Journal of Mathematical Psychology

Assessing cross-modal interference in the detection response task

--Manuscript Draft--

Manuscript Number:	YJMPS-D-19-00078R2
Article Type:	VSI: Multisensory modelin:Research Paper
Keywords:	Cross-modal interference; detection response task; systems factorial technology; workload capacity; multi-tasking; Cognitive Modeling
Corresponding Author:	Alexander Thorpe University of Newcastle Callaghan, NSW AUSTRALIA
First Author:	Alexander Thorpe
Order of Authors:	Alexander Thorpe
	Reilly Innes
	James Townsend
	Rachel Heath
	Keith Nesbitt
	Ami Eidels
Abstract:	The detection response task (DRT) is a measure of workload that can assess the cognitive demands of real-world multitasking. It can be configured to present simple stimuli of several modalities, including auditory and visual signals. However, the concurrent presentation of the DRT stimuli alongside another task could cause dual-task interference, and the extent of this interference could be different based on the DRT's configuration. It is necessary to consider the characteristics of the DRT stimulus, such as modality, to identify a minimally intrusive stimulus. Fifty participants completed a computer-based one-dimensional tracking task alongside a DRT. The DRT's stimuli varied in their modality (visual/auditory), while the tracking task varied in its workload demand (low/high). DRT performance was modelled using a shifted-Wald model, while the tracking task was assessed using systems factorial technology (SFT), a non-parametric methodology capable of capturing a cognitive system's workload capacity. To allow the latter's use, we developed a method of transforming continuous tracking data into a discrete form akin to response times. Analysis of DRT data found little evidence that the DRT's modality affected processing efficiency, while SFT analysis found limited-capacity processing on the tracking task across both DRT modalities. These findings suggest DRT modality had little effect on the level of interference between the two tasks.
Suggested Reviewers:	Joseph Houpt joseph.houpt@wright.edu
	Daniel Little daniel.little@unimelb.edu.au
	Cheng-Ta Yang yangct@mail.ncku.edu.tw
Response to Reviewers:	

Dear Dr Diederich, Editor in Chief, *Journal of Mathematical Psychology*
Dr Colonius, Action Editor and Guest Editor of the special issue

Thank you for considering our revised manuscript "Assessing cross-modal interference in the detection response task", for publication in the special issue of the *Journal of Mathematical Psychology* on computational and mathematical approaches to multisensory integration. We thank the Reviewers again for their thoughtful comments and provide our detailed response below. Edits to latest round of reviews are marked YELLOW in the revised manuscript. For completeness, we kept changes from the previous round marked as well (GREEN). We believe the manuscript is now ready for publication in the *Journal of Mathematical Psychology* and hope you agree.

Sincerely,
Alex Thorpe,
The University of Newcastle

Response to Editor and Reviewer comments:

Reviewer #1: One issue: The current manuscript could benefit from integrating the relevant recent research on the topic of workload capacity and the role of distractors (Little, Eidels, Fifić, M., & Wang, 2015; Little, Eidels, Fifić, & Wang, 2018) that are highly relevant for the theoretical considerations on the page 11, second to the last paragraph).

>>> We thank the reviewer for this suggestion. Information regarding workload capacity in the presence of distractors and the resilience function has been added to the Introduction section (p. 10).

The authors responded to all of my suggestions in a satisfactory manner. Overall, I think that the manuscript describes an interesting research project that could further improve our understanding about the perceptual mechanism underlying the detection response task (DRT and the system's capacity in the cognitive operations. There is a notable integrative effort to combine previously separated approaches, such as the measures of workload capacity (as provided by the system's factorial technology) and the classical detection task, thus increasing the validity of the findings. These crossovers between different technologies can be expected to have a facilitatory effect of the future work on this matter.

>>> Thank you.

Reviewer #2: I cannot support publication of this study. I will not go into details again (I already did this in my first review), I will just list a number of aspects that illustrate the overall impression I got from the study.

>>> We believe we have addressed the Reviewer's concerns and implemented her/his suggestions in the last round. For the remaining points, please see our response below.

1. "Mean response times (RT) and hit rates (HR) were compared using repeated-measures ANOVA and Bayesian repeated-measures ANOVA. Bayesian tests were applied only in cases of non-significant or ambiguous results, to present evidence in favour of no difference; this cannot be achieved using frequentist tests alone." (p. 19)

If the F-test is not significant, the authors look further what the Bayesian test says. This two-stage procedure is statistically flawed. In statistics, "the more the merrier" does not hold. The analysis should follow from the hypotheses.

>>> There are many arguments in favour of Bayesian statistical analysis (articulated clearly in work by EJ Wagenmakers, Richard Morrey, Jeff Rouder, John Kruschke, Dani Navarro, and others), and in theory we could have reported only Bayesian tests. Yet despite its appeal, Bayes Factors have not yet become the 'industry standard' in reporting psychological results, and it is therefore not quite clear whether it is sufficient to report Bayes Factors alone, while the majority of students worldwide are still taught to interpret frequentists tests. For this reason, we see merit in reporting frequentist tests at baseline, and supplementing them with Bayes Factors when we are seeking evidence in favour of the null hypothesis -- which Freq' tests cannot provide. This is the approach taken in the present manuscript. We could have discarded all frequentists tests and reported only Bayes Factors, but we feel this change would be too radical and would not serve the best interest of the JMP readership. The approach we have taken is quite different from, say, running two sets of analyses

and choosing to report the test-outcome that best supports our hypothesis, which we whole-heartedly agree would have been a flawed process.

2. Pseudo-RTs for the tracking data: In my first review I have demonstrated that this measure is flawed and I have shown that there is no monotonical relation to performance. Instead of a convincing counterargument, the authors kept the analysis of this flawed measure and "fixed" the problem by running a few simulations and sensitivity analyses and copying my comment into the paper.

We thank Reviewer 2 for reiterating their concern re- pseudo-RTs for the tracking data. In the lack of more concrete text in this round of revisions we went back to her/his original review, and will assume the relevant concern is this:

"*Regarding the pseudo-RT transformation: If RT is used as a performance measure, better performance should be associated with lower RTs. Consider a very good participant that shows perfect tracking behavior. The needle of this participant would never leave the yellow area, and you would end up with no data. Now consider a participant that randomly moves the needle up and down very fast. This participant would have very fast responses, although "performance" is actually very poor. The first example already illustrate that the pseudo-RT-transformation does not work. The second example may not be applicable to the present experiment because the speed of the needle is probably kept constant by the program, but the "idea" of a tracking task (i.e., with a proper response device) would actually allow different speed.*

>>> The same point was indeed raised by the Reviewer in the first round. We already agreed in our first revision that the more lenient the threshold, the fewer 'pseudo trials' we record, and acknowledged the issue in the text. We undertook an analysis exercise to assess how much of an impact this issue might have on the outcome of capacity analyses. Appendix B (p. 48) and Figure 14 show relatively stable capacity estimates across different data-set sizes. But, that is not the main point with respect to the Reviewer's current concern. The main point, rather, is that capacity estimates were stable irrespective of the tracking error *tolerance-level*. That is, we repeated the analysis of the same tracking-error data reported in Results, for different levels of error-tolerance (epsilon). And, when doing so, the overall pattern of results (namely, limited capacity) changed very little, as can be seen in Figure 14.

Now, consider a very 'good' subject, as in the example provided by the Reviewer. This subject can track the moving target accurately and stay close to the central needle. Yet, even a very good subject cannot track the needle perfectly across the entire course of the trial (this is not an assumption, but rather an empirical fact). If we 'zoom in' the analysis, and minimize the error tolerance level, then at some points in time we will discover deviations of the tracking marker from the target location. If the subject is as good as suggested, they should be able to quickly get closer to the target and will manifest speedy pseudo response times, as expected. But the more interesting test-case is this: What if a very good subject can follow the target very closely, but never quite be on top of it? In that case, with a relatively large epsilon we might record very long response times for a subject who is actually performing quite well, theoretically confounding the analysis. This is where the analysis reported in Figure 14 becomes useful: if that were the case, then systemically varying the tolerance level should, at some critical point, be affected by the constant yet small error. As soon as the error-tolerance index, epsilon, gets smaller than the subject's error, these deviations should manifest as errors, captured as frequent and quick pseudo response times, and affect the capacity analysis. Our dedicated analysis, however, shows this does not happen (see Figure 14).

We suspect, based on the Reviewer's comment that this explanation did not quite come through in the original text. We apologise and have now improved the clarity of exposition [p. 35]. There could be other possible solutions for this issue. For example, one could normalise the amount of time spent out of the error-tolerance range by overall task duration (so that good subjects score low, as they spend little normalised time out of the 'permitted' range, and poor subjects score high). Or, for another example, our own team member, Rachel Heath, is exploring methods that subject tracking error to sophisticated fractal analysis based on her non-linear dynamics approach. But, these and other techniques offer completely different analyses, and importantly, do not produce the desired pseudo-response times that allow us to connect tracking data with response-time analysis tools.

3. Summary of the main findings: Participants were slower in condition A than in condition B. The modality of the distractor did not matter. Instead of honestly reporting this modest finding, the reader is flooded with poorly motivated models (Wald) and extra analyses of capacity coefficients that are nothing more than log transforms of RT distributions. Since the log transform is a monotonical transformation, we have K_A(t) < K_B(t) whenever we have the same relation for the survivors S_A(t) < S_B(t) and vice-versa, and from S_A(t) < S_B(t) follows the same relation for mean RT. As a side note, the many numerical problems for t = 0 still exist in Figure 11, 14, 17, 18. In my last review I illustrated the problem (basically, log S(t) = -inf for t -> 0).

The Reviewer pointed out the numerical problems associated with computing a ratio of log transform (e.g., integrated hazard functions) in the last round of revisions. In the Reviewer's own words "*C(t) suffers from an in-built design flaw because it is defined as the ratio (!) of two logarithms C(t) := log S_AB(t) / [log S_A(t) + log S_B(t)] instead of the difference, C*(t) := log S_AB(t) MINUS [log S_A(t) + S_B(t)], which would be much more natural to use with logs anyway… I tend to consider this as a minimum requirement for a revision. In fact, the authors noted this problem themselves at a later page.*"

>>> In response, we modified in R1 (the previous revision) the entire analysis to a difference form, as suggested. This change was also implemented in Figures 10, 13, 16, and 17. We believe this response should have satisfied the reviewer's concerns, and in fact find it a bit unfair to bring it up again after the requested changes to the analysis had been made. As a side note, any numerical problem for t=0 might be important mathematically, but practically inconsequential for behavioural data; the common assumption is that humans simply cannot respond so quickly and such responses are typically censored anyway.

4. Cohen's d, eta-square, z-standardization of capacity coefficients: The inclusion of dimensionless (and therefore meaningless) effect measures did not in any way improve the paper. In fact, response time has a nice SI unit, milliseconds, which is easily communicated. Translating meaningful measures to standard deviations of an arbitrarily chosen experimental condition does not add any information. Did anyone ever enter a grocery store to buy half a standard deviation of milk?

>>> We have included eta-square values at the specific request of Reviewer 1. As for Cz, the measure developed by Houpt and Townsend in 2012 was added to aid inference. By now Cz is considered common practice in SFT analysis (e.g., cited in Google Scholar more than 120 times, 24/4/2020). Its main contribution is supplementing inference from capacity analysis, which used to be based only on visual inspection of the capacity coefficient plots. It appears to now be accepted as a legitimate and useful statistical tool.

5. Figure 1 presents a linear ballistic accumulator, whereas a diffusion model is used in the text. I am again wondering why none of the 6 coauthors noticed such inconsistencies.

