropriately identify the model. Thus, future research should consider additional ways to measure attention consistency to enhance the first-order structure of the hierarchical model. As discussed in Empirical Paper 2 (Welhaf & Kane, 2022a), changes in pupil dilation provide another potential way to assess sustained attention. Variation in pupil size is proposed to be an indirect measure of locus coeruleus-norepinephrine (LC-NE)

[17] This discussion focuses on ways to improve the validity of assessments of in-the-moment RTsd by using periods that only occur before TUTs or during “out of the zone” periods. An additional approach, which might help improve TUT rate validity, is to only count TUT reports toward TUT rates if they also follow periods of relatively high RT variability (e.g., greater than the subject’s median RTsd). Here, then, one would only use TUT reports where there was also corresponding behavioral evidence of the subject’s attention being off task.

functioning, which is import for regulating arousal based on current attentional demands (Cohen et al., 2004). Critically, for the current proposed measurement approach, individual differences in pupil size variation (both pre-trial variability and variability in task-evoked pupillary responses [TEPRs]) correlate with objective and subjective measures of attention consistency.

Unsworth & Robison (2017a), for example, reported a modest negative correlation between pupil size variability (a latent factor reflecting shared variance between pre-trial and TEPR variability) and TUT rates (*r* = .23), suggesting that people who had more variable pupil size were also more likely to mind wander in simple attention tasks. Note that the correlation between *mean* baseline pupil and TUTs was nonsignificant, suggesting that pupil instability, and not average size, is likely a more indicative measure of attentional consistency. Likewise, Unsworth et al. (2020) found similar negative correlations between pre-trial pupil variability and the slowest 20% of trials and TUT rates in the PVT (*r*s = –.30 and –.22, respectively). Future studies of attention consistency measurement should therefore consider adding pupil measures alongside objective performance and subjective self-report measures. Doing so would add a third first-order factor to the hierarchical sustained attention model. That is, the hierarchical model would now model the individual-differences covariation between objective, subjective, and physiological indicators of attention consistency, which may help identify the hierarchical model proposed in Empirical Paper 2 without fixing paths to be equal (Welhaf & Kane, 2022a).

As discussed in the Integrated Introduction, objective indicators of attention consistency can be RT- or accuracy-based. In Empirical Paper 2 (Welhaf & Kane, 2022a), our objective factor was primarily RT-based, but some of our measures were accuracy based (e.g., SART omissions). A possible way forward for studies using only objective and subjective indicators could be to split the objective factor into separate accuracy- and RT-based factors (obviously requiring enough tasks from which to draw these). Studies would then also be able to build factors that more closely reflect the separate attentional states described by Cheyne et al. (2009; see also Unsworth et al., 2021): A state of *focal inattention* (State 1) characterized by brief periods of high instability of attention corresponding to increased errors, near misses, and variable responding, a state of *global inattention* (State 2) where automatic, "mindless," responding and processing overrides top-down control resulting in anticipatory responses, and a final state of *behavioral disengagement* (State 3) where subjects' attention is so withdrawn from the task that they completely fail to respond, resulting in errors of omission.

Likewise, future studies could use different thought-probe menus in different tasks to assess TUT propensity more broadly. Mind wandering studies have asked about a variety of dimensions of TUTs, including temporal orientation (e.g., Stawarczyk et al., 2011; 2013), emotional valence (e.g., Banks et al., 2016), and content descriptors (e.g., daydreams vs. personal worries; e.g., Kane et al., 2016; McVay & Kane, 2012b). Using a few different probe types across tasks, and creating different latent variables for each probe type, could again allow for a larger number of first-order sustained attention factors to allow the higher-order factor to be identified.

Limitations and Future Directions

Below I discuss some future lines of work in both the nomothetic span and the construct representation approaches that would add to the current set of studies and improve the field's evaluation of the construct validity of sustained attention measures.

Nomothetic Span Considerations

One limitation of the nomothetic span studies presented here (e.g., Empirical Papers 1 and 2) is that we were limited to what nomological network associations we could examine. That is, because we relied on previously collected data, we could only analyze associations with the collected constructs. In future work, however, several other nomological network constructs (and additional nomothetic span approaches [e.g., group comparisons]) might provide further construct validity evidence for the proposed covariation approach to assessing sustained attention.

Attention Control

One question that remains from the current studies is how “attention control” and sustained attention abilities are related. More specifically, is the ability to sustain attention simply another way of describing, and measuring, a specific component of attention control abilities? Theories of executive attentional control propose two primary dimensions, typically referred to as goal maintenance and competition resolution (e.g., Engle & Kane, 2004), or proactive and reactive control processes (e.g., Braver et al., 2007). Goal maintenance is often defined as one’s ability to activate and maintain task goals in the presence of (and in advance of) conflict (e.g., Kane & Engle, 2003). Likewise, proactive control is proposed to reflect how “goal relevant information is actively maintained in a *sustained* (emphasis added) manner, before the occurrence of cognitive demanding events, to optimally bias attention, perception and action systems in a goal-driven manner” (Braver, 2012, p. 106).

People with better goal maintenance/proactive control ability perform better on tasks because they can better activate and maintain task goals ahead of expected conflict. However, goals are not always strongly maintained over the course of a task, or even during the full course a single task trial, which may lead to errors or relatively long, or variable, RTs (e.g., Meier & Kane, 2017; Meier et al., 2018; Unsworth & Robison, 2020). Thus, momentary failures of goal maintenance/proactive control and failures of sustained attention are identical to one another.

This might explain why we were unable to examine the relationship between attention control and sustained attention in Empirical Paper 2 (Welhaf & Kane, 2022a). Recall that in Study 1, the attention control factor, which was mainly comprised of goal-maintenance type tasks (e.g, antisaccade, Stroop), correlated > 1.0 with the second-order sustained attention factor and led to model misfit and estimation issues forcing us to drop it from the CFA. Thus, it's possible that when an attention control factor is defined primarily by tasks that heavily require goal maintenance, it will be impossible to dissociate attention control from sustained attention constructs.

Attention consistency and goal maintenance might be indistinguishable due to their linked neural pathways. Consistency of attention is proposed to be regulated by the locus-coeruleus norepinephrine system (LC-NE). The LC has widespread projections to other areas of the brain including the fronto-parietal network (FPN) which has a been linked to goal-maintenance and proactive control abilities (e.g., Szabadi, 2013). The LC also has major inputs from the prefrontal cortex suggesting a bidirectional connection between LC-NE and FPN (Rajkowski et al., 2000). Indeed, some have recently argued that attention control (goal maintenance) errors can be explained by dysregulation of LC-NE functioning (manifesting in moment-to-moment fluctuations in attention consistency) and downstream fluctuations in the FPN activity (i.e., goal-maintenance failures; Unsworth & Robison, 2017). Thus, LC activity can determine moment-to-moment task activity levels by biasing FPN activity, which leads to accurate, and consistent, task performance.

What then could future work do to examine the link between goal maintenance/proactive control and attention consistency? One interesting approach would be to design a battery of tasks in which attention control was exclusively assessed in tasks that place a high demand on goal maintenance/proactive control to regulate conflict (e.g., antisaccade, high-congruency Stroop, AX-CPT) while modeling attention consistency from tasks that present little conflict (e.g., PVT, MRT, simple RT and low-choice RT tasks). If goal-maintenance and sustained attention abilities are two sides of the same coin, then even when making the tasks as independent as possible, we might expect a strong, if not perfect, correlation between these factors. However, if these two factors were to correlate less strongly (i.e., < .70), this might suggest that there is some possibility of dissociating goal maintenance and sustained attention abilities.

An alternative or additional approach would be to look for discriminant validity evidence between attention consistency and the response competition/reactive control components of attention control (rather than proactive control; Braver, 2012; Kane & Engle, 2003). These components of attention control can be thought of as "late-selection" mechanisms that are brought online only after conflict has been recognized or when no predictive information is available and needs to be corrected just-in-time to avoid an incorrect response (Braver, 2012). Based on prior work of reactive control, future studies could model a competition resolution/reactive control factor from flanker conflict tasks, a mostly incongruent Stroop task, and AX-CPT tasks with reactive strategy instructions (Cooper et al., 2017; Gonthier et al., 2016; Kane and Engle, 2003). I would propose that this competition resolution/reactive control factor would be still somewhat correlated with sustained attention ability because there is still a requirement for consistent task focus for successful completion of the task, but much less so than a goal maintenance/proactive control factor.

The Role of Motivation in Sustained Attention

Empirical Paper 2 (Welhaf & Kane, 2022a) attempted to investigate the individual differences link between self-reported motivation and sustained attention ability. However, we found that self-reported task motivation was too strongly correlated with the higher-order sustained attention ability (i.e., > 1.0), leading to severe misfit of the model. I do not think that motivation and attentional consistency are isomorphic constructs, however. First, the zero-order correlations indicated only a modest to moderate relationship between task-specific motivation and attention consistency indicators. For example, motivation in the PVT correlated with PVT slowest 20% and TUTs with *r*s = –.36 and –.46, respectively, and Choice RT and Continuous Tracking Motivation correlated modestly with objective indicators from those tasks (*r*s = –.21 and–.28, respectively). Thus, at the task level, motivation cannot fully explain attention consistency.