>>> We thank the reviewer for noting the model presented in Figure 1 is an LBA and not the Wald model, yet at the same time note the figure is said to provide a general illustration of “the process of making a decision as described by a sequential sampling model” [SIC; and we don’t mean it here as the acronym for Survivor Interaction Contrast]. Due to its simplicity the LBA provides a convenient platform for illustrating Evidence Accumulation and the associated parameters. This was not an inconsistency, but rather a deliberate choice to enhance the clarity of the exposition. It may work for some but perhaps not for all, and we understand the Reviewer’s concern. To avoid confusion, we removed this figure from the manuscript, as all relevant information it depicted is described in-text.

6. The purpose of the study is not clear. There's no real theory other than replication of Wicken's findings, and the methodological contribution is questionable (see 2 and 3). In some sense, the authors acknowledge this in the last sentence, "Our findings also present novel methods for applying cognitive models to more real-world data sets in the future."

>>> That is a fair point, and we now have improved the exposition of the goals in the first page of the Introduction to ensure the purpose is stated clearly and early.

Highlights

- Detection response task (DRT) is a measure of workload processing capacity
- To assess the extent to which DRT interferes with other tasks's performance or processing efficiency based on DRT stimulus modality, DRT was presented alongside a computer-based tracking task
- DRT presence interfered with tracking task performance, but modality and salience of DRT stimulus had no differential interference effect
- Processing efficiency was assessed using shifted-Wald model for DRT data, and systems factorial technology (SFT) for primary task.
- Continuous tracking task data was transformed to a discrete form to allow SFT analysis of data from a continuous task
- High tracking task load led to lower processing efficiency in both tasks

Abstract

Assessing Cross-Modal Interference in the Detection Response Task

Alexander Thorpe[1], Reilly Innes, James Townsend, Rachel Heath, Keith Nesbitt, and Ami Eidels

[1]University of Newcastle, Australia
alexander.thorpe@uon.edu.au

The detection response task (DRT) is a measure of workload that can assess the cognitive demands of real-world multitasking. It can be configured to present simple stimuli of several modalities, including auditory and visual signals. However, the concurrent presentation of the DRT stimuli alongside another task could cause dual-task interference, and the extent of this interference could be different based on the DRTs configuration. It is necessary to consider the characteristics of the DRT stimulus, such as modality, to identify a minimally intrusive stimulus. Fifty participants completed a computer-based one-dimensional tracking task alongside a DRT. The DRTs stimuli varied in their modality (visual/auditory), while the tracking task varied in its workload demand (low/high). DRT performance was modelled using a shifted-Wald model, while the tracking task was assessed using systems factorial technology (SFT), a non-parametric methodology capable of capturing a cognitive systems workload capacity. To allow the latter's use, we developed a method of transforming continuous tracking data into a discrete form akin to response times. Analysis of DRT data found little evidence that the DRT's modality affected processing efficiency, while SFT analysis found limited-capacity processing on the tracking task across both DRT modalities. These findings suggest DRT modality had little effect on the level of interference between the two tasks.

Assessing Cross-Modal Interference in the Detection Response Task

Alexander Thorpe[1], Reilly Innes[1], James Townsend[2], Rachel Heath[1], Keith Nesbitt[1], and Ami Eidels[1]

[1]University of Newcastle

[2]Indiana University

Author Note

This research was supported by an Australian Research Council grant DP160102360 to AE and JTT, and by an Australian Government Research Training Program (RTP) Scholarship awarded to Alexander Thorpe. Correspondence concerning this article should be addressed to Alexander Thorpe, School of Electrical Engineering and Computing, University of Newcastle, Callaghan NSW 2308, Australia; Email: alexander.thorpe@newcastle.edu.au

Abstract

The detection response task (DRT) is a measure of workload that can assess the cognitive demands of real-world multitasking. It can be configured to present simple stimuli of several modalities, including auditory and visual signals. However, the concurrent presentation of the DRT stimuli alongside another task could cause dual-task interference, and the extent of this interference could be different based on the DRT's configuration. It is necessary to consider the characteristics of the DRT stimulus, such as modality, to identify a minimally intrusive stimulus. Fifty participants completed a computer-based one-dimensional tracking task alongside a DRT. The DRT's stimuli varied in their modality (visual/auditory), while the tracking task varied in its workload demand (low/high). DRT performance was modelled using a shifted-Wald model, while the tracking task was assessed using systems factorial technology (SFT), a non-parametric methodology capable of capturing a cognitive system's workload capacity. To allow the latter's use, we developed a method of transforming continuous tracking data into a discrete form akin to response times. Analysis of DRT data found little evidence that the DRT's modality affected processing efficiency, while SFT analysis found limited-capacity processing on the tracking task across both DRT modalities. These findings suggest DRT modality had little effect on the level of interference between the two tasks.

Assessing Cross-Modal Interference in the Detection Response Task

When driving a car, we all know we should keep our hands on the wheel and off our mobile devices and other distractions. These distractions place extra *workload* on the user, drawing resources away from the more important task of driving safely. But are all distractions made equal? Texting and driving is an obvious risk, as both tasks demand the focus of the driver's eyes, but the modern car also places demands on the driver's ears, with such features as voice assistant technology. How does multi-tasking across modalities differ from multi-tasking when both tasks appeal to the same sense? The current study aims to use mathematical models of participants' responses to assess how measures of workload may interfere with user performance, and whether a multi-modal paradigm limits such interference. This aim is complicated by the trial-by-trial structure of most psychological experiments, which does not reflect the continuous nature of real-world tasks such as driving. To this end, our second aim was to develop and test a new technique for transforming data from a continuous task into a form that allows the use of these mathematical models.

Background

Multi-tasking has a deleterious effect on performance due to people's limited *capacity* for processing information (Kahneman, 1973). Tasks that require more resources interfere to a greater extent than others, while *automated* tasks require no resources, and therefore have no impact on other tasks (Wickens, 2002). However, a unitary model of processing capacity where all tasks draw on a single pool of resources cannot explain findings relating to *multi-modal* multi-tasking— situations in which tasks are presented using different sensory modalities (Wickens, 2002).

Treisman and Davies (1973) found that dual-task performance was faster when the two tasks were presented through different modalities than when they were presented through the same modality. They theorised that perceptual capacity may operate at multiple levels—different senses process inputs from different modalities by *peripheral* systems, with a single *central* system processing the information flowing from the periphery. Navon and Gopher (1979) explained the differential effect of multi-modal

multi-tasking in terms of multiple resources, whereby separate mechanisms within the perceptual and processing systems draw on separate resources. This explanation predicts that a pair of tasks that appeal to the same perceptual system will draw resources from the same pool, where tasks appealing to separate senses will not.

Wickens (1980), in developing Multiple Resource Theory, identified structural patterns to dual-task interference within and between modalities. This was manifest in a four-dimensional model of multiple resources; peripheral systems are responsible for perceiving and processing stimuli for each modality, therefore the resources used for this processing are not shared across modalities, whereas responding to stimuli is handled by a *central* system with a single resource pool. Multi-modal stimulus perception is theorised under this model to draw on separate resources from each stimulus modality at a peripheral level, similarly to Treisman and Davies' model, as an explanation for the relatively smaller impact of multi-tasking across modalities (Wickens, 2002). Under this model, dual-task interference occurs primarily at a central processing level (Bonnel & Hafter, 1998), as a result of the limited availability of working memory (Wickens, 2002).

Multi-tasking performance may also be limited not by participants' ability to perceive stimuli, but by their ability to ignore them. Lavie's (1995) Perceptual Load Theory proposed a perceptual mechanism with limited capacity which is responsible for perceiving all stimuli presented. As this mechanism has a limited capacity, it may be overloaded under conditions of high *perceptual load*, or situations where high amounts of information are presented simultaneously, with new stimuli being missed. This theory was expanded to include a *cognitive control* mechanism which is responsible for selective attention, whereby attention is directed to task-relevant stimuli, and restricted to exclude irrelevant stimuli (Lavie, Hirst, de Fockert & Viding, 2004). Overloading this mechanism causes selective attention to break down, and perceptual information to be processed regardless of its relevance to the task at hand. In multi-tasking situations with high cognitive and perceptual load, it is theorised that selective attention will fail, causing interference between tasks due to irrelevant information being processed, followed by a state of inattentional blindness, where new information is not processed at

all. In the latter state, distractor interference actually decreases, due to the distracting stimuli not being perceived.

Theories of perceptual load and cognitive control have been applied to cross-modal multi-tasking. Multi-sensory integration is negatively affected by increased perceptual load with performance on both a complex speech integration task and a simpler multi-sensory detection task decreasing as perceptual load increased (Gibney et al., 2017). Additionally, increased visual perceptual load in an air traffic control task had a negative effect on auditory signal detection (Causse, Imbert, Giraudet, Jouffrais & Tremblay, 2016), suggesting cross-modal task interference. There may therefore be two components to task interference in multi-tasking—a perceptual component, whereby two tasks that share a modality will cause greater interference than two tasks that do not, and a processing component, whereby the participant's limited capacity to process and respond to multiple information sources is overloaded, regardless of the modality of that information.

Detection Response Task

The study of user workload, and the consequences of overloading a user's limited capacity, are of relevance to the domain of human machine interaction. Depending on the manner and time that information is delivered, users may be hindered rather than helped by information delivery (Haapalainen, Kim, Forlizzi & Dey, 2010). Indeed, some systems designed to help users, such as hands-free phone interfaces in cars, may endanger drivers and other road users by placing unnecessary workload on drivers (Strayer, Cooper, Turrill, Coleman & Hopman, 2017).

The detection response task (DRT) is an objective measure of workload designed for use in real-world settings such as driving (International Organization for Standardization, 2016). It is a simple signal detection task that measures *residual capacity*, or those resources that are not occupied by other tasks at a given time. This can be considered a subset of a person's total capacity (see Table 1 for a glossary of theoretical and methodological terms). The demands of the task do not change across

trials, therefore changes in performance on the DRT, defined as mean response time (RT), hit rate, or lapse rate, indicate changes in the user's workload state across different conditions of a primary task—slower responses and lower accuracy indicate less residual capacity, and hence higher workload, while faster responses and higher accuracy indicate more residual capacity is available. Its low-profile apparatus allows it to be used in situations such as driving tasks, where other hardware would not be practical. Software-based versions of the DRT have also been used, whereby the methodological specifications of the ISO standard are applied to a computer-based task (Thorpe, Nesbitt & Eidels, 2019). In an automotive setting, it has been used to assess the cognitive impact of mobile phone use (Strayer et al., 2015), and in-car conversations (Tillman, Strayer, Eidels & Heathcote, 2017), and has been used in lab as well as in highway driving tasks (Hsieh, Seaman & Young, 2015). The DRT has also been employed in aviation research, to examine the effect of added information in helicopter flight (Innes, Howard, Eidels & Brown, 2018).

Table 1

Glossary of terms used throughout this text.

Capacity coefficient	An estimate of a process's *workload capacity*, calculated by comparing hazard rates of responses across different levels of task demands.
Cross-correlation	A method of correlating two time series by time-shifting one time series.
Detection response task	A simple signal detection task designed to index a participant's workload state by measuring *residual capacity*. Abbreviated as DRT.
Drift rate	A parameter of the shifted-Wald model, denoted as v, representing the rate of evidence accumulation.
Modality	The sensory system to which a stimulus is presented.
Non-decision time	A parameter of the shifted-Wald model, denoted t_0, representing any time spent not accumulating evidence for a decision.
Pseudo-RT	A temporally sensitive variable akin to RT that reflects time taken on a task.
Residual capacity	Any cognitive resources not currently dedicated to other cognitive processes.
Shifted-Wald model	A single-bound evidence accumulation model that can estimate the drift-rate, threshold and non-decision time of a single-choice RT task.
Systems factorial technology	A non-parametric methodology that can estimate the *workload capacity* of a cognitive process by estimating *capacity coefficients*. Abbreviated as SFT.
Threshold	A parameter of the shifted-Wald model, denoted as a, representing how much evidence is required for a decision-making process to complete.
Workload capacity	The efficiency of a process under increased task demands.

Sequential Sampling Models

The assessment of DRT performance need not be limited to comparisons of mean RT. Sequential sampling models have been frequently used as a tool to characterize individuals' response processes, whereby to make a decision, one accumulates evidence towards a response threshold (Laming, 1968). The effect of manipulating variables is not measured by surface-level behaviour such as changes in RT, but rather by changes in the latent *parameters* that sequential sampling models estimate, such as the value of the response threshold and the rate at which the response threshold is approached.

Typically, sequential sampling models investigate choice RTs in terms of evidence accumulation rate, or the speed with which information is processed, referred to hereafter as 'v', which is in turn composed of mean drift rate and variance around that mean; response threshold, or the amount of evidence required to make a decision, referred to hereafter as 'a'; starting point, which represents how biased towards a choice one is at the start of the decision-making process; and non-decision time, generally defined as the time spent encoding and responding to the stimulus, referred to hereafter as 't_0' (Brown & Heathcote, 2008; Leite & Ratcliff, 2010). Variation in each parameter, or a combination of parameters, could underpin changes in RT, but the rate of evidence accumulation and response threshold are the parameters most commonly studied in recent literature related to the DRT.

Sequential sampling models have been applied to DRT response time data from automotive studies to assess the effect of in-car conversation while driving on the driver's workload state, (Tillman et al., 2017), and to assess the effect of performing a concurrent working memory task while driving (Castro, Strayer, Matzke & Heathcote, 2019). Higher cognitive load has been associated with higher response threshold, and hence higher caution (Castro et al., 2019; Tillman et al., 2017), with the effect increasing with time pressure (Palada, Neal, Strayer, Ballard & Heathcote, 2019). However, there have been mixed results with regard to drift rate. Drift rate represents the rate of evidence accumulation, and Schmiedek, et al. (2007) found that mean drift rate on a choice-RT task was positively correlated with working memory capacity. It is

therefore an intuitive assumption that drift rate corresponds closely to the amount of cognitive resources available. However, Tillman et al. (2017) did not find lower drift rate in DRT responses as cognitive load increased. Slower DRT RTs could therefore be attributed to increased caution, rather than slower information processing as a result of sharing a limited resource such as working memory (Heathcote et al., 2015). Given this alternative explanation, as well as the possibility this or other findings are statistical outliers, it is therefore unclear which model parameters or combinations of parameters relate to changing levels of cognitive load and further research is required. More generally, the relationship between model parameters and experimental manipulations may not be straightforward. In the case of choice RT experiments, it has been found that models may be so flexible as to generate any pattern of empirical data, raising the issue of model mimicry, whereby a process can be equally explained by multiple disparate models (Jones & Dzhafarov, 2014).

A challenge to modelling the DRT is the simple nature of the task itself. Sequential sampling models such as the drift diffusion model (Ratcliff, 1978) and the linear ballistic accumulator (Brown & Heathcote, 2008) were devised to assess data from two-alternative forced choice experiments. Such experiments produce RT and accuracy data for each of the possible choices participants could make. The DRT, meanwhile, is a simple detection task. Without an "incorrect" response available, accuracy can only be considered in terms of hits, late misses, and false alarms. This requires any potential sampling model to be simpler than traditional models. A single-choice version of the drift diffusion model and has been developed (Ratcliff & Van Dongen, 2011; Schwarz, 1989) and this model has been applied to DRT RTs (Ratcliff & Strayer, 2014). The Wald model (Heathcote, 2004) has also been used to model DRT responses as it is suitable for single-choice tasks (Castro et al., 2019; Palada et al., 2019; Tillman et al., 2017). The Wald distribution, or Inverse Gaussian distribution, describes the passage of time for Brownian motion with a positive drift rate towards a positive threshold. The *shifted-Wald* model differs from the Wald model, as it uses t_0 to describe the shifted response time distribution, so that the model still

maintains a threshold and rate, yet accounts for non-decision time by shifting the response time distribution. In contrast, the Wald model cannot estimate t_0 without adding an independent process. The single-bound drift diffusion model is not able to differentiate between the effects of drift rate and threshold or estimate non-decision time under certain restrictions (Ratcliff & Strayer, 2014), and Palada et al. 2019 found that the linear ballistic accumulator did not perform as well in fitting single-choice data as the Wald model. Given the limitations of the LBA and single-drift diffusion models in prior research, the shifted-Wald model therefore appears to be the most appropriate of the three models to apply to DRT data.

While the DRT presents information about the user's workload state, it does not examine the demands of the primary task itself. Given that the DRT only indexes residual capacity rather than the intrinsic workload demands of the primary task, its use may not be sufficient to fully understand the participants' cognitive state. It is therefore important to consider the properties of the cognitive processes used to complete the primary task in a dual-task paradigm. Sequential sampling models have been applied to primary tasks in such paradigms (Palada et al., 2019), but this approach can only be undertaken when the data from the primary task is discrete, in the form of RT distributions. Additionally, sequential sampling models require properties of the RT distributions and their underlying parameters to be estimated or assumed. One of these issues can be addressed by applying a *non-parametric* model to data, which requires no such assumptions. One method for achieving this is through the application of *systems factorial technology*.

Systems Factorial Technology

Systems factorial technology (SFT) is a powerful mathematical methodology that allows the properties of cognitive processes to be examined (Townsend & Nozawa, 1995). Through SFT, empirical data can be compared to benchmark models with given properties, with the best-matching model giving the best account for the data. As mentioned above, SFT is non-parametric, and is therefore agnostic to the distributional

properties of the data examined (Houpt, Blaha, McIntire, Havig & Townsend, 2014), lending the methodology a great deal of versatility. Central to the application of SFT to experimental data is the *redundant target paradigm*, an experimental design in which participants may be presented with a single stimulus to be processed, or a pair of stimuli, either of which may be processed to complete the task. To take the example of a simple signal detection task, a single stimulus such as a light or tone would constitute a single-target trial, whereas two stimuli conveying the same information would constitute a double-target trial. By manipulating the presence and salience of these two targets, the properties of the process in question can be estimated (Townsend & Nozawa, 1995).

Of particular interest in the current study is the property of *workload capacity*, or the efficiency of a process under increased task demands (Townsend & Eidels, 2011). As depicted in Figure 2, a process with *limited capacity* reduces in efficiency as load increases, whereas an *unlimited-capacity* process retains its efficiency across levels of task load. A process is considered *super-capacity* if it operates at a greater efficiency under higher task load. Workload capacity is assessed using the *capacity coefficient*, which compares hazard functions of levels of task load to estimate an empirical measure of efficiency (Townsend & Eidels, 2011). This function can be compared to a theoretical benchmark, which represents performance under an unlimited-capacity, independent and parallel (UCIP) model. Different UCIP models are associated with different stopping rules (Houpt et al., 2014), but the interpretation of these models is the same—a capacity coefficient equal to the UCIP model represents an unlimited-capacity process, while capacity coefficients above or below the UCIP benchmark represent super- and limited-capacity processes respectively.

SFT can also be used to assess the presence of irrelevant stimuli on a process's efficiency. The presence of distractors can have a deleterious effect on a system's capacity to process relevant information. As a consequence, if irrelevant stimuli are present while a participant responds to a single task (the upper half of Figure 2), performance on this task may be worse than it would be in the absence of irrelevant stimuli, thereby distorting any capacity coefficients calculated using this data (Little,

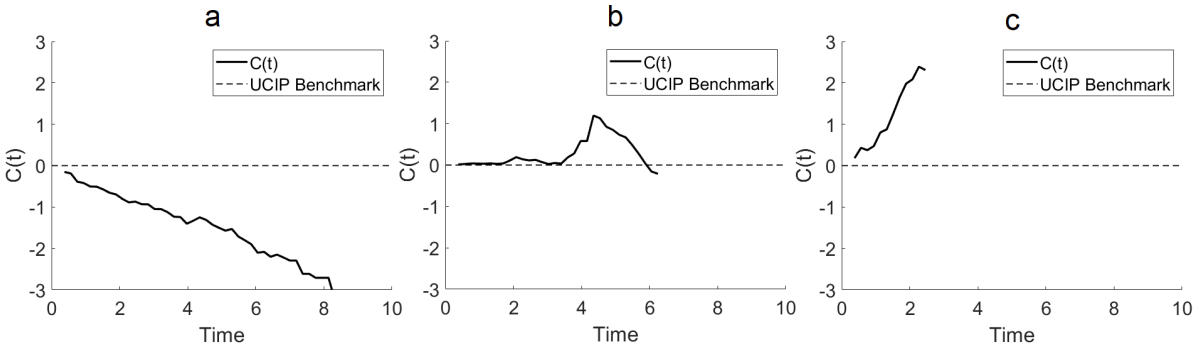

Figure 1. Representations of limited-capacity (a), unlimited-capacity (b) and super-capacity (c) coefficients compared to the $C_{OR}(t)$ benchmark model.

Eidels, Fific & Wang, 2015). The presence of distractors in a task does not preclude the use of SFT to estimate workload capacity—the *resilience function* can be calculated similarly to the capacity coefficient, whereby double-target data is compared to single-target data, but those single targets are assumed to have been presented alongside distractors (Houpt & Little, 2017; Little, Eidels, Fific & Wang, 2018). An alternative capacity analysis is the *single-target self-terminating* capacity coefficient, which compares target processing when the target is presented alone against the same process in the presence of a non-target stimulus (Blaha, 2011). The UCIP model associated with this measure assumes that the presence of a non-target does not reduce the efficiency with which targets are processed. Compared to this benchmark, a limited-capacity process would be negatively affected by the presence of non-target stimuli, whereas a super-capacity process would improve its processing efficiency when non-targets are present. In a dual-task paradigm, it is assumed that only the target stimulus must be processed before a response can be made (Blaha & Houpt, 2015).

The benchmark model in its "difference" form is defined as

$$C_{\mathrm{STST}}(t) = K_{\mathrm{A,\,X}}(t) - K_{\mathrm{A}}(t)$$

This model compares the reverse hazard function of the target channel, labeled 'A', with the reverse hazard function of the target channel in the presence of a task-irrelevant stimulus, labeled 'X'. The value of the UCIP benchmark is zero, with limited capacity performance characterised by a capacity coefficient of less than zero.

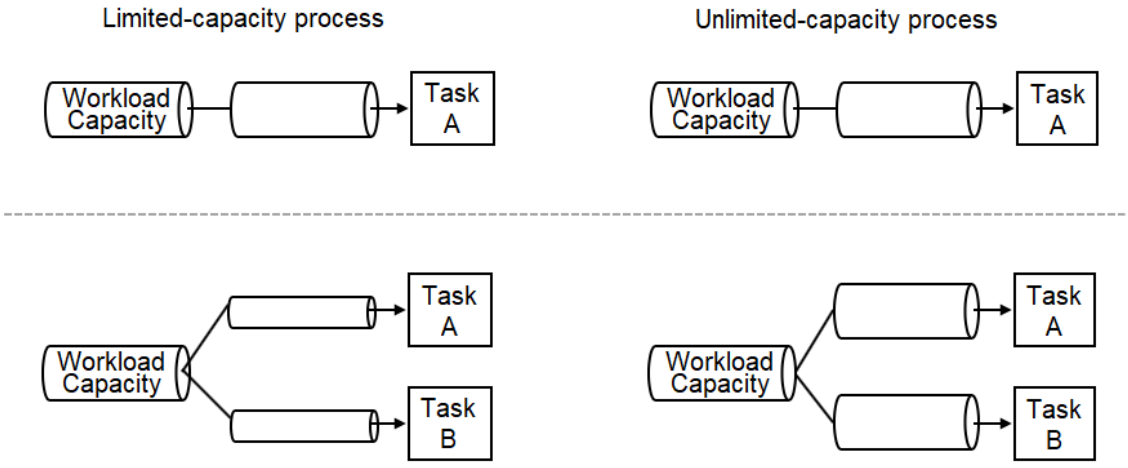

Figure 2. Representations of limited- and unlimited-capacity processes, with pipe diameter representing resource availability for each channel.