Recent work in the cognitive control literature has emphasized that performance is not only determined by one's ability (or capacity) but also by one's motivation (e.g., Braver et al., 2014; Shenhav et al., 2017; 2021). This interplay between cognition and motivation is also likely important for understanding attention consistency. The current dissertation has argued that sustained attention is an *ability* that is important for successfully completing a range of tasks, with some people having a better or worse ability than others. However, it could be that sustained attention ability is largely similar for everyone, but how well someone implements this ability is determined by how motivated they are to perform. Like other forms of cognitive control, engaging sustained attention comes with costs (e.g., a depletion of cognitive resources, or increased levels of stress, boredom, and general mental workload; see Thomson et al., 2015; Warm et al., 2008). With the appropriate benefits, however, one may be more willing to sustain attention to meet the demands of a given task. Indeed, recent work has argued that individuals balance exerting cognitive effort (i.e., engaging in sustained goal-directed thought) with disengaging (i.e., mind wandering) or resting from the current task (Kool & Botvinick, 2013, 2014; Sripada, 2018) and that variability in task performance may reflect a combination of cognitive ability and cognitive motivation (Westbrook et al., 2013).

How then can we begin to understand the relationship between motivation and attention consistency? Recent research using discounting paradigms have investigated whether there is a domain-general construct of cognitive motivation by examining how costs impact decision-making (Crawford et al., 2022; see also Westbrook et al., 2013). For example, a "cognitive effort discounting" paradigm has been used to investigate how participants make decisions between participating in high-effort tasks that payout a high reward versus low-effort tasks for less reward. Using this paradigm, participants make a series of choices until a level of subjective equivalence is reached (i.e., a point where low-effort and high-effort choices are equally rewarding). Researchers can vary how demanding the high-effort task is which can lead to more rewarding, but more costly (effortful) decisions. This might be a useful approach to better assess how willing people are to engage in sustained attention as a cognitively demanding activity, rather than simply asking about subjects' motivation to perform. For example, future studies could examine whether people who find more subjective costs in engaging in cognitively effortful tasks show less attentional consistency via objective and subjective indicators.

ADHD Symptoms

Attention-hyperactivity deficit disorder (ADHD) is defined by three clusters of symptoms including inattention, hyperactivity, and impulsivity (American Psychiatric Association, 2013). A frequent observation about individuals with ADHD is that they are "consistently inconsistent" (Karalunas, 2010). Indeed, the most prominent symptoms associated with the inattention cluster are poor sustained attention and distractibility. While usually investigated in children, some studies have investigated how ADHD-related sustained attention deficits play out into young adulthood.

First, with respect to objective measures of attention consistency, adults with diagnosed ADHD show greater RT variability compared to typically developing adults (*g* = 0.46; Kofler et al., 2013; Shahar et al., 2016). Additionally, children with ADHD, compared to controls, appear to show a clear periodicity of long RTs, with relatively long RTs manifesting roughly every 20 s, suggesting that lapses of attention occur in cycles (Castellanos et al., 2005; Vaurio et al., 2009). Further, self-reported ADHD symptoms correlate with a latent variable measure of RT variability (*r* = .23; Brydges et al., 2021; see also Keith et al., 2017). Thus, individuals with greater self-reported, or diagnosed, attention problems tend to show poorer sustained attention performance in simple attention tasks than those with fewer self-reported symptoms or healthy controls.

Second, individuals who self-report more (or more severe) ADHD symptoms also report more TUTs during basic attention tasks than do those with fewer self-reported symptoms or healthy controls (e.g., Franklin et al., 2017; McVay & Kane, 2013; Meier, 2021; but see Kane, Smeekens et al., 2021). Collectively, self-reported ADHD symptoms appear to be related to both objective and subjective indicators of sustained attention, separately (*r*s ≈ .15–.25 and .25–.35, respectively), and should thus be correlated with their covariation (i.e., with a common sustained attention factor). If ADHD symptoms are truly related to general sustained attention ability, then I would predict that the correlation between these two factors would be stronger than the correlation between ADHD symptoms and both objective and subjective indicators (i.e., a correlation ≈ .25 with the general factor). Much like how the correlations with other self-report measures (e.g., neuroticism, cognitive failures, positive schizotypy) in the current nomothetic span studies appear to average out when looking at the correlation with the higher-order sustained attention factor, I would expect self-reported ADHD symptoms to follow the same pattern.

Learning and Memory

The ability to maintain consistent focus during a task should be related to how well people learn new information from it. That is, being consistently focused should allow people better to encode information as it is being encountered; failing to sustain attention (i.e., thinking about task-unrelated topics) should predict poorer learning. Indeed, previous research has found that both objective and subjective measures of attention consistency predict learning in a variety of tasks and contexts (for a review, see Blondé et al., 2022). For example, deBettencourt et al. (2018) had subjects complete a go/no-go task using indoor vs. outdoor images followed by a surprise recognition memory test. Sustained attention was indexed as the average RT on the three trials preceding the test item during the go/no-go test. The findings indicated that slower pre-test item RTs predicted better performance, suggesting that better in-the-moment sustained attention allowed for better encoding of the items (see also Smallwood et al., 2006).[18]

Between-subject findings parallel those of deBettencourt et al. (2018): Creating a latent "attention" factor using pupil variability, commission errors, and RT variability from a gradCPT task (a go/no-go task with gradual rather than abrupt stimulus transitions), Madore et al., (2020)

[18] deBettencourt et al.'s (2018) measurement of sustained attention does raise some concerns. Although others have argued that faster responses during continuous performance tasks are associated with attentional lapses, as these reflects short periods of mindless, habitual, responding, rather than careful processing of stimulus characteristics (Robertson et al., 1997), pre-target *M* RT might reflect changes in in-the-moment strategy instead of sustained attention engagement. As such, I don't agree that this is a suitable measure of in-the-moment attention consistency and instead would suggest using variability in RT as an improved measurement approach.

found a strong negative correlation with performance on episodic memory tasks (r = –0.52). This "attention" factor loosely resembles how sustained attention was (and could be) assessed in the current nomological network studies and suggests that better sustained attention should be related to better learning and memory performance.

Likewise, TUT rates during encoding predict poorer learning performance, such that people who mind wander more during the encoding period also tend to recall less at test. This pattern of results has been found in a range of contexts, including traditional lab memory tasks (rs ≈ –.15 to –.35; e.g., Garlitch & Wahlheim, 2020; Miller & Unsworth, 2021; Thomson et al., 2014; Xu & Metcalfe, 2016), in more naturalistic tasks like video-lecture learning (rs ≈ –.30 to –.50; e.g., Hollis & Was, 2016; Jing et al., 2016; Kane et al., 2017; Welhaf et al., 2022), and in live classroom settings (rs ≈ –.15 to –.20; e.g., Kane, Carruth, et al., 2021; Wammes, Seli et al., 2016). Collectively, then, both objective and subjective indicators of attention consistency are related to individual differences in learning ability. I would therefore predict that a higher-order sustained attention factor would correlate moderately with learning ability (r ≈ –.35 to –.45).

Negative Affect

Emotions can impact memories, perception, and other goal-directed behaviors (Forgas, 2008). It should not be surprising, then, that negative emotional states likely impact attentional consistency, with more negative affect leading to more variable performance and more frequent TUTs. Two possible reasons for this are that: (a) changes in affect may reduce the attentional resources that one can devote to a task or, (b) negative emotional states capture attention to a greater degree which disrupts ongoing task performance (Jefferies et al., 2008; Lazarus, 1999; Olivers & Nieuwnhuis, 2005). Further, previous research has argued that stress-induced TUTs may require top-down suppression to reduce their impact on the primary task (Klein & Boals, 2001; see also Wegner, 1994). In this view, a more negative emotional state should be related to greater RT variability and increased TUT rates, and thus poorer general sustained attention ability.

Negative affect predicts objective measures of attention consistency at both within- and between-subject levels. At the within-subject level, Sliwinski et al. (2006) reported that on days where subjects reported feeling more stressed (compared to days with little to no stress), they also had an increase in relatively long RTs in a 2-back task. As well, subjects who underwent a negative mood induction (compared to a positive or neutral induction) showed greater RT variability following the induction (Irrmischer et al., 2018). Between-subject analyses also suggest that trait negative emotionality predicts sustained attention failures in simple lab tasks: People who report higher levels of depression or trait negative affect make more errors in sustained attention tasks (Farrin et al., 2003; Mrazek et al., 2012) and have more variable RTs (Ode et al., 2011, Studies 3 and 4).

The link between mood state and TUTs (i.e., subjective measures of attention consistency) has also been well established at the within- and between-subject levels. For example, inducing negative affect in the lab leads to increased TUT rates (e.g., Marcusson-Clavertz et al., 2020; Smallwood et al., 2009; Smallwood & O'Connor, 2011; Stawarczyk et al., 2013). As well, in daily life studies of mind wandering, subjects report increased levels of negative affect during times of mind wandering versus on-task thinking (e.g., Kane et al., 2007, 2017; Killingsworth & Gilbert, 2010; Poerio et al., 2013; Song & Wang, 2012). At the between-subject level, self-reported negative affect correlates with TUT rate in simple attention and working memory tasks, with higher levels of negative affect predicting greater TUT rate (e.g., Banks & Welhaf, 2022; Robison et al., 2020; Ruby et al., 2013). Collectively, then, negative affect may be negatively related to the general ability to sustain attention, such that people with higher trait levels of negative mood also show poorer attention consistency.

Healthy Aging

Healthy aging is associated with declines in multiple cognitive processes, including processing speed, inhibitory control, and general executive control (e.g., Braver & West, 2008; Hasher & Zacks, 1988; Salthouse, 1996). Theories of cognitive aging might therefore predict that sustained attention ability should also worsen with age. However, the literature appears to be mixed: Some studies find that RT variability tends to increase with adult age (e.g., Der & Deary, 2006; Hultsch et al., 2002; Robison et al., 2022; West et al., 2002), whereas others report equivalent or even reduced levels of RT variability in older versus younger adults (Moran et al., 2021; Nicosia & Balota, 2021; Waugh et al., 1973). The link between mind wandering and aging is more one-sided—and surprising: Older adults report fewer TUTs during lab tasks compared to younger adults (meta-analytic estimate $g = -.89$, Jordão et al., 2019), and report fewer mind wandering experiences in everyday life (Maillet et al., 2018), suggesting that older adults may *not* experience more sustained attention failures.

What might explain why the trajectories of objective and subjective indicators of sustained attention in older adults appear to go in opposite directions? A hallmark finding in the cognitive aging literature is that older adults generally have slower processing speed compared to younger adults (Salthouse, 1996). Given the strong correlation between mean RT (which is often used as an indicator for processing speed) and RT variability, it is not surprising that older adults often show greater RT variability compared to younger adults. When attention consistency is measured by objective indicators, then, it might be picking up on age-related differences in, or effects of, processing speed, rather than actual sustained attention differences.

Alternatively, or in addition, older adults might intentionally slow down during laboratory tasks to ensure more accurate performance. That is, in tasks that rely on both speed and accuracy, older adults might favor accuracy over speed (Vallesi et al., 2021). This intentional slowing should also selectively contribute to RT variability but not TUT rate. Thus, because processing speed and speed-accuracy trade-offs selectively impact RT variability (with no impact on TUT rates), any studies of attention consistency in older adults that relies *only* on objective indicators may not present accurate conclusions about the effects of aging on attention consistency. Rather, I suggest that it is necessary to look at the covariation between objective and subjective indicators of attention consistency, as this measure is not influenced by processes that selectively influence either objective or subjective indicators.

Discriminant Validity Evidence

If sustained attention is a fundamental process for the regulation of thought and behavior, it may be difficult to find constructs that are *unrelated* (i.e., $r = 0.00$) to measures of attention consistency. However, there may be some constructs that are *relatively* weakly associated with attention consistency, providing evidence for discriminant validity. Empirical Paper 2 (Welhaf & Kane, 2022a) presented some initial evidence for discriminant validity of the higher-order sustained attention factor: Correlations with certain personality traits such as agreeableness and conscientiousness were nonsignificant with the higher-order factor despite correlating weakly with the individual subjective factor. Replicating this pattern (i.e., null correlations with the higher-order factor despite significant correlations with one of the first-order factors) would be useful in confirming this discriminant validity evidence.

While sustained attention appears to be related to WMC, it may be more weakly correlated with short-term memory (STM) measures. “Simple” STM tasks do not have a strong executive attention requirement, as do traditional “complex span” measures of WMC (e.g., Engle et al., 1999; Kane et al., 2004; Unsworth & Engle, 2007). Thus, STM is *more* reflective of simple storage (and less of executive control), whereas measures of WMC are *more* reflective of executive control abilities (and less of simple storage). Attention consistency might therefore have weaker associations with STM ability than with WMC. No studies that I’m aware of have yet investigated how RT variability or mind wandering relate to STM.

Previous research has indicated that attention is critical to creative cognition. For example, the *controlled attention* theory of creativity (e.g., Beaty et al., 2014; Benedek et al., 2012) argues that focused attention can allow for better memory search and generating more novel ideas (e.g., Gilhooly et al., 2007, Nusbaum & Silvia, 2011; Zabelina et al., 2016). Other work has also argued, in contrast, that a *lack* of attention (i.e., mind wandering) may support creative cognition. Here, mind wandering supports unconscious associative thinking which allows for spreading activation and subsequent generation of more creative ideas, perhaps especially during periods of incubation, where individuals are not fixated on a problem (e.g., Baird et al., 2012).

The ability to maintain consistent focus (i.e., sustain attention) may be useful in supporting creative thinking, but the current evidence is mixed. Generating creative ideas appears to sometimes correlate positively with mind wandering propensity, with people who are more prone to mind wandering generating more creative ideas (e.g., Gable et al., 2019). In contrast, measures of creativity are sometimes weakly *negatively* correlated with TUT rate (Frith et al., 2021; Hao et al., 2015; Murray et al., 2021) or produce null associations (Smeekens & Kane, 2016; Steindorf et al., 2021). There also appear to be no studies explicitly linking objective indicators of attention consistency with measures of creativity or divergent thinking.

Frith et al. (2021) included SART intra-individual RT variability as an indicator in their attention control factor, but zero-order correlations between SART RTsd and creativity ratings were weakly negative (*r*s –.25 to –.10), indicating that greater variability was related to poorer divergent thinking. Given the mixed empirical results regarding TUTs and creativity, and the lack of studies examining RT variability and creativity, future work should consider exploring the creativity–attention consistency connection as evidence for discriminant validity of the higher-order sustained attention factor.

Considerations for Future Construct Representation Studies

While the nomothetic span studies focused on correlational evidence for the construct validity of sustained attention measures, the construct representation studies presented in Empirical Paper 3 took an experimental approach (combined with individual differences) to understand and identify the cognitive processes that cause variation in task performance. We hypothesized that minimizing the sustained attention demands by implementing theoretically derived manipulations would reduce (or even eliminate) the covariation between objective and subjective indicators of attention consistency. However, in both studies, this hypothesis was not supported: We found no significant decrease in the correlation between objective and subjective indicators in the minimized tasks, despite medium-sized experimental reductions in mean levels of RT variability and TUT rates. Empirical Paper 3 discusses some possible explanations for why we failed to find a weaker correlation in the minimized conditions and so I will only briefly revisit them here. Subsequently, I will discuss additional methods that might be used to not only further minimize sustained attention demands for "minimized" tasks, but also to maximize demands for "maximized" tasks, in order to sufficiently impact the covariation between objective and subjective measures of attention consistency.

Methods to Further Minimize Sustained Attention Demands

A promising manipulation to reduce the sustained attention demands of a task is increasing subjects' motivational state. Under highly motivating conditions, subjects show improved objective performance measures (e.g., lower RT variability and fewer errors) and subjective measures (e.g., fewer TUT reports). Providing subjects with attainable goals (e.g., Esterman et al., 2014; Seli et al., 2019) or instructions to increase their effort (e.g., Unsworth et al., 2022), or giving them rewards or feedback on their performance (e.g., Massar et al., 2016, 2019; Robison et al., 2021), or combinations of these, should further facilitate subjects remaining focused for the duration of the task. To maximize these motivational effects, it may be necessary to provide such feedback or incentives relatively frequently (even on a trial-by-trial basis) if we are interested in reducing trial-to-trial variability due to sustained attention lapses. That is, more frequent rewards or goal reminders may be necessary for improving sustained attention as currently defined (see Hood & Hutchinson, 2021, for an example of frequent goal reminders improving Stroop task performance).

Altering the frequency of thought probes may be another way to exogenously redirect subjects' attention back to the task. First, more frequent probing will provide subjects with additional task breaks. That is, subjects only must maintain their optimal focus for a brief time, before getting another break in the task occurs and allows them to reset. These shortened durations of task engagement may minimize lapses by not allowing enough time between probes for subjects' minds to wander or for falling into repetitive, mindless responding. Additionally, probes, themselves, may remind subjects that they need to keep their thoughts on task, and so more frequent probing will increase the frequency of those "stay on task" reminders.

Whereas some studies have found that frequent probing reduces TUT rates in a task (e.g., Greve & Was, 2022; Seli, Carriere et al., 2013; Schubert et al., 2019; but see Robinson et al., 2019), evidence for the impact of thought probe frequency on objective measures of sustained attention is less clear. If increasing the frequency of thought probes impacts the underlying mechanisms of sustained attention, then we would expect to find that both indicators, and importantly their covariation, are similarly impacted. However, if thought probe frequency only alters the likelihood of reporting TUTs, as found by Seli, Carriere et al. (2013), then it might not be a strong enough manipulation to impact sustained attention generally. To see whether thought probe frequency affects attention consistency, future research should consider how frequent versus infrequent probing changes the covariation between objective and subjective indicators.

Another way to redirect attention back towards a task may be through providing subjects with periodic alerting or warning signals of upcoming critical task events. For example, using the SART, some studies have found that providing reliable (compared to unreliable) warning cues about upcoming important task events (i.e., rare no-go trials) reduced commissions errors and produced faster RTs (Dang et al., 2022; Finkbeiner et al., 2015; Helton et al., 2011). Likewise, in the PVT, warning tones (both reliable and unreliable) could be added toward the end of each ISI to see whether such manipulations reduce lapses or long RTs as well as TUTs. If warning signals can serve as a temporary reminder that a critical trial is upcoming or that a trial is about to start, then we would expect to see reduced RT variability and TUT rates (and critically their covariation) in the warning signal task compared to the control task.