Compatibility Issues and the Current Study

The DRT operates by indexing the effect of another task on the user's workload state—it measures the extent to which the primary task impacts on DRT performance. The prospect of the primary task interfering with DRT performance is therefore expected. However, it is necessary to consider the potential impact of the DRT on the primary task. The DRT has not been found to increase participants' overall perceived workload (Strayer et al., 2015), though if the effect of the DRT's presence was small, participants may not have been sufficiently sensitive to changes in their own workload state to detect the change between conditions. Such a small impact of the DRT on primary task performance may be an acceptable cost in exchange for the knowledge gained from its use. Nevertheless, it would be prudent for researchers to identify and utilise the least invasive configuration of the DRT for a given experimental paradigm.

Of particular note is the potential impact of presenting the DRT using the same modality as the primary task. It may be, as has been theorised, that the DRT may impact primary-task performance less when the two tasks do not share resources (Bonnel & Hafter, 1998; Treisman & Davies, 1973; Wickens, 1980). However, Wickens (2008) noted that the predictions of Multiple Resource Theory may not extend to

residual capacity as the latter is measured by the DRT, but only to the state of overload and the relative extent to which performance fails as a result. If the combined load of the DRT and primary task is sufficiently low, it may be of no benefit for the two tasks to be presented using different modalities.

For the DRT to be well-controlled, the primary task used to impose workload should be continuous, as opposed to a discrete trial-by-trial task, so as to impose a consistent level of workload over time. Consider a task traditionally used in cognitive research, such as a signal detection task. Depending on which phase of a trial a DRT probe is presented, a participant could be perceiving stimuli, making a response, or sitting idly in an inter-trial interval. Each of these states could impose different levels of workload, and hence elicit different levels of performance on the DRT. For this reason, the DRT has not been used extensively alongside established methods of assessing workload in cognitive research. Palada et al. (2019) addressed this by presenting a primary task with a relatively slow required response alongside the DRT, and only analysing trials that occurred concurrently. Both tasks in this paradigm produced discrete data appropriate for cognitive modelling, but this approach is limited to tasks with naturally discrete structure—a continuous task such as driving does not naturally produce discrete data. SFT is relatively flexible in the form of data to which it can be applied, but the requirements of its experimental design limit its ecological validity, as it requires a redundant target paradigm. The ideal primary task for use with the DRT would be continuous in nature, while the ideal task for the application of SFT would be a dual task that generates response time distributions. A continuous task that is compatible with SFT would allow for a more sophisticated understanding of the workload demands of the task, as well as extending the use of SFT to a novel application.

The current study aimed to investigate the effect of cognitive load across DRT stimulus modality, and to identify the least invasive DRT configuration in the context of a visual task. Beyond comparing DRT performance and group-level main-task performance, we aimed to apply theoretical models of workload to both the DRT and

the primary task using complementary modelling techniques, to assess the workload demands of the primary task and the differential effect of interference in single-modal and multi-modal multi-tasking in both tasks. It was predicted that participants would exhibit faster and more accurate performance on both tasks when the DRT was presented using a different modality to the primary task, and slower, less accurate performance when the two tasks shared the visual modality. It was further predicted that processing efficiency in both tasks would decrease under high primary task load.

Method

Participants

Fifty undergraduates from the University of Newcastle (F=25, M=25) participated in the study. Mean age of participants was 22.4 years (SD=5.6). Participants were remunerated with course credit. The study was approved by the University of Newcastle Human Research Ethics Committee.

Design

The task had two independent variables: task load, which was manipulated by changing the number of targets to be tracked and had three levels (High load - two targets to track simultaneously, Low_{LEFT} - track only one moving target, on the left side, and $\text{Low}_{\text{RIGHT}}$ - track only the right target), and DRT salience, which had three levels (high/low/absent), resulting in a 3x3 factorial within-subjects design. Figure 3 depicts the objects used to calculate primary task performance—the yellow boxes are the targets to be tracked, and the black horizontal lines are the *needles*, or the objects that participants controlled. The dependent variable for the primary task was tracking error, defined as the absolute distance between the needle and the centre of the corresponding *target* at any point in time, measured in degrees of viewing angle as calculated from the participants' sitting distance. The dependent variables measured with the secondary task, when it was present, were mean response time (RT) and hit rate defined as the proportion of positive responses to presented stimuli. A

between-subjects factor of DRT modality was also used, with half of the participants presented with a visual stimulus, and the other half presented with an auditory stimulus. DRT modality was presented as a between-subjects factor to mitigate participant fatigue as an equivalent experiment with all factors presented within-subjects would have taken around three hours to complete.

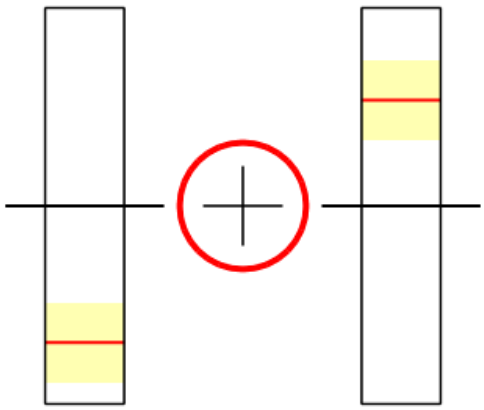

Figure 3. Example stimulus screen with tracking task gauges (left and right) and DRT stimulus (centre). Each gauge is composed of a target (yellow square with red centre line), response needle (black horizontal line) and target bounds (black rectangle).

Apparatus & Stimuli

Figure 3 shows an example screen from the task. Two gauges were presented, each subtending 5 degrees of viewing angle from bottom to top at a sitting distance of 80cm. The total viewing area of stimuli for both gauges and the secondary task was 5x5 degrees of viewing angle. Both gauges were presented whether or not they were to be attended to. Each gauge featured a target to be tracked, depicted in Figure 3 as the red horizontal line in the centre of the yellow box, which moved at a constant rate towards randomly-chosen destinations on the gauge. Upon reaching a destination, it would immediately begin moving towards a newly sampled destination. If the new destination required the target to continue moving in the same direction, the target's behaviour would not appear to change. Otherwise, the target would change direction to reach the

new destination. Each gauge also featured a tracking needle, to be used by the participant to track the corresponding target. These needles were controlled using a computer keyboard, with the 'a' and 'z' keys moving the left needle up and down, while the apostrophe and forward-slash keys moved the right needle up and down. When the response key was released, the needle would come to rest.

The DRT stimulus was presented as either a visual or auditory stimulus. The visual stimulus was a red circle around the fixation cross, while the auditory stimulus was a 2000Hz sine wave. As seen in Figure 4, the salience of the visual stimulus was manipulated by reducing the contrast between the stimulus and the background colour to 5% in the low salience condition, while the amplitude of the auditory stimulus as produced by the experiment's software was reduced to 5% in the low salience condition.

The DRT stimulus was presented for a duration of 1,000ms, at random intervals between 3,000-5,000ms from the end of the previous trial. A trial began when the stimulus was presented, and ended either when the participant responded, or after 2,500ms from presentation. A trial was considered a 'hit' when the participant responded within this 2,500ms window, and a 'miss' if the participant did not respond. A response was recorded as a false alarm if it was made outside the 2,500ms response window. Participants responded to the DRT stimulus using a PCsensor FS1-P USB footswitch.

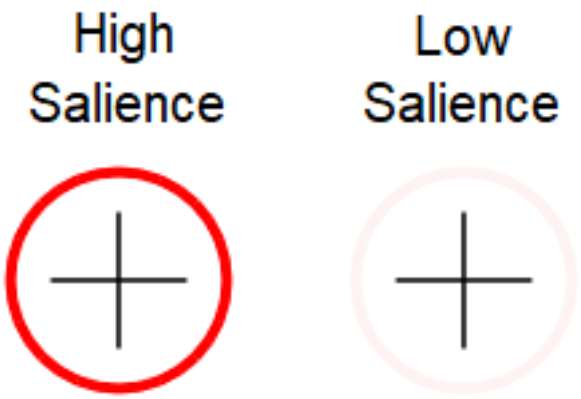

Figure 4. High and low salience DRT visual stimuli.

The task was computer-based, coded using Psychopy v1.90.3 and presented at a

60Hz refresh rate on a Dell S2240Lc LCD monitor, with a screen width of 53cm and 1920x1080 resolution. Audio stimuli were presented using Sennheiser HD-201 headphones.

Procedure

Participants were asked to report their age, dominant hand, English language proficiency, and visual and aural acuity. They were seated in a quiet room and positioned in front of the testing monitor, at a distance of 80cm. The footswitch used to respond to the DRT was placed beneath the participant's dominant foot. If the DRT stimulus was auditory, participants were fitted with headphones and presented with both high- and low-salience stimuli to ensure they could hear both tones. Participants completed an interactive instruction phase, in which the tasks were explained. Participants were not directed to give a higher priority to either task, but to complete both tasks to the best of their ability. They then completed a 30-second practice trial, before completing the experimental phase.

The main experiment was presented in three sections, with the three DRT salience conditions counterbalanced across participants. In each part, fifteen trials were presented, each of which lasted 60 seconds. Each of the three load condition was presented for five trials, with instructions for which gauge(s) to attend being presented at the beginning of each trial. Upon completing a trial, participants were required to wait for 15 seconds before continuing, to prevent fatigue. This procedure was presented both with and without the DRT present, the order of DRT-present and DRT-absent trials counterbalanced across participants. After completing each section, participants were asked to wait for 60 seconds before completing the next section. The experiment took approximately 90 minutes to complete. Upon completion, participants were debriefed.

Data Analysis

DRT performance was assessed both by comparing mean RT and hit rate by condition, and by modelling DRT responses. To assess the workload capacity demands

of the primary task, we intended to estimate capacity coefficients from primary task data. This analysis traditionally uses RT distributions, a discrete data form, for its calculations. As well as estimating capacity coefficients from distributions of tracking error, we aimed to transform the continuous tracking error data to a discrete form. To this end, we used two methods to derive discrete data from the tracking error time series—cross-correlation between the target and response time series, and a novel method of deriving distributions of *pseudo-response time*.

Cross-Correlation. We used cross-correlation to generate a time-based measure of performance by comparing target movement data to response movement data (Heath, 2000). Given the response time series in a given experimental block is the participant's reaction to the target time series in the same block (in the current study, the needle moves in response to the target), the former time series can be shifted forwards in time to find the time lag at which the two time series best correlate. The extent to which the time series needs to be shifted reflects the latency of the participant's response. Figure 5 shows how two time series can be cross-correlated by shifting one time series—the left panel represents raw time series data, while the right panel represents the time series shifted to best correlate. The number of lags that produces the best correlation can then be multiplied by the resolution of the data (in the current study, the sample rate of the display), producing a value that represents the time the participant took to respond to the target on average. By doing so, we produced discrete RT data from a continuous task, which could then be compared to continuous results.

Pseudo-RT Distributions. The left panel of Figure 6 depicts the stimuli participants were asked to track. The yellow box around each target was included as a guide, but for the purpose of data transformation its limits were defined as the limits of acceptable performance - when the participants' needles were within this box, which subtended 0.5°on either side of the target, they were deemed to be performing the task to an acceptable degree. When tracking error exceeded this error threshold, an "error" event began, which ended when the needle returned to within the acceptable performance limits. The right panel of Figure 6 depicts an example time series of

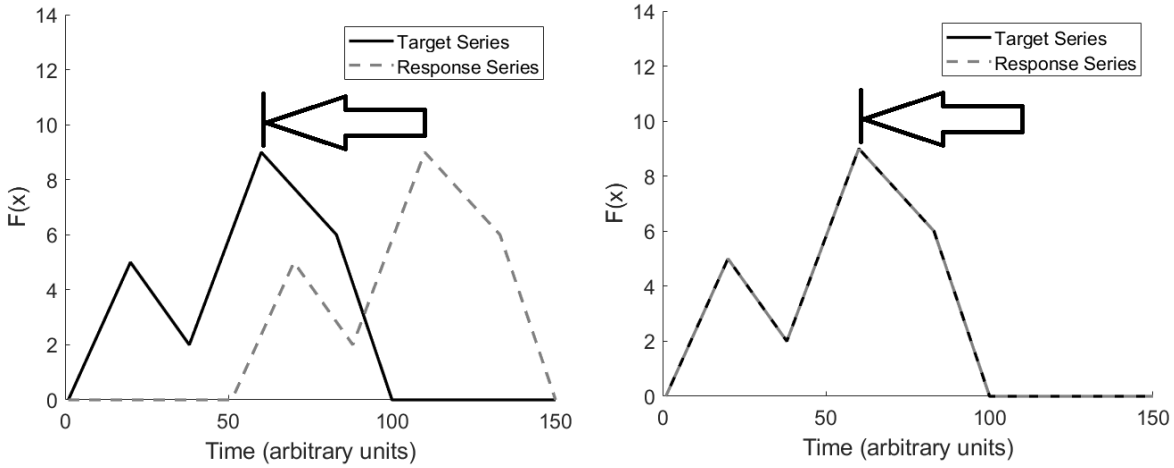

Figure 5. Example cross-correlation, showing raw time series data (left) and data where the response times series was time-shifted to the point of best correlation (right).

tracking error, with the threshold of acceptable performance represented by the dotted line. The time between an error event beginning, represented in Figure 6 as the point the time series crosses above the threshold, and the end of this error event, represented as the point the time series returns below the threshold, was treated as a pseudo-response time. The red lines in Figure 6 represent the length of each pseudo-RT. Pseudo-RT distributions for each load condition were calculated using this method.

Results

Seven participants were excluded due to technical malfunction, low DRT accuracy (<33% overall hit rate), and non-responses to primary tracking task trials.

DRT

To address the issues of whether the software-based DRT was sensitive to load, and whether DRT modality and salience affected DRT responses, mean response times (RT) and hit rates (HR) were compared using repeated-measures ANOVA and Bayesian repeated-measures ANOVA. Bayesian tests were applied only in cases of non-significant or ambiguous results, to present evidence in favour of no difference; this cannot be achieved using frequentist tests alone. The reported Bayes Factors are Inclusion Bayes Factors, which reflect the strength of evidence in favour of including a factor in an

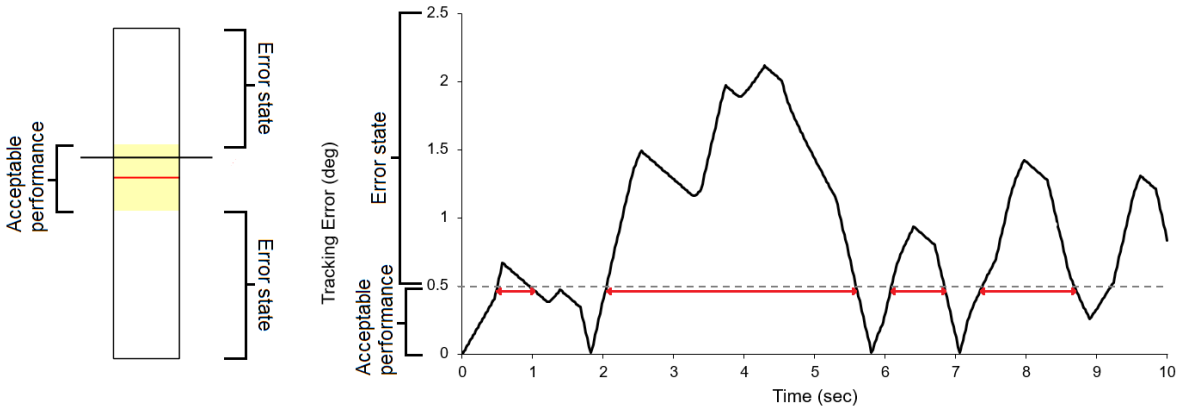

Figure 6. Tracking task stimuli, with the limits of acceptable performance used for pseudo-RT transformation (left), and example tracking error time series with threshold of acceptable performance represented by the dotted line (right). The length of each error event (red lines) represents the length of each pseudo-RT.

explanatory model of the data. Unlike the more commonly used BF_{10}, which compare the effect of a factor to a null model, the $BF_{\text{Inclusion}}$ takes into account the relative likelihood of other factors having an effect on the data. However, these Bayes Factors can be interpreted the same as BF_{10}, according to the classification scheme presented by Lee and Wagenmakers (2014), from Jeffreys (1961). Effect sizes were interpreted according to Cohen's (1988) conventions.

A main effect of primary tracking task load on DRT RT was found. As the left panel of Figure 7 shows, when primary tracking task load was high, mean RT was slower (M=585ms, SD=138ms) than when load was low (M=506ms, SD=98ms), F(1, 41)=29.25, p<.001, η^2=.10. This effect size was moderate, but consistent for both the visual DRT condition, F(1, 22)=32.90, p<.001, η^2=.17, and the auditory DRT condition, F(1, 19)=8.47, p=.009, η^2=.08. The same main effect was found for mean HR. Under high load conditions, participants responded to fewer probes (M=95%, SD=7%) than under low load (M=97%, SD=7%), F(1, 41)=4.33, p=.044, η^2=.01. When separating data by DRT modality, this small main effect was only found for the auditory DRT condition, F(1, 19)=5.05, p=.037, η^2=.03. In light of this, Bayesian analysis was carried out, which indicated ambiguous evidence against including primary

tracking task load in an explanatory model of the data, $BF_{\text{Inclusion}}$=0.55. Figure 7 also shows the main effect of DRT modality on mean RT in the left panel. Participants who were presented auditory DRT stimuli responded faster (M=512ms, SD=145ms) than those who were presented visual stimuli (M=579ms, SD=97ms), $F(1, 41)$=5.38, p=.025, η^2=.12. No significant interaction between primary tracking task load and DRT modality was found.

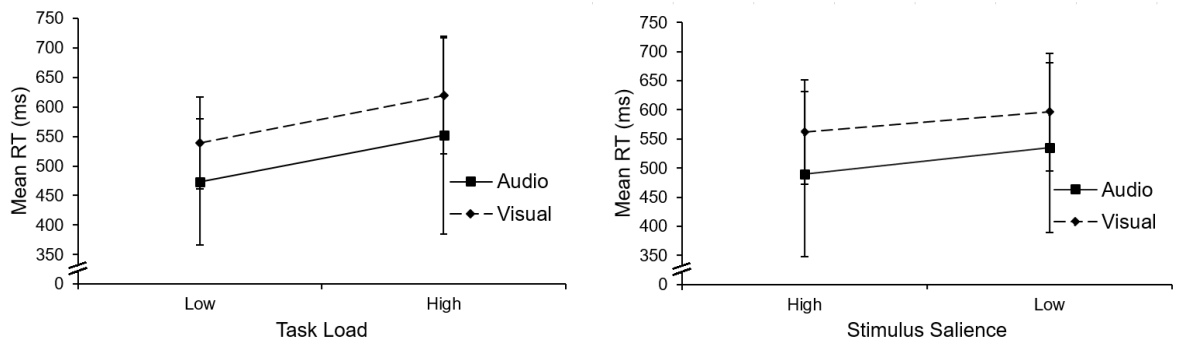

Figure 7. Mean response times to DRT across primary tracking task load (left), stimulus salience (right) and DRT modality. Error bars represent one standard deviation.

The salience of the DRT stimulus also affected DRT performance. Participants responded faster to high salience stimuli (M=525ms, SD=120ms) than to low salience stimuli (M=566ms, SD=120ms), $F(1, 41)$=12.61, p<.001, η^2=.03. The right panel of Figure 7 shows this small effect. Participants who were presented auditory stimuli were faster to respond to high salience stimuli (M=489ms, SD=146ms) than to low salience stimuli (M=535ms, SD=142ms), $F(1, 19)$=6.65, p=.018, η^2=.03. Participants in the visual group were also faster to respond to high salience stimuli (M=561ms, SD=101ms) than to low salience stimuli, (M=596ms, SD=90ms), $F(1, 22)$=5.78, p=.025, η^2=.03. No significant differences were found in mean HR between DRT modality or salience conditions. Bayesian analysis presented ambiguous evidence in favour of including salience as an explanatory factor, $BF_{\text{Inclusion}}$=1.31, and weak evidence against including modality, $BF_{\text{Inclusion}}$=0.27. These findings constitute evidence in favour of the hypothesis that DRT modality and salience affect DRT mean

RT, but mixed evidence regarding the effect of primary tracking task load and salience on HR, and evidence against modality affecting HR.

DRT Model. Following this initial analysis, we fit a shifted-Wald model to the data using the JAGS module developed by Steingroever, Wabersich and Wagenmakers (2018). Similar to previous studies that applied this model to DRT data (Castro et al., 2019; Palada et al., 2019; Tillman et al., 2017), we fit a single-bound model to response time data from the DRT to investigate the latent processes contributing to the response trends discussed above.

We separately fit 8 single bound diffusion models to the DRT data. The 8 models represented the full parameter space where modality (of the DRT) and primary tracking task load—i.e. one or two stimuli to track—were able to vary for the 7 combinations of parameters. We also fit a null model with no variability (i.e. three group level parameters which did not vary with modality or primary tracking task load). Table 2 provides a comparison of the deviance information criterion (DIC) values for each model, which shows that the full model, which allowed a, v and t_0 to vary, had the best fit to the data. This model assumed that a and v and t_0 varied with modality and primary tracking task load, so that individual parameters were drawn from unique group distributions. As 2 shows, this model had the most *effective parameters*, a value which represents the sum of all free parameters in the model across the individual and group-level data sets. Nevertheless, its low DIC value indicated it was the winning model. A full breakdown of the prior distributions can be seen below. Additionally, we opted not to include salience effects as a factor in the model as we showed there was no evidence to suggest a change in responding between low and high salience stimuli. To ensure we had not overlooked this as a confounding variable in the model, we fit the winning model to the low salience and high salience data separately and found no evidence of a parameter shift between these conditions.

The winning model shows that primary tracking task load had a clear effect on both the rate of evidence processing, the decision threshold and the non-decision time. Furthermore, it appears that modality of the stimulus also affects response patterns.

Table 2

DIC values for each model variant. The winning model is in bold.

	DIC Value	Effective Parameters[a]
$a + v + t_0$	**-21386**	283
$a + v$	-21139	236
$a + t_0$	-21034	236
$v + t_0$	-21202	236
a	-20753	189
v	-21118	189
t_0	-20525	189
Null	-20130	142

[a] "Effective parameters" represent the sum of all free parameters in the model across individual and group-level data sets.