Methods to Maximize Sustained Attention Demands

The methodological approach used in the construct representation studies (Empirical Paper 3) focused exclusively on trying to *minimize* the sustained attention demands of prototypical tasks. However, our “maximized” tasks—reflecting the standard implementations of prototypical sustained attention asks—may not have actually *maximized* their sustained attention demand. As discussed briefly in Empirical Paper 3, to create demand-maximized tasks, we could have implemented the opposing manipulations used in the minimized tasks (e.g., requiring more frequent responding in the SART by reducing no-go trials, increasing time between trials in the MRT, increasing the length, or variability, of SOAs in the PVT). Future studies should consider minimizing *and* maximizing sustained attention demands within a study to see such effects. Below I suggest some additional methods for maximizing demands.

As previously discussed, manipulations of emotional state appear to affect both objective (Irrmischer et al., 2018) and subjective (Marcusson-Clavertz et al., 2020; Smallwood et al., 2009; Smallwood & O’Connor, 2011; Stawarczyk et al., 2013) indicators of attention consistency. Specifically, under negative mood manipulations (i.e., inducing sadness or stress/anxiety) subjects show more variable and erroneous responding and greater TUT rates compared to control conditions. Future work, then, could use similar mood manipulations before sustained attention tasks to compare how the covariation between objective and subjective indicators changes compared to control condition.

As an extension of manipulating affect, cuing subjects’ personal concerns may be another approach to maximize sustained attention demands. Cuing of personal concerns should increase the demands on sustained attention (compared to low- or non-cued condition) because such concerns should create greater interference with task goals. Indeed, previous lab work had found that cuing subjects with their personal goals or concerns (compared to non-goal related cues) can increase TUT rates (e.g., Kopp et al., 2015; McVay & Kane, 2013; Vannucci et al., 2017). These cuing manipulations have primarily been used in the context of mind wandering research, but one study (McVay & Kane, 2013) reported null effects on objective sustained attention indicators. Across four experiments, McVay & Kane (2013) found no significant difference on SART no-go accuracy for personal-goal cues (*M* = .45) vs. other-goal cues (*M* = .44). Thus, cuing of subjects' goals might be enough to alter subjects' TUTs but not their performance. However, it is possible that the effect of cuing subjects' goals might be there if more appropriate (or subtle) objective sustained attention measures were used (e.g., RTsd or *tau*). Future work should investigate the impact that personal cues that are embedded within tasks have on objective measures more reflective of attention consistency, and critically, the covariation between objective and subjective measures.

A final approach to maximizing the demands of the sustained attention tasks could be to make the tasks longer or less interesting, which should reduce subjects' motivation or effort to perform. Previous work has argued that mind wandering increases with time on task because subjects' motivation to perform wavers (Esterman et al., 2016; Thomson et al., 2015). From an opportunity cost perspective (Kurzban et al., 2013; Thomson et al., 2015) subjects may gain more information about the costs of engaging in longer, monotonous tasks, especially when tasks start being interpreted by subjects as long (e.g., over 30 min). Subjects typically don't know how long a task will last, and so their motivation and effort may remain stable for some time; at some point, however, they may decide it is no longer worth the cost to put in the same effort. In these cases, mind wandering should increase, and performance should become more variable. For example, Brosowksy et al., (2020) found that within-subjects, MRT RRT and mind wandering increased with time-on-task, and this was coupled with a decrease in motivation. Future research could examine how task length impacts the correlation between objective and subjective indicators of attention consistency by varying how long subjects must complete a task for (i.e., a 10 min SART vs. a 25 min SART).

Conclusions

The ability to sustain attention is a fundamental cognitive function to basic and complex intellectual processes and outcomes. Current measurement approaches, however, limit our understanding of attention consistency by relying on separate "objective" or "subjective" approaches that have their own unique source of measurement error. The current integrated dissertation argues that, because both objective (performance-based) and subjective (self-report-based) measures of sustained attention capture only some variance related to sustained attention, a more construct valid measurement approach is to assess their covariation. It is here—in the individual-differences overlap between objective and subjective measures—where we should best capture sustained attention ability, independent of sources of measurement error unique to either of these indicator types. The current dissertation studies highlight the importance of improved measurement and discusses strengthens, outstanding concerns, and future directions to further improve the proposed measurement approach.

REFERENCES

Albert, D. A., Ouimet, M. C., Jarret, J., Cloutier, M. S., Paquette, M., Badeau, N., & Brown, T. G. (2018). Linking mind wandering tendency to risky driving in young male drivers. *Accident Analysis & Prevention*, *111*, 125-132.

Allen, M., Smallwood, J., Christensen, J., Gramm, D., Rasmussen, B., Gaden Jensen, C., ... & Lutz, A. (2013). The balanced mind: the variability of task-unrelated thoughts predicts error-monitoring. *Frontiers in Human Neuroscience*, *7*, 743.

Anderson, T., Petranker, R., Lin, H., & Farb, N. A. S. (2021). The metronome response task for measuring mind wandering: Replication attempt and extension of three studies by Seli et al. *Attention, Perception, & Psychophysics, 83*(1), 315–330.

Antrobus, J. S. (1968). Information theory and stimulus-independent thought. *British Journal of Psychology*, *59*(4), 423-430.

Antrobus, J. S., Singer, J. L., & Greenberg, S. (1966). Studies in the stream of consciousness: experimental enhancement and suppression of spontaneous cognitive processes. *Perceptual and Motor Skills*, *23*(2), 399-417.

Anwyl-Irvine, A. L., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, 1407-1425.

Anwyl-Irvine, A.L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*, 388–407.

Ariga, A., & Lleras, A. (2011). Brief and rare mental "breaks" keep you focused: Deactivation and reactivation of task goals preempt vigilance decrements. *Cognition*, *118*(3), 439-443.

Arnau, S., Löffler, C., Rummel, J., Hagemann, D., Wascher, E., & Schubert, A. L. (2020). Inter-trial alpha power indicates mind wandering. *Psychophysiology*, *57*(6), e13581.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience, 28,* 403–450.

Bakker, M., & Wicherts, J. M. (2014) Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS ONE, 9.*

Baldwin, C. L., Roberts, D. M., Barragan, D., Lee, J. D., Lerner, N., & Higgins, J. S. (2017). Detecting and quantifying mind wandering during simulated driving. *Frontiers in Human Neuroscience*, *11*, 406.

Banks, J. B., Welhaf, M. S., Hood, A. V., Boals, A., & Tartar, J. L. (2016). Examining the role of emotional valence of mind wandering: All mind wandering is not equal. *Consciousness and Cognition*, *43*, 167-176.

Bargh, J. A. (1982). Attention and automaticity in the processing of self-relevant information. *Journal of Personality and Social Psychology*, *43*(3), 425.

Bastian, M., & Sackur, J. (2013). Mind wandering at the fingertips: automatic parsing of subjective states based on response time variability. *Frontiers in Psychology*, *4*, 573.

Baumeister, A.A., & Kellas, G. (1968). Intrasubject response variability in relation to intelligence. *Journal of Abnormal Psychology, 73*, 421-423.

Bedi, A., Russell, P. N., & Helton, W. S., (2022). Go-stimuli probability influences response bias in the sustained attention to response task: A signal detection theory perspective. *Psychological Research,* https://doi.org/10.1007/s00426-022-01679-7

Bellgrove, M. A., Hester, R., & Garavan, H. (2004). The functional neuroanatomical correlates of response variability: Evidence from a response inhibition task. *Neuropsychologica, 42*, 1910-1916.

Bertelson, P., & Joffe, R. (1963). Blockings in prolonged serial responding. *Ergonomics*, *6*(2), 109-116.

Bills, A. G. (1931). Blocking: A new principle of mental fatigue. *The American Journal of Psychology*, *43*(2), 230-245.

Bills, A. G. (1935). Fatigue, oscillation, and blocks. *Journal of Experimental Psychology*, *18*(5), 562.

Bornovalova, M. A., Choate, A. M., Fatimah, H., Peterson, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis on psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry, 88,* 18-27.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061.

Broadbent, D. E., Cooper, P. F., FitzGerald, P., & Parkes, K. R. (1982). The cognitive failures questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology*, *21*(1), 1-16.

Brosowsky, N. P., DeGutis, J., Esterman, M., Smilek, D., & Seli, P. (2020). Mind wandering, motivation, and task performance over time: Evidence that motivation insulates people from the negative effects of mind wandering. *Psychology of Consciousness: Theory, Research, and Practice*.

Bunce, D., MacDonald, S. W., & Hultsch, D. F. (2004). Inconsistency in serial choice decision and motor reaction times dissociate in younger and older adults. *Brain and Cognition*, *56*(3), 320-327.

Burgoyne, A. P., & Engle, R. W. (2020). Attention control: A cornerstone of higher-order cognition. *Current Directions in Psychological Science, 29(6),* 624-630.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81.

Cheyne, J. A., Carriere, J. S., & Smilek, D. (2006). Absent-mindedness: Lapses of conscious awareness and everyday cognitive failures. *Consciousness and Cognition*, *15*(3), 578-592.

Cheyne, J. A., Solman, G. J., Carriere, J. S., & Smilek, D. (2009). Anatomy of an error: A bidirectional state model of task engagement/disengagement and attention-related errors. *Cognition*, *111*(1), 98-113.

Chiew, K. S., & Braver, T. S. (2013). Temporal dynamics of motivation-cognitive control interactions revealed by high-resolution pupillometry. *Frontiers in Psychology*, *4*, 15.

Christoff, K., Fox, K. C. R. (2018). The Oxford Handbook of Spontaneous Thought: Mind-Wandering, Creativity, and Dreaming, 1st ed.; Oxford University Press: United States.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences*, *106*(21), 8719-8724.