Figure 8 shows the posterior distributions for the model parameters. To assess between-groups effects, we compared the posterior parameter distributions for each modality condition. The p-values below represent the probability that the difference between posterior parameter distributions was less than or equal to 0. Thresholds were comparable across load and modality conditions, with the only difference shown between the auditory and visual conditions when under low workload (low visual-high visual; p=.599, low auditory-high auditory; p=.919, high visual-high auditory; p=.104, low visual-low auditory; p=.001). There were some differences observed across rates of evidence accumulation, especially across workload conditions (high visual-low visual; p=.057, high auditory-low auditory; p=.001, high visual-high auditory; p=.503, low visual-low auditory; p=.297). There were comparable estimates of non-decision time across all conditions, with the exception of the difference between modalities in the low workload condition (low visual-high visual; p=.768, low auditory-high auditory; p=.311,

high visual-high auditory; p=.591, low auditory-low visual; p=.041).

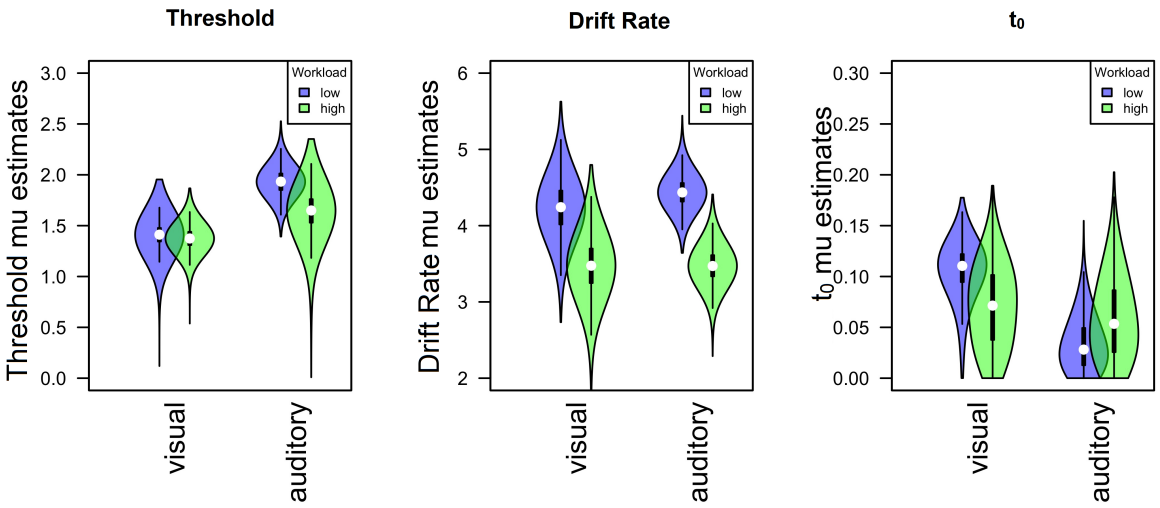

Figure 8. Posterior distributions of threshold, drift rate and non-decision time across DRT modality and primary tracking task load.

There is a clear effect of primary tracking task load on rate across modalities, with higher task load seeming to reduce the rate of evidence accumulation. This is in line with previous research, such as (Castro et al., 2019; Tillman et al., 2017), which similarly showed large differences in drift as a result of task load manipulations. In addition to this, the interaction effect seen, shows that this effect is greater in the auditory channel, with a more marked decrease to rate observed in the high load condition. An effect of modality is shown, with the auditory group setting higher thresholds than the visual stimulus group. This suggests that participants in the auditory condition were more cautious than those in the visual condition.

Primary Tracking Task

Task Trade-off. The DRT can be considered an index of residual capacity, when taken in the context of primary tracking task performance. However, the finding that DRT performance was slower or less accurate in high-load does not necessarily imply lower residual capacity. An alternative explanation is that participants traded off between the two tasks, deliberately allocating resources from the DRT to the primary tracking task. In this scenario, resources would still be shared between the two tasks,

but changes in DRT performance could not be taken as a reliable indicator of changes in the user's workload state. If this was the case, we would expect to see primary tracking task performance improve, or at least not decrease, under high load conditions. To address this question, primary task tracking error was analysed using repeated-measures ANOVA.

As the left panel of Figure 9 shows, mean tracking error was found to be higher when primary tracking task load was high (M=0.53°, SD=0.18°) than when task load was low (M=0.31°, SD=0.08°), $F(1, 41)$=141.33, p<.001, η^2=.39. This pattern of results suggests participants were not trading off between tasks, as performance was worse on both tasks under high primary tracking task load than under low load.

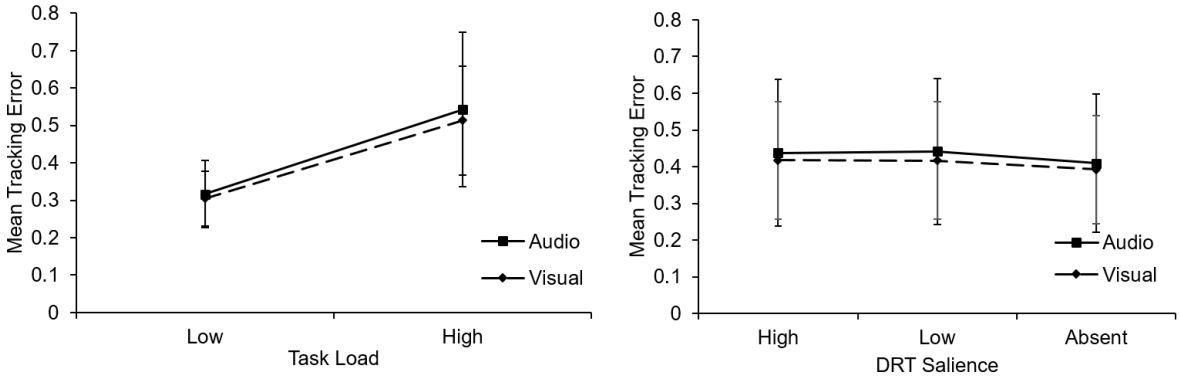

Figure 9. Mean tracking error across primary tracking task load (left) and DRT salience conditions (right).

DRT Interference on Primary Tracking Task. The questions of whether the presence of the DRT interfered with primary tracking task performance, and what differential effect the DRT's modality and salience had on primary tracking task performance, were addressed by comparing primary task tracking error using repeated-measures ANOVA and Bayesian repeated-measures ANOVA.

As the right panel of Figure 9 shows, mean tracking error was affected by the salience of the DRT stimulus, $F(2, 82)$=9.50, p<.001, η^2<.01. In light of a very small main effect size, post-hoc contrasts were carried out which showed that conditions in which the DRT was present (DRT-High and DRT-Low) had significantly higher mean tracking error (M=0.43°, SD=0.18°) than conditions in which the DRT was not

presented (M=0.40°, SD=0.17), $t(42)$=4.37, p<.001. However, the contrast of the two DRT-present conditions indicated no significant difference between mean tracking error when the DRT stimulus was high-salience (M=0.43°, SD=0.18) and when the DRT stimulus was low-salience (M=0.43, SD=0.18), $t(42)$=0.17, p=.863. Bayesian analysis only offers ambiguous evidence for whether the factor of DRT salience should be included in an explanatory model of the data, $BF_{\text{Inclusion}}$=1.28. These findings suggest the main effect of DRT salience on mean tracking error was driven by the presence of the DRT stimulus, rather than its salience.

Mean tracking error of participants who were presented auditory DRT stimuli was not significantly different to those who were presented visual DRT stimuli, $F(1, 41)$=0.31, p=.58, η^2<.01. Bayesian analysis indicated some evidence against the hypothesis that the factor of modality, depicted as the two lines in Figure 9, should be included in an explanatory model, $BF_{\text{Inclusion}}$=0.47. This finding is consistent with the hypothesis that DRT stimuli from different modalities do not interfere with the primary tracking task to different degrees.

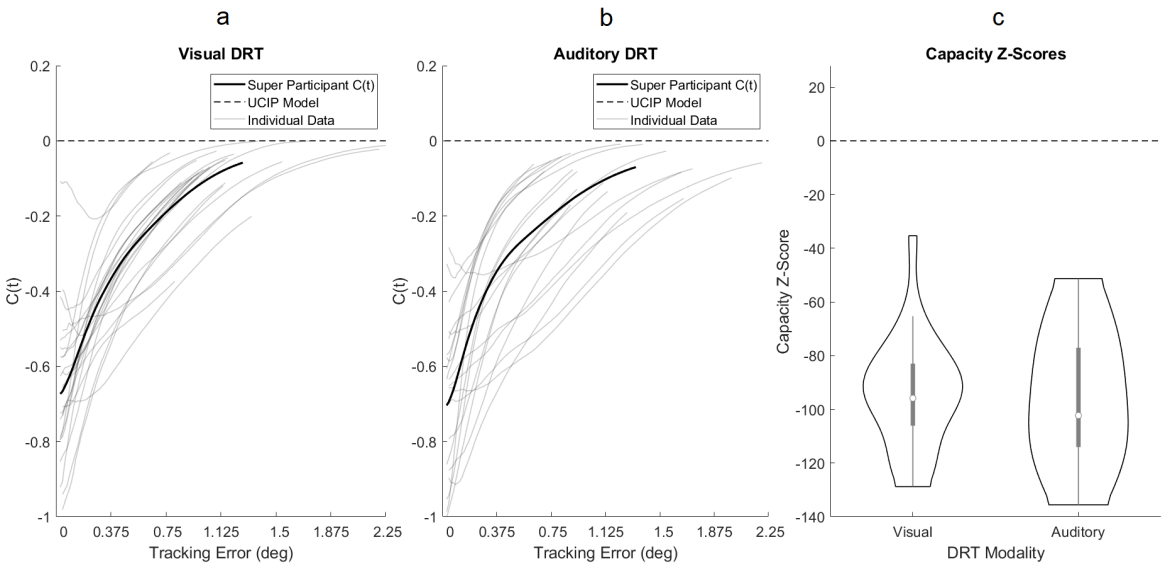

Figure 10. Capacity coefficients for participants from visual (a) and auditory (b) modalities, and capacity Z-scores (c).

Workload Capacity. Tracking error distributions from the single- and double-target conditions were compared at the participant level using the *single-target*

self-terminating benchmark model, which assumes performance on each task is unaffected by the presence of a distractor or second task. The difference form of the capacity coefficient is used below as this form allows for significance testing of the capacity coefficient against the UCIP benchmark (Houpt & Townsend, 2012).

Capacity coefficients were calculated for each participant, but for the sake of illustrating group-level trends, we also aggregated data from each DRT modality group into a single *super-participant*, and calculated capacity coefficients for the two aggregate data sets. Figure 10a and 10b show the resulting capacity coefficients from these analyses. Individual capacity coefficients are also presented alongside the super-participant for each group, to illustrate within-group variance.

At no point did any capacity coefficients exceed the UCIP benchmark, depicted in Figure 10 as the dashed line, indicating a limited-capacity system. This pattern was consistent across both modalities, and for all individual capacity coefficients as well as both super-participants. To supplement this analysis, significance testing on individual capacity coefficients was carried out using Houpt, et al.'s (2014) SFT package in R. This test assesses whether any difference between the capacity coefficient function and the UCIP benchmark is significantly different to zero. In all cases across both modalities, capacity coefficients were significantly below the UCIP benchmark. The distribution of z-scores from these tests are presented in Figure 10c, which shows that all participants from both modalities showed highly significant capacity limitations.

It should be noted that, because tracking error was sampled at 60 data points per second across several minutes of testing for each condition, the resulting data sets were very large compared to those collected in traditional SFT studies. We would therefore expect any significance testing to return a highly significant result, due simply to the number of observations. The magnitude of the z-scores in Figure 10c reflect the high significance of the statistical tests, rather than an effect size. These results should therefore be considered in terms of the trends depicted by the capacity coefficients in Figure 10a and 10b, their consistency across participants and modalities, and our confidence in these trends as a result of significance testing.