Cohen, J. D., Aston-Jones, G., & Gilzenrat, M. S. (2004). A systems-level perspective on attention and cognitive control: Guided activation, adaptive gating, conflict monitoring, and exploitation vs. exploration. In M. I. Posner (Ed.), Cognitive neuroscience of attention (pp. 71–90). New York: Guilford Press.

Coyle, T.R. (2001). IQ is related to the worst performance rule in a memory task involving children. *Intelligence*, *29*, 117-129.

Coyle, T.R. (2003a) A review of the worst performance rule: Evidence, theory, and alternative hypotheses. *Intelligence*, *6*, 567-587.

Coyle, T.R. (2003b) IQ, the worst performance rule, and Spearman's law: A reanalysis and extension. *Intelligence, 31*, 473-489.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Dang, J. S., Figueroa, I. J., & Helton, W. S. (2018). You are measuring the decision to be fast, not inattention: the Sustained Attention to Response Task does not measure sustained attention. *Experimental brain research*, *236*(8), 2255-2262.

De Jong, R., Berendsen, E., & Cools, R. (1999). Goal neglect and inhibitory limitations: Dissociable causes of interference effects in conflict situations. *Acta Psychologica*, *101*(2-3), 379-394.

Denzin, N. K. (1970). *The Research Act in Sociology.* Chicago: Aldine.

Dinges, D. F., & Powell, J. W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior research methods, instruments, & computers*, *17*(6), 652-655.

Doebler, P., & Scheffler, B. (2016). The relationship of choice reaction time variability and intelligence: A meta-analysis. *Learning and Individual Differences, 52,* 157-166.

Duncan, T.E., Duncan, S.C., & Strycker, L.A. (2006). An introduction to latent variable growth curve modeling: Concepts, issues, and applications. Erlbaum: New Jersey, U.S.A.

Dutilh, G., Vandekerckhove, J., Ly, A., Matzke, D., Pedroni, A., Frey, R., Rieskamp, J., Wagenmakers, E. J. (2017). A test of diffusion model explanation of the worst performance rule using preregistration and blinding. *Attention, Perception, & Psychophysics, 79,* 713-725.

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods, 22*, 541-562.

Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor models for predicting criteria by general and specific factors: Problems of nonidentifiability and alternative solutions. *Journal of Intelligence, 6(3)*, 42. https://doi.org/10.3390/jintelligence6030042

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343-368.

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science, 11(1),* 19–23. https://doi.org/10.1111/1467-8721.00160

Engle, R.W., & Kane, M.J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation, volume 44* (pp. 145 - 199). New York: Academic Press.

Esterman, M., & Rothlein, D. (2019). Models of sustained attention. *Current Opinion in Psychology, 29*, 174-180.

Esterman, M., Noonan, S. K., Rosenberg, M., & DeGutis, J. (2013). In the zone or zoning out? Tracking behavioral and neural fluctuations during sustained attention. *Cerebral cortex*, *23*(11), 2712-2723.

Esterman, M., Reagan, A., Liu, G., Turner, C., & DeGutis, J. (2014). Reward reveals dissociable aspects of sustained attention. *Journal of Experimental Psychology: General, 143*, 2287–2295

Esterman, M., Rosenberg, M. D., & Noonan, S. K. (2014). Intrinsic fluctuations in sustained attention and distractor processing. *Journal of Neuroscience*, *34*(5), 1724-1730.

Fernandez, A.B., Fagot, C.D., Dirk, E.F., & de Ribaupierre, G.H. (2014). Generalization of the worst performance rule across the lifespan. *Intelligence, 42*, 31-43.

Fiske, D. W., & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin*, *52*(3), 217.

Fortenbaugh, F. C., DeGutis, J., & Esterman, M. (2017). Recent theoretical, neural, and clinical advances in sustained attention research. *Annals of the New York Academy of Sciences*, *1396*(1), 70.

Franklin, M. S., Mooneyham, B. W., Baird, B., & Schooler, J. W. (2014). Thinking one thing, saying another: The behavioral correlates of mind-wandering while reading aloud. *Psychonomic Bulletin & Review*, *21*(1), 205-210.

Frischkorn, G.T., Schubert, A., Neubauer, A.B., Hagemann, D. (2016). The worst performance rule as moderation: New methods for worst performance analysis. *Journal of Intelligence, 4*, 1-22.

Giambra, L. M. (1995). A laboratory method for investigating influences on switching attention to task-unrelated imagery and thought. *Consciousness and Cognition*, *4*(1), 1-21.

Gibson, W. A. (1959). Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika, 24,* 229–252.

Gollwitzer, P.M., & Bargh, J.A. (Eds.). (1996). The psychology of action: Linking cognition and motivation to behavior. New York: Guilford Press.

Gopher, D., Armony, L., & Greenshpan, Y. (2000). Switching tasks and attention policies. *Journal of Experimental Psychology: General, 129*, 308–339.

Grodsky, A., & Giambra, L.M. (1990–1991). The consistency across vigilance and reading tasks of individual differences in the occurrence of task-unrelated and task-related images and thoughts. *Imagination, Cognition and Personality, 10,* 39–52.

Hawkins, G. E., Mittner, M., Forstmann, B. U., & Heathcote, A. (2019). Modeling distracted performance. *Cognitive Psychology*, *112*, 48-80.

He, J., Becic, E., Lee, Y. C., & McCarley, J. S. (2011). Mind wandering behind the wheel: performance and oculomotor correlates. *Human Factors*, *53*(1), 13-21.

Head, J., & Helton, W. S. (2014). Sustained attention failures are primarily due to sustained cognitive load not task monotony. *Acta Psychologica*, *153*, 87-94.

Head, J., & Helton, W. S. (2018). The troubling science of neurophenomenology. *Experimental Brain Research*, *236*(9), 2463-2467.

Heathcote, A., Brown, S., & Cousineau, D. (2004). QMPE: Estimating Lognormal, Wald, and Weibull RT distributions with a parameter-dependent lower bound. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 277-290.

Helton, W. S. (2009). Impulsive responding and the sustained attention to response task. *Journal of Clinical and Experimental Neuropsychology*, *31*(1), 39-47.

Helton, W. S., & Russell, P. N. (2015). Rest is best: The role of rest and task interruptions on vigilance. *Cognition*, *134*, 165-173.

Helton, W. S., & Russell, P. N. (2017). Rest is still best: The role of the qualitative and quantitative load of interruptions on vigilance. *Human Factors*, *59*(1), 91-100.

Helton, W. S., Weil, L., Middlemiss, A., & Sawers, A. (2010). Global interference and spatial uncertainty in the Sustained Attention to Response Task (SART). *Consciousness and Cognition*, *19*(1), 77-85.

Hollenbeck, J. R., Ilgen, D. R., Tuttle, D. B., & Sego, D. J. (1995). Team performance on monitoring tasks: An examination of decision errors in contexts requiring sustained attention. *Journal of Applied Psychology*, *80*(6), 685.

Hollis, R. B., & Was, C. A. (2016). Mind wandering, control failures, and social media distractions in online learning. *Learning and Instruction*, *42*, 104-112.

Hultsch, D. F., MacDonald, S. W., & Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(2), P101-P115.

Hurlburt, R. T., & Heavey, C. L. (2001). Telling what we know: describing inner experience. *Trends in Cognitive Sciences, 5*, 400–403.

Hutchison, K. A., Moffitt, C. C., Hart, K., Hood, A. V., Watson, J. M., & Marchak, F. M. (2020). Measuring task set preparation versus mind wandering using pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 280.

Jackson, J. D., & Balota, D. A. (2012). Mind-wandering in younger and older adults: converging evidence from the Sustained Attention to Response Task and reading for comprehension. *Psychology and Aging*, *27*(1), 106.

Jackson, J. D., Weinstein, Y., & Balota, D. A. (2013). Can mind-wandering be timeless? Atemporal focus and aging in mind-wandering paradigms. *Frontiers in Psychology*, *4*, Article 742. doi:10.3389/fpsyg.2013.00742

James, W. (1890). The principles of psychology, Vol. 1. Henry Holt and Co.

Jensen, A. R. (1987a). Individual differences in the Hick paradigm. In Vernon, P. A. (Ed.), *Speed of information processing and intelligence*. Norwood, NJ: Ablex.

Jensen, A. R. (1992). The importance of intraindividual variation in reaction time. *Personality and Individual Differences*, *13*(8), 869-881.

Jensen, A.R. (1987). Individual differences in the Hick paradigm. In Speed of information-processing and intelligence; P.A. Vernon Ed.; Ablex: New Jersey, U.S.A.

Jensen, A.R. (1982). Reaction time and psychometric "g". In A Model for Intelligence; H.J. Eysenck Ed.; Plenum: New York, U.S.A.

Jones, A. D., Cho, R. Y., Nystrom, L. E., Cohen, J. D., & Braver, T. S. (2002). A computational model of anterior cingulate function in speeded response tasks: Effects of frequency, sequence, and conflict. *Cognitive, Affective, & Behavioral Neuroscience*, 2, 300-317.

Jordano, M. L., & Touron, D. R. (2017). Stereotype threat as a trigger of mind-wandering in older adults. *Psychology and Aging*, *32*(3), 307.

Kam, J. W. Y., & Handy, T. C. (2013). The neurocognitive consequences of the wandering mind: A mechanistic account of sensory-motor decoupling. *Frontiers in Psychology, 4,* 725.

Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeys, I., & Kwapil, T. R. (2007). For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. *Psychological Science*, *18*(7), 614-621.

Kane, M. J., Carruth, N. P., Lurquin, J. H., Silvia, P. J., Smeekens, B. A., von Bastian, C. C., & Miyake, A. (2021). Individual differences in task-unrelated thought in university classrooms. *Memory & Cognition*, 1-20.

Kane, M. J., Gross, G. M., Chun, C. A., Smeekens, B. A., Meier, M. E., Silvia, P. J., & Kwapil, T. R. (2017). For whom the mind wanders, and when, varies across laboratory and daily-life settings. *Psychological Science*, *28*(9), 1271-1289.

Kane, M. J., & McVay, J. C. (2012). What mind wandering reveals about executive-control abilities and failures. *Current Directions in Psychological Science*, *21*(5), 348-354.

Kane, M. J., Meier, M. E., Smeekens, B. A., Gross, G. M., Chun, C. A., Silvia, P. J., & Kwapil, T. R. (2016). Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General*, *145*(8), 1017.

Kane, M. J., Smeekens, B. A., Meier, M. E., Welhaf, M. S., & Phillips, N. E. (2021). Testing the construct validity of competing measurement approaches to probed mind-wandering reports. *Behavior Research Methods*, 1-40.

Kane, M. J., Smeekens, B. A., Von Bastian, C. C., Lurquin, J. H., Carruth, N. P., & Miyake, A. (2017). A combined experimental and individual-differences investigation into mind wandering during a video lecture. *Journal of Experimental Psychology: General*, *146*(11), 1649.

Kawagoe, T. (2022). Executive failure hypothesis explains the trait-level association between motivation and mind wandering. *Scientific reports*, *12*(1), 1-9.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*(6006), 932-932.

Klein, R. J., & Robinson, M. D. (2019). Neuroticism as mental noise: evidence from a continuous tracking task. *Journal of Personality*, *87*(6), 1221-1233.

Kline, R.B. (2011). Principles and practice of structural equation modeling (5th ed., pp. 3-427). New York: The Guilford Press.

Klinger, E. (1971). *Structure and functions of fantasy.* New York: Wiley.

Klinger, E. (1978). Modes of normal conscious flow. In *The stream of consciousness* (pp. 225-258). Springer, Boston, MA

Klinger, E. (2009). Daydreaming and fantasizing: Thought flow and motivation. In K. D. Markman, W. M. P. Klein, & J. A. Suhr (Eds.), *Handbook of imagination and mental simulation* (pp. 225–239). Psychology Press.

Klinger, E., & Cox, W. M. (1987–88). Dimensions of thought flow in everyday life. *Imagination, Cognition and Personality, 7,* 105–128.

Kovacs, K., & Conway, A. R. A. (2016). Process Overlap Theory: A unified account of the general factor of intelligence. *Psychological Inquiry, 27*, 151-177.

Kranzler, J.H. (1992). A test of Larson and Alderton's (1990) worst performance rule of reaction time variability. *Personality & Individual Differences, 13*, 255-261.

Kucyi, A. (2018). Just a thought: How mind-wandering is represented in dynamic brain connectivity. *Neuroimage*, *180*, 505-514.

Laflamme, P., Seli, P., & Smilek, D. (2018). Validating a visual version of the metronome response task. *Behavior Research Methods*, *50*(4), 1503-1514.

Langner R., & Eickoff, S.B. (2013). Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention. *Psychological Bulletin, 139*,870–900.

Larson, G. E., & Alderton, D. L. (1990). Reaction time variability and intelligence: A "worst performance" analysis of individual differences. *Intelligence*, *14*(3), 309-325.

Larson, G.E. & Saccuzzo, D.P. (1989). Cognitive correlates of general intelligence: Toward a process theory of g. *Intelligence, 13*, 5-31.

Leys, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49,* 764-766. http://dx.doi.org/10.1016/j.jesp.2013.03.013

Lim, J., & Dinges, D. F. (2008). Sleep deprivation and vigilant attention. In D. W. Pfaff & B. L. Kieffer (Eds.), *Molecular and biophysical mechanisms of arousal, alertness, and attention* (pp. 305–322). Blackwell Publishing.

Lindquist, S. I., & McLean, J. P. (2011). Daydreaming and its correlates in an educational environment. *Learning and Individual Differences*, *21*(2), 158-167.

Locke, H. S., & Braver, T. S. (2008). Motivational influences on cognitive control: behavior, brain activation, and individual differences. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(1), 99-112.

Löffler, C., Frischkorn, G. T., Rummel, J., Hagemann, D., & Schubert, A. L. (2021). Do attentional lapses account for the worst performance rule?. *Journal of Intelligence*, *10*(1), 2.

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood?. *Multivariate Behavioral Research*, *41*(4), 499-532.

Macdonald, J. S. P., Mathan, S., & Yeung, N. (2011). Trial-by-trial variations in subjective attentional state are reflected in ongoing prestimulus EEG alpha oscillations. *Frontiers in Psychology*, *2*, 82.

Mackworth, N. H. (1950). *Researches on the measurement of human performance* (Medical Research Council Special Report Series No. 268). London, UK: Her Majesty's Stationery Office.

Marcusson-Clavertz, D., Cardeña, E., & Terhune, D. B. (2016). Daydreaming style moderates the relation between working memory and mind wandering: Integrating two hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*, 451-464.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Wandering minds: the default network and stimulus-independent thought. *Science*, *315*(5810), 393-395.

Massar, S. A., Poh, J. H., Lim, J., & Chee, M. W. (2020). Dissociable influences of implicit temporal expectation on attentional performance and mind wandering. *Cognition*, *199*, 104242.

Massidda, D. Retimes: Reaction time analysis (Version 0.1-2). Retrieved from https://CRAN.R-project.org/package=retimes

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review, 16*, 798-817.

McCormack, P. & Wright, N. (1964). The positive skew observed in reaction time. *Canadian Journal of Psychology, 18*, 43-51.

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 196.

McVay, J. C., & Kane, M. J. (2012a). Drifting from slow to “d'oh!”: Working memory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 525.

McVay, J. C., & Kane, M. J. (2012b). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *Journal of Experimental Psychology: General*, *141*(2), 302.

McVay, J. C., & Kane, M. J. (2013). Dispatching the wandering mind? Toward a laboratory method for cuing “spontaneous” off-task thought. *Frontiers in Psychology*, *4*, 570.

McVay, J. C., Kane, M. J., & Kwapil, T. R. (2009). Tracking the train of thought from the laboratory into everyday life: An experience-sampling study of mind wandering across controlled and ecological contexts. *Psychonomic Bulletin & Review*, *16*(5), 857-863.

McVay, J.C., & Kane, M. J. (2010). Does mind wandering reflect executive function or executive failure? Comment on Smallwood and Schooler (2006) and Watkins (2008). *Psychological Bulletin, 136*, 188-197.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*, 108-141.

Megemont, M., McBurney-Lin, J., & Yang, H. (2022). Pupil diameter is not an accurate real-time readout of locus coeruleus activity. *eLife, 11*, article e70510. DOI: https://doi.org/10.7554/eLife.70510

Meiran, N., & Shahar, N. (2018). Working memory involvement in reaction time and intelligence: An examination of individual differences in reaction-time distributions. *Intelligence, 69,* 176-185.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*, 8–14.

Mrazek, M. D., Smallwood, J., Franklin, M. S., Chin, J. M., Baird, B., & Schooler, J. W. (2012). The role of mind-wandering in measurements of general aptitude. *Journal of Experimental Psychology: General*, *141*(4), 788.

Nieuwenhuis, S., Yeung, N., Van Den Wildenberg, W., & Ridderinkhof, K. R. (2003). Electrophysiological correlates of anterior cingulate function in a go/no-go task: effects of response conflict and trial type frequency. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(1), 17-26.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231.

Norman, D. A. (1981). Categorization of action slips. *Psychological Review, 88*(1), 1–15

O'Connell, R. G., Dockree, P. M., Bellgrove, M. A., Turin, A., Ward, S., Foxe, J. J., & Robertson, I. H. (2009). Two types of action error: electrophysiological evidence for separable inhibitory and sustained attention neural mechanisms producing error on go/no-go tasks. *Journal of Cognitive Neuroscience*, *21*(1), 93-104.

Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. In J. Robertson & M. Kaptein (Eds.), Modern statistical methods for HCI (pp. 275–287). Cham, Switzerland: Springer International Publishing.

Parasuraman, R. (1986). Vigilance, monitoring, and search. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance, Vol. 2. Cognitive processes and performance* (pp. 1–39). John Wiley & Sons.

Parasuraman, R., & Davies, D. R. (1977). A taxonomic analysis of vigilance performance. In *vigilance* (pp. 559-574). Springer, Boston, MA.

Parasuraman, R., Warm, J. S., & See, J. E. (1998). Brain systems of vigilance. In R. Parasuraman (Ed.), *The attentive brain*(pp. 221–256). The MIT Press.

Peebles, D. & Bothell, D (2004). Modelling performance in the sustained attention to response task. *Proceedings of the sixth international conference on cognitive modeling*, Carnegie Mellon University/University of Pittsburgh, Pittsburgh, PA, pp. 231-236

Peiris, M. T., Jones, R. D., Davidson, P. R., Carroll, G. J., & Bones, P. J. (2006). Frequent lapses of responsiveness during an extended visuomotor tracking task in non-sleep-deprived subjects. *Journal of Sleep Research*, *15*(3), 291-300.