Cross-Correlation. Primary tracking task performance was also assessed based on cross-correlated response times as described in the Method section. Time series of target direction changes were cross-correlated with equivalent response needle direction changes. Direction changes were used for the cross-correlation rather than raw time series data as the latter were auto-correlated. Cross-correlation times below zero were removed as the relevant relationship was the extent to which target direction changes predicted responses, not the reverse. Cross-correlation times greater than ten seconds were also removed as the time distance was deemed too great to be meaningful. Group-level analysis was then carried out across load and salience conditions.

As the left panel of Figure 11 shows, mean cross-correlated RT was significantly slower under high load (M=709ms, SD=343ms) than under low load (M=446ms, SD=53ms), $F(1, 41)=28.36$, $p<.001$, $\eta^2=.13$. However unlike the tracking error data, a small main effect of salience was found in this data set, $F(2, 82)=3.80$, $p=.026$, $\eta^2=.02$, though Bayesian analysis presented ambiguous evidence against including salience as an explanatory factor, $BF_{\text{Inclusion}}=0.78$.

As the right panel of Figure 11 shows, the significant contrast between DRT-present and DRT-absent conditions that was present in the tracking error data was not found here, though it approached significance, $t(42)=1.86$. $p=.067$. As before, no significant difference was found between high-salience and low-salience conditions, $t(42)=1.646$, $p=.322$, $BF_{10}=0.43$. No significant effect of DRT modality was found, $F(1, 41)=0.52$, $p=.477$, $\eta^2=.01$ with Bayesian analysis presenting some evidence against including DRT modality as a factor, $BF_{\text{Inclusion}}=0.25$. This pattern of results was similar to those found from tracking error data, though the strength of these findings was lower, perhaps due to the relatively small data set used when compared to the very large tracking error data set.

Pseudo-RT Analysis. The capacity analysis above was carried out using the distributions of tracking error as substitutes for RT distributions. An issue with this analysis is that capacity coefficients have previously only been estimated from RT distributions, rather than continuous data sets. To address this, we derived pseudo-RT

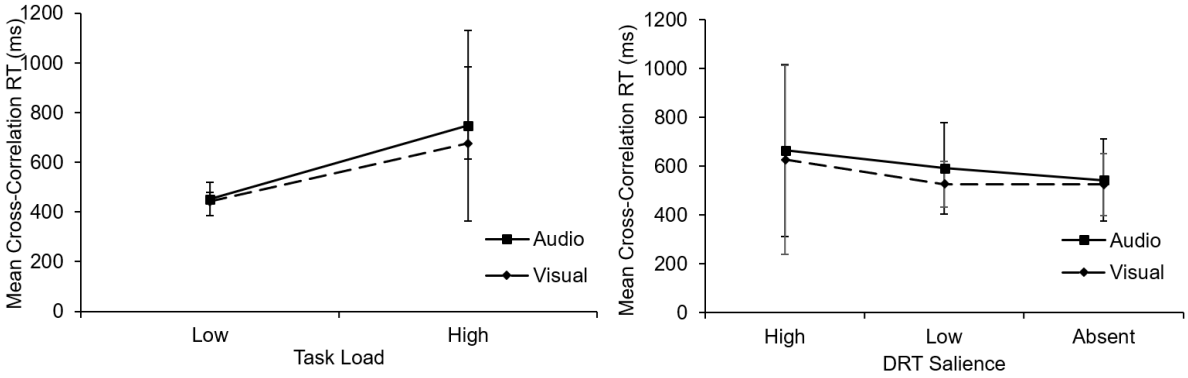

Figure 11. Mean cross-correlation RT across primary tracking task load (left) and DRT salience conditions (right).

distributions as described in the Method section, and estimated capacity coefficients from these discrete distributions.

Before applying a capacity analysis to the pseudo-RT distributions, a group-level analysis of means was carried out, to assess whether the pseudo-RTs depicted a similar trend of results to the tracking error data. As with tracking error, mean RT was significantly slower under high load (M=1016ms, SD=355ms) than under low load (M=594ms, SD=156ms), $F(1, 41)=121.89$, $p<.001$, $\eta^2=.37$. As Figure 12 shows, the trend seen in the tracking error data was replicated; the only significant difference observed was between conditions in which the DRT was present (M=816ms, SD=350ms) and conditions in which it was absent (M=784ms, SD=338ms), $t(42)=2.32$, $p=.023$. No significant difference was found between high-salience (M=810ms, SD=351ms) and low-salience conditions (M=822ms, SD=351ms), $t(42)=0.75$, $p=.453$, and Bayesian analysis presented evidence against salience as a factor in an explanatory model, $BF_{\text{Inclusion}}=0.12$. Additionally, no significant difference between DRT modality groups was found, with Bayesian analysis presenting evidence against including DRT modality as a factor, $BF_{\text{Inclusion}}=0.45$. These results suggest the pseudo-RT data reflects similar trends to the tracking error data from which it was derived, although the strength of these findings was lower than with the tracking error data, as reflected in the higher p-values.

The workload capacity analysis outlined above was repeated using the derived

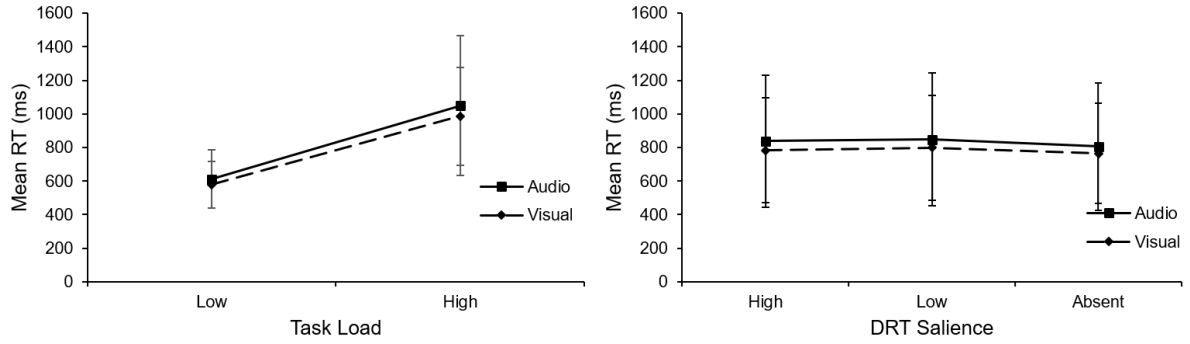

Figure 12. Mean pseudo-RT across primary tracking task load (left) and DRT salience conditions (right).

pseudo-RT distributions, with results presented in Figure 13. These results suggest a limited-capacity process, similar to the tracking error data. However, the individual capacity coefficients are more variable in their spread around the super-participant coefficient, particularly in the left and right tails of the coefficients. This, along with the capacity z-scores being closer to zero, reflect the smaller number of data points used in this analysis relative to the very large tracking error data sets. Nevertheless, this pattern of results suggests that the pseudo-RT distributions produce capacity results consistent with the tracking error data—using the same analytic tools, we come to the same conclusion of limited capacity using both data sets.

Discussion

The DRT produced results in line with those expected from previous studies, with faster responses under conditions of low load than high load. The result of faster responses to high-salience stimuli compared to low-salience stimuli was also predicted. The primary finding from mean DRT RT was that participants who were presented an auditory signal responded faster than those who were presented a visual signal. This could be a result of the auditory pathway generally producing faster RTs than the visual pathway (Green & Gierke, 1984). Another potential explanation is the split of resources theorised under Multiple Resource Theory—the auditory DRT stimulus may have drawn upon a different resource than the visual primary tracking task, whereas in the visual condition both tasks drew from the same pool of resources.

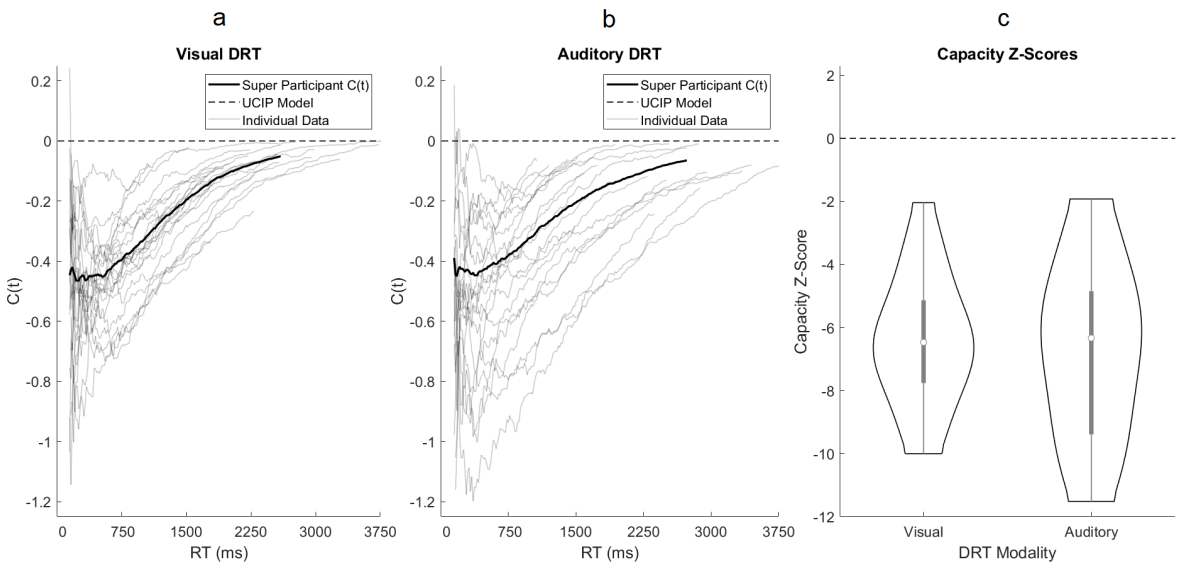

Figure 13. Capacity coefficients for visual DRT (a) and auditory DRT groups (b) derived from pseudo-RT distributions, with capacity z-score distributions (c).

Although participants responded to auditory DRT probes faster than visual probes, there was no significant difference in the level of interference from the DRT on the primary tracking task when the DRT's modality was manipulated. This is in contrast with the prediction that presenting two tasks using the same modality causes greater interference than presenting tasks using different modalities. These findings, when taken together, suggest it is the mere presence of a secondary task, rather than its salience or modality, that drives the interference effect. It is intuitive that the DRT causes some interference; for the DRT to measure residual capacity it must place some workload on the user, if only a relatively small amount. It is surprising, however, that a lower-salience stimulus did not impose a higher workload on users than a high-salience stimulus. This may be because the difference in workload between perceiving stimuli of different salience is too small to have a differential effect on the primary tracking task, which places a relatively high load on the user. It may also be that the difference between the salience conditions was not sufficient to produce the effect, and stimuli with more marked salience differences may be necessary. It was similarly surprising that a DRT stimulus presented in the same modality as the primary tracking task did not interfere more than a stimulus presented in a different modality to the primary tracking

task. As with salience, it may be that the differential interference effect, if it existed, was too small to have an effect on primary tracking task performance.

The use of a continuous variable, in this case tracking error, in place of RT in a capacity analysis was novel, as was the use of and converted pseudo-RT distributions in the same analysis. Both the tracking error and pseudo-RT data produced limited-capacity coefficients across all participants. Capacity z-scores derived from the tracking error series were markedly higher in magnitude than those derived from pseudo-RT, not because the effect was greater for the former data-set, but because the relative richness of the data-set increased the magnitude of the difference between the conditions relative to the within-group variance. As mentioned in the Results section, this difference between measures should not be taken as a sign of increased power in one than the other. Nevertheless, the distributions of capacity z-scores across DRT modalities within a measure can be compared, at least visually, with no visible difference between the groups on either measure. This suggests a relatively stable capacity limitation, whether in the presence of an auditory or visual DRT.