Pham, T., Lau, Z. J., Chen, S. H. A., & Makowski, D. (2021). Heart rate variability in psychology: A review of HRV indices and an analysis tutorial. *Sensors, 21*, article 3998. https://doi.org/10.3390/s21123998

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling* (No. 157). Sage.

Rajkowski, J., Lu, W., Zhu, Y., Cohen, J. D., & Aston-Jones, G. (2000). Prominent projections from the anterior cingulate cortex to the locus coeruleus (LC) in rhesus monkey. *Society for Neuroscience Abstracts, 26*, 2230.

Ralph, B. C., Onderwater, K., Thomson, D. R., & Smilek, D. (2017). Disrupting monotony while increasing demand: benefits of rest and intervening tasks on vigilance. *Psychological Research*, *81*(2), 432-444.

Rammsayer, T.H., & Troche, S.J. (2016). Validity of the worst performance rule as a function of task complexity and psychometric g: On the crucial role of g saturation. *Journal of Intelligence, 4,* 5.

Randall, J.G., Oswald, F.L., & Beier, M.E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin, 140*, 1411-1431.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, *60*(3), 127-157.

Reason, J. T. (1984). Lapses of attention in everyday life. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 515–549). Orlando, FL: Academic Press.

Reason, J. T. (1990). *Human error.* Cambridge, England: Cambridge University Press.

Reason, J. T., & Lucas, D. (1984). Absent-mindedness in shops: Its incidence, correlates and consequences. *British Journal of Clinical Psychology*, *23*(2), 121-131.

Reason, J. T., & Mycielska, K. (1982). *Absent minded? The psychology of mental lapses and everyday errors.* Englewood Cliffs, NJ: Prentice Hall

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, *35*(6), 747-758.

Robertson, I. H., & O'Connell, R. (2010). Vigilant attention. *Attention and time*, *79*, 88.

Robinson, M. D., & Tamir, M. (2005). Neuroticism as mental noise: a relation between neuroticism and reaction time standard deviations. *Journal of Personality and Social Psychology*, *89*(1), 107.

Robison, M. K., Gath, K. I., & Unsworth, N. (2017). The neurotic wandering mind: An individual differences investigation of neuroticism, mind-wandering, and executive control. *The Quarterly Journal of Experimental Psychology*, *70*(4), 649-663.

Robison, M. K., Miller, A. L., & Unsworth, N. (2019). Examining the effects of probe frequency, response options, and framing within the thought-probe method. *Behavior Research Methods*, *51*(1), 398-408.

Robison, M. K., Miller, A. L., & Unsworth, N. (2020). A multi-faceted approach to understanding individual differences in mind-wandering. *Cognition*, *198*, 104078.

Robison, M. K., & Unsworth, N. (2015). Working memory capacity offers resistance to mind-wandering and external distraction in a context-specific manner. *Applied Cognitive Psychology*, *29*(5), 680-690.

Robison, M. K., & Unsworth, N. (2018). Cognitive and contextual correlates of spontaneous and deliberate mind-wandering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(1), 85.

Robison, M. K., Unsworth, N., & Brewer, G. A. (2021). Examining the effects of goal-setting, feedback, and incentives on sustained attention. *Journal of Experimental Psychology: Human Perception and Performance*, *47*(6), 869.

Rosenberg, J.M., Beymer, P.N., Anderson, D.J., Van Lissa, C.J., & Schmidt, J.A. (2018). tidyLPA: An R Package to Easily Carry Out Latent Profile Analysis (LPA) Using Open-Source or Commercial Software. *Journal of Open Source Software, 3(30)*, 978.

Rosenberg, M. D., Finn, E. S., Constable, R. T., & Chun, M. M. (2015). Predicting moment-to-moment attentional state. *Neuroimage*, *114*, 249-256.

Rosenberg, M., Noonan, S., DeGutis, J., & Esterman, M. (2013). Sustaining visual attention in the face of distraction: a novel gradual-onset continuous performance task. *Attention, Perception, & Psychophysics*, *75*(3), 426-439.

Rummel, J., Hagemann, D., Steindorf, L., & Schubert, A. L. (2021). How consistent is mind wandering across situations and tasks? A latent state–trait analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Sanders, A. F., & Hoogenboom, W. (1970). On the effects of continuous active work on performance. *Acta Psychologica*, *33*, 414-431.

Schmidt-Hansen, M., & Honey, R. C. (2009). Working memory and multidimensional schizotypy: dissociable influences of the different dimensions. *Cognitive Neuropsychology*, *26*(7), 655-670.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*(3), 414.

Schooler, J. W., Reichle, E. D., & Halpern, D. V. (2004). *Zoning out while reading: Evidence for dissociations between experience and metaconsciousness*. MIT press.

Schubert, A. L. (2019). A meta-analysis of the worst performance rule. *Intelligence*, *73*, 88-100.

Schubert, A., Frischkorn, G. T., & Rummel, J. (2020). The validity of the online thought-probing procedure of mind wandering is not threatened by variations of probe rate and probe framing. *Psychological Research*, *84,* 1846-1856.

Schweizer, K., & Moosbrugger, H. (2004). Attention and working memory as predictors of intelligence. *Intelligence, 32,* 329 –347. http://dx.doi.org/ 10.1016/j.intell.2004.06.006

Seli, P., Carriere, J. S., Levene, M., & Smilek, D. (2013). How few and far between? Examining the effects of probe rate on self-reported mind wandering. *Frontiers in psychology*, *4*, 430.

Seli, P., Cheyne, J. A., & Smilek, D. (2013). Wandering minds and wavering rhythms: linking mind wandering and behavioral variability. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 1.

Seli, P., Cheyne, J. A., Xu, M., Purdon, C., & Smilek, D. (2015). Motivation, intentionality, and mind wandering: Implications for assessments of task-unrelated thought. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(5), 1417.

Seli, P., Jonker, T. R., Cheyne, J. A., Cortes, K., & Smilek, D. (2015). Can research participants comment authoritatively on the validity of their self-reports of mind wandering and task engagement?. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 703.

Seli, P., Kane, M. J., Smallwood, J., Schacter, D. L., Maillet, D., Schooler, J. W., & Smilek, D. (2018). Mind-wandering as a natural kind: A family-resemblances view. *Trends in Cognitive Sciences, 22*, 479-490.

Seli, P., Schacter, D. L., Risko, E. F., & Smilek, D. (2019). Increasing participant motivation reduces rates of intentional and unintentional mind wandering. *Psychological Research*, *83*(5), 1057-1069.

Shahar, N., Teodorescu, A.R., Usher, M., Pereg, M., Meiran, N. (2014). Selective influences of working memory load on exceptionally slow reaction times. *Journal of Experimental Psychology: General, 143*, 1837-1860.

Shaw, T., Finomore, V., Warm, J., & Matthews, G. (2012). Effects of regular or irregular event schedules on cerebral hemovelocity during a sustained attention task. *Journal of Clinical and Experimental Neuropsychology*, *34*(1), 57-66.

Sheppard, L.D., Vernon, P.A. (2008). Intelligence and speed of information-processing: A review of 50 years of research. *Personality & Individual Differences, 44*, 535-551.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... & Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337-356.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. Retrieved from https://ssrn.com/abstract=2694998

Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychological Bulletin*, *132*(6), 946.

Smallwood, J., & Schooler, J. W. (2015). The science of mind wandering: empirically navigating the stream of consciousness. *Annual Review of Psychology*, *66*, 487-518.

Smallwood, J., Davies, J. B., Heim, D., Finnigan, F., Sudberry, M., O'Connor, R., & Obonsawin, M. (2004). Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. *Consciousness and Cognition*, *13*(4), 657-690.

Smallwood, J., McSpadden, M., & Schooler, J. W. (2007). The lights are on but no one's home: Meta-awareness and the decoupling of attention when the mind wanders. *Psychonomic Bulletin & Review*, *14*(3), 527-533.

Smallwood, J., McSpadden, M., Luus, B., & Schooler, J. (2008). Segmenting the stream of consciousness: The psychological correlates of temporal structures in the time series data of a continuous performance task. *Brain and Cognition*, *66*(1), 50-56.

Smallwood, J., Nind, L., & O'Connor, R. C. (2009). When is your head at? An exploration of the factors associated with the temporal focus of the wandering mind. *Consciousness and Cognition*, *18*(1), 118-125.

Smilek, D., Carriere, J. S., & Cheyne, J. A. (2010). Failures of sustained attention in life, lab, and brain: ecological validity of the SART. *Neuropsychologia*, *48*(9), 2564-2570.

Smith, A. C., Brosowsky, N. P., Ralph, B. C., Smilek, D., & Seli, P. (2022). Re-examining the effect of motivation on intentional and unintentional task-unrelated thought: Accounting for thought constraint produces novel results. *Psychological research*, *86*(1), 87-97.

Soemer, A., & Schiefele, U. (2019). Text difficulty, topic interest, and mind wandering during reading. *Learning and Instruction*, *61*, 12-22.

Song, X., Wang, X. (2012) Mind Wandering in Chinese Daily Lives – An Experience Sampling Study. *PLOS ONE 7(9)*: e44423.

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*, 103445.

Stawarczyk, D., Majerus, S., Maj, M., Van der Linden, M., & D'Argembeau, A. (2011). Mind-wandering: Phenomenology and function as assessed with a novel experience sampling method. *Acta Psychologica*, *136*(3), 370-381.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2017). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–12.

Steinborn, M. B., Langner, R., Flehmig, H. C., & Huestegge, L. (2016). Everyday life cognitive instability predicts simple reaction time variability: analysis of reaction time distributions and delta plots. *Applied Cognitive Psychology*, *30*(1), 92-102.

Steinmayr, R., Ziegler, M., & Träuble, B. (2010). Do intelligence and sustained attention interact in predicting academic achievement?. *Learning and Individual Differences*, *20*(1), 14-18.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual review of clinical psychology*, *5*, 1-25.

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences*, *110*(16), 6313-6317.

Thomson, D. R., Besner, D., & Smilek, D. (2015). A resource-control account of sustained attention: Evidence from mind-wandering and vigilance paradigms. *Perspectives on Psychological Science*, *10*(1), 82-96.

Tomporowski, P. D., & Tinsley, V. F. (1996). Effects of memory demand and motivation on sustained attention in young and older adults. *The American Journal of Psychology*, *109*, 187-204.

Ulrich, R., & Miller, J. (1993). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General, 123*, 34-80.

Unsworth, N. (2015). Consistency of attentional control as an important cognitive trait: A latent variable analysis. *Intelligence*, *49*, 110-128.

Unsworth, N., & Engle, R. W. (2008). Speed and accuracy of accessing information in working memory: An individual differences investigation of focus switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 616-630.

Unsworth, N., & McMillan, B. D. (2013). Mind wandering and reading comprehension: Examining the roles of working memory capacity, interest, motivation, and topic experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 832–842.

Unsworth, N., & McMillan, B. D. (2014). Similarities and differences between mind-wandering and external distraction: A latent variable analysis of lapses of attention and their relation to cognitive abilities. *Acta Psychologica, 150,* 14–25.

Unsworth, N., & McMillan, B. D. (2017). Attentional disengagements in educational contexts: A diary investigation of everyday mind-wandering and distraction. *Cognitive research: Principles and Implications*, *2*(1), 1-20.

Unsworth, N., & Miller, A. L. (2021). Individual differences in the intensity and consistency of attention. *Current Directions in Psychological Science*, *30*, 391-400.

Unsworth, N., & Robison, M. K. (2015). Individual differences in the allocation of attention to items in working memory: Evidence from pupillometry. *Psychological Bulletin & Review, 22*, 757-765

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective & Behavioral Neuroscience*, *16*, 601–615.

Unsworth, N., & Robison, M. K. (2017b). The importance of arousal for variation in working memory capacity and attention control: A latent variable pupillometry study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1962.

Unsworth, N., & Robison, M. K. (2020). Working memory capacity and sustained attention: A cognitive-energetic perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(1), 77.

Unsworth, N., & Robison, M.K. (2017a). A Locus Coeruleus-Norepinephrine account of individual differences in working memory capacity and attention control. *Psychonomic Bulletin & Review, 24*, 1282-1311.

Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory & Language, 62,* 392-406.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2012). Variation in cognitive failures: An individual differences investigation of everyday attention and memory failures. *Journal of Memory and Language, 67,* 1–16.

Unsworth, N., McMillan, B. D., Brewer, G. A., & Spillers, G. J. (2012). Everyday attention failures: An individual differences investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1765.

Unsworth, N., Miller, A. L., & Aghel, S. (2022). Effort mobilization and lapses of sustained attention. *Cognitive, Affective, & Behavioral Neuroscience, 22*, 42-56.

Unsworth, N., Redick, T. S., Lakey, C. E., & Young, D. L. (2010). Lapses in sustained attention and their relation to executive and fluid abilities: An individual differences investigation. *Intelligence, 38,* 111–122. http://dx.doi.org/10.1016/j.intell.2009.08.002

Unsworth, N., Redick, T. S., Spillers, G. J., & Brewer, G. A. (2012). Variation in working memory capacity and cognitive control: Goal maintenance and microadjustments of control. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 65,* 326– 355. http://dx.doi.org/10.1080/17470218.2011.597865

Unsworth, N., Robison, M. K., & Miller, A. L. (2018). Pupillary correlates of fluctuations in sustained attention. *Journal of Cognitive Neuroscience*, *30*(9), 1241-1253.

Unsworth, N., Robison, M. K., & Miller, A. L. (2021). Individual differences in lapses of attention: A latent variable analysis. *Journal of Experimental Psychology: General*.

Unsworth, N., Spillers, G. J., Brewer, G. A., & McMillan, B. (2011). Attention control and the antisaccade task: A response time distribution analysis. *Acta Psychologica, 137*, 90-100.

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston- Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science, 283,* 549–554.

Van Den Brink, R. L., Murphy, P. R., & Nieuwenhuis, S. (2016). Pupil diameter tracks lapses of attention. *PLoS One*, *11*(10), e0165274.

Van Zandt, T.V. How to fit a response time distribution. Psychol. Bull. & Rev. 2000, 7, 424-465

Wagenmakers, E. J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological Review*, *114*(3), 830.

Wammes, J. D., Seli, P., Cheyne, J. A., Boucher, P. O., & Smilek, D. (2016). Mind wandering during lectures II: Relation to academic performance. *Scholarship of Teaching and Learning in Psychology*, *2*(1), 33.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, *50*(3), 433-441.

Weinstein, Y. (2018). Mind-wandering, how do I measure thee with probes? Let me count the ways. *Behavior research methods*, *50*(2), 642-661.

Weinstein, Y., De Lima, H. J., & Van Der Zee, T. (2018). Are you mind-wandering, or is your mind on task? The effect of probe framing on mind-wandering reports. *Psychonomic Bulletin & Review*, *25*(2), 754-760.

Weissman, D. H., Roberts, K. C., & Woldroff, M. G. (2006). The neural basis of momentary lapses in attention. Nature Neuroscience, 9, 971–978.

Welhaf, M. S., & Kane, M. J. (2022a). A Nomothetic Span Approach to the Construct Validity of Sustained Attention Measurement: Re-Analyzing Two Latent-Variable Studies of Performance Variability and Mind-Wandering Self-Reports. https://psyarxiv.com/4qk56/

Welhaf, M.S., & Kane, M. J. (2022b). A Combined Experimental–Correlational Approach to the Construct Validity of Performance-Based and Self-Report-Based Measures of Sustained Attention. https://psyarxiv.com/znbp7/

Welhaf, M. S., Meier, M. E., Smeekens, B. A., Silvia, P. J., Kwapil, T. R., & Kane, M. J. (2022). A “Goldilocks Zone” for mind wandering reports? A secondary data analysis of how few thought probes are enough for reliable and valid measurement. *Behavior Research Methods*.

Welhaf, M. S., Smeekens, B. A., Gazzia, N. C., Perkins, J. B., Silvia, P. J., Meier, M. E., ... & Kane, M. J. (2020a). An exploratory analysis of individual differences in mind wandering content and consistency. *Psychology of Consciousness: Theory, Research, and Practice*, *7*(2), 103-125.

Welhaf, M. S., Smeekens, B. A., Meier, M. E., Silvia, P. J., Kwapil, T. R., & Kane, M. J. (2020b). The worst performance rule, or the not-best performance rule? Latent-variable analyses of working memory capacity, mind-wandering propensity, and reaction time. *Journal of Intelligence*, *8*(2), 25.

Wessel, J. R. (2018). Prepotent motor activity and inhibitory control demands in different variants of the go/no-go paradigm. *Psychophysiology*, *55*(3), e12871.

Wickham, H., Francois, R., Henry, L., Müller, K. Dplyr: A grammar of data manipulation. Retrieved from https://CRAN.R-project.org/package=dplyr

Wiemers, E. A., & Redick, T. S. (2019). The influence of thought probes on performance: Does the mind wander more if you ask it? *Psychological Bulletin & Review, 26*, 367-373.

Wilhelm, O., & Oberauer, K. (2006). Why are reasoning ability and working memory capacity related to mental speed? An investigation of stimulus-response compatibility in choice reaction time tasks. *European Journal of Cognitive Psychology, 18*, 18-50.

Wilson, K. M., Finkbeiner, K. M., De Joux, N. R., Russell, P. N., & Helton, W. S. (2016). Go-stimuli proportion influences response strategy in a sustained attention to response task. *Experimental Brain Research*, *234*(10), 2989-2998.

Wilson, K. M., Head, J., De Joux, N. R., Finkbeiner, K. M., & Helton, W. S. (2015). Friendly fire and the sustained attention to response task. *Human Factors*, *57*(7), 1219-1234.

Yamashita, A., Rothlein, D., Kucyi, A., Valera, E. M., Germine, L., Wilmer, J., ... & Esterman, M. (2021). Variable rather than extreme slow reaction times distinguish brain states during sustained attention. *Scientific Reports, 11(1)*, 1-13.

Yanko, M. R., & Spalek, T. M. (2014). Driving with the wandering mind: The effect that mind-wandering has on driving performance. *Human factors*, *56*(2), 260-269.

Young, M. E., Sutherland, S. C., & McCoy, A. W. (2018). Optimal go/no-go ratios to maximize false alarms. *Behavior Research Methods*, *50*(3), 1020-1029.

Zanesco, A. P. (2020). Quantifying streams of thought during cognitive task performance using sequence analysis. *Behavior Research Methods*, *52*(6), 2417-2