We can only be confident in our findings if we are also confident that a capacity analysis of pseudo-RTs is sensitive to processes with different workload capacity. To address this, we simulated pseudo-RT data to represent limited-, unlimited-, and super-capacity processes. We found that capacity coefficients estimated from these simulated data sets correctly identified the workload capacity of each set (see Appendix A for details of simulation and findings).

The results from capacity analysis suggest the concurrent tracking of the two moving targets resulted in interference between the two channels. This may be an obvious conclusion based on mean tracking error on the primary task increasing under high load, but a decrement in performance on the double-target condition does not necessarily imply the channels processing the targets are not operating independently and in parallel, as discussed in the Introduction. However, given the UCIP benchmark for the single-target self-terminating model assumes no decrement in performance in the presence of a distractor, it is safe to conclude performance in the current study was at

best limited-capacity, if not severely limited. Overall, our analysis suggests interference between the two channels. Given that both parts of the primary tracking task were identical and hence both visual, it is theorised under Multiple Resource Theory that the tasks would interfere with one another, both at the peripheral perceptual level, and the central response level.

The finding of capacity limitation from our analysis of primary tracking task data corresponds to the primary finding of the DRT response time modelling, which showed drift rate was heavily affected by task load. This is an unsurprising finding, as the DRT's sensitivity to interference from other tasks is the central premise of the measure. This effect was stable across modality, suggesting multi-modal multi-tasking was not more efficient than multi-tasking within the same modality. Indeed, response thresholds were generally higher in the auditory group, suggesting greater caution as a result of multi-modal multi-tasking. The drop in thresholds under high load suggests participants dealt with the effect of task load by becoming less cautious in their responses to the DRT. These two parameters alone did not provide the best account for the data, so non-decision time was included in the model. However, non-decision time was poorly estimated, so findings regarding non-decision time should be regarded with caution.

The winning model generally provided a good fit of the data, but the model overestimated RTs in the 90th quantile. This was a result of the model failing to account for go-failure in the empirical data, where the accumulation process does not start as expected. Responses longer than 2,500ms were recorded as late misses in the empirical data, but no such limitation was placed on data generated by the model. This resulted in a very small number of generated RTs that were much slower than any empirical data points, which in turn resulted in more substantial right tails being predicted by the model.

In the current study, the most complex model was chosen as the winning model, similar to Castro et al.'s (2019) study. Our finding that DRT drift rate was affected by task load is in contrast to Tillman et al.'s (2017) finding, but otherwise in line with other findings from modelling studies. A point of divergence from the literature came in

our finding that response threshold did not increase with task load. In other studies, threshold was either found to vary with task load (Castro et al., 2019; Tillman et al., 2017), or it was not included in the winning model for DRT responses (Palada et al., 2019). This difference could be explained by the demands of the primary task. Palada et al. (2019) found that the effect of task load on both response threshold and non-decision time interacted with the response deadline of the primary task, with a more urgent primary-task deadline resulting in DRT response threshold remaining constant across load conditions. A continuous task with a constant response pressure may have a similar effect, though this is not supported by Castro, et al.'s (2019) finding, which used a similar one-dimensional tracking task as the primary task. More generally, the current results did not support Multiple Resources Theory. One would expect to see lower drift rates for the visual group, as both the DRT and primary tracking task drew from the same peripheral resource. This was not found, though there was greater variance in the drift rate parameter estimates for visual-group participants than for auditory-group participants. The only finding that suggests multi-modal multi-tasking carried some advantage over dual visual tasks came from non-decision time, which was lower for the auditory group. Again, this is not predicted by Multiple Resource Theory, which theorises that resources for executing responses are central, and therefore should not be different across modalities (Wickens, 2002). However as noted above, our findings related to non-decision should be treated with some degree of caution.

Perceptual Load Theory predicts task interference as a result of task complexity and high perceptual load, without making specific predictions about the modality of stimuli (Lavie, 1995). It could therefore explain the effect of increased primary tracking task load, both on primary tracking task and DRT performance, while also accounting for the relatively uniform interference effect from the DRT on primary tracking task performance. If the cognitive demand of attending to both the primary tracking task and the DRT has a similar impact on the cognitive control mechanism regardless of DRT stimulus modality, then interference effect could manifest in a similar way, as we saw in our findings.

Considerations

The conversion of tracking error to pseudo-RT distributions allowed the use of SFT, but the manner in which these pseudo-RTs were derived was somewhat arbitrary. The threshold for 'acceptable' performance was meaningful to the task, as it had been explicitly depicted by the stimulus; there was a perceptual difference between the needle being within the bounds of the target box, and hence within the 'acceptable' threshold, and the needle being outside these bounds. However, participants were not given any instructions regarding the threshold, so what participants deemed acceptable performance may not have matched the threshold chosen. Indeed, choosing a different threshold of acceptable performance was possible, if not as meaningful to the current task. If participants did not attempt to track the target as closely as possible, but rather considered a different proximity to the target to be "acceptable", capacity estimates would change for different threshold values. To investigate this possibility, we collected pseudo-RT distributions from multiple thresholds, from 0.2°to 1.0°, and used these distributions for the capacity analysis outlined in the Results section alongside the original threshold of 0.5°. Figure 14 shows the distributions of capacity Z-scores for these threshold values. This analysis indicates capacity estimates were stable across different threshold values. There were subtle differences between the distributions at lower threshold values, though capacity estimates remained limited. The most marked difference is between lower values and the two highest values, 0.9°and 1.0°. This is not necessarily a limitation of the method used; if one distribution contains a higher proportion of fast RTs than another, then truncating the left tail of that distribution would selectively affect the fast distribution, leading to more homogeneous distributions and therefore changing the outcome of any comparison between the two distributions. The trend of limited capacity was also apparent with these more lenient threshold, though to a lesser extent compared to more strict thresholds. Nevertheless, this suggests that the choice of threshold value has some effect on the outcome of the analysis. For this method of data conversion to be valid, the threshold value should at least be meaningful to the task, so as to reflect a threshold that the participants may have

perceived.

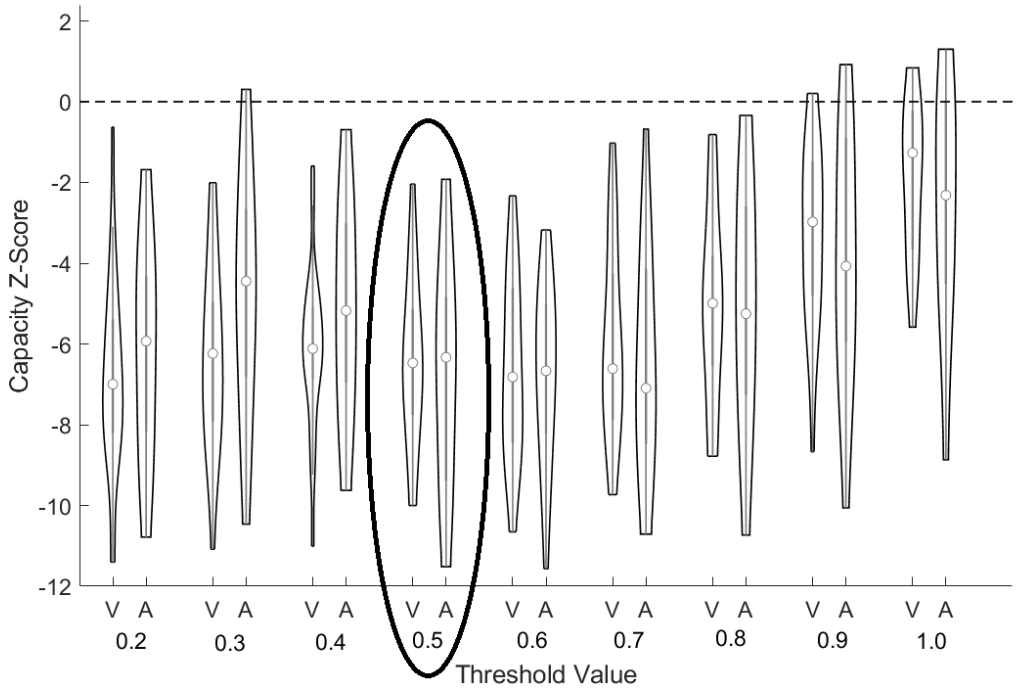

Figure 14. Capacity z-score distributions for multiple threshold values. Original threshold value used for the analysis was 0.5°(encircled), determined by the size of the target stimulus.

Another issue relating to the size of the target threshold of acceptable performance is that changing the size of the threshold also changes the number of observations gathered. Consider an experimental block in which the participant maintains "acceptable" performance, as defined by an overly generous threshold, for the entire block. This block would not produce any pseudo-RTs, as the needle would not travel outside the threshold at any time. Indeed, one individual who exhibits higher overall performance than another individual would necessarily produce fewer pseudo-RTs over the same amount of participation time. A paucity of pseudo-RT data represents an issue with this transformation technique, though capacity coefficients estimated from smaller data sets may still recover the underlying workload capacity of a system (see Appendix B). Two solutions to this issue are readily available. First, a researcher could pre-test several thresholds of acceptable performance while developing

a task, to identify an optimal value which renders the task difficult enough to produce pseudo-RTs, while also addressing the issue of truncating distributions discussed above. A second solution would be to simply ensure that experimental sessions run long enough to produce sufficient data. Given that participant fatigue is a common issue, and one that factored into the current study's design, this may present its own issues in practice.

A limitation of the current study's design was that no psycho-physical pretesting was carried out to establish the levels of salience to be used in the main experiments. The decision to limit the low-salience DRT stimulus to 5% of the high-salience stimulus on a single parameter—colour contrast for the visual stimuli and amplitude for the auditory modality—was arbitrary, with the goal of merely being sufficiently extreme to elicit a salience effect. It may be the case that our manipulation did not reduce the salience of the stimulus as much as expected, resulting in the unexpected lack of differential interference on the primary tracking task. This could be controlled in future research by using a photometer and decibel reader to ensure stimulus levels are controlled. Additionally, stimuli in both modality conditions were reduced by the same relative amount—95%—which assumes that auditory and visual perception are equally sensitive to changes in stimulus intensity. This issue could be addressed by carrying out psycho-physical pretesting to ascertain which salience levels should be used.

Although the current study utilised the capacity coefficient, this is not the only analysis that was available under the SFT framework. The effect of task-irrelevant stimuli being presented alongside targets could be measured using the *resilience function* (Houpt & Little, 2017). As described in the Introduction, this measure is similar to the capacity coefficient in that it compares double-target performance to equivalent single-target performance, but differs in that it utilises single-target trials that are presented alongside distractors. It therefore predicts a different benchmark model to the capacity coefficient, as it takes into account the resources required to process distractors. As with other measures within the SFT framework, processes with different architectures, stopping rules and capacity limitations produce different resilience functions, which can be compared to empirical data to identify a given

process's properties. Unlike the capacity coefficient, some inferences can be made regarding architecture and stopping rule without the need to manipulate the salience of stimuli. However, multiple combinations of architecture and stopping rule predict the same resilience function, and can therefore only be differentiated using a full double-factorial design (Fific & Little, 2017; Houpt & Little, 2017). As the current study did not contain a salience manipulation for the primary tracking task, there would be no way to assess the relative strength of non-target signals, so findings from this method could have been ambiguous. Additionally, the resilience function, like $C_{OR}(t)$ and $C_{AND}(t)$ forms of the capacity coefficient, assume a system whereby two input channels feed a single response. Although participants may process these two channels independently, they only provide one response to these two channels. By contrast, the current study required independent responses for each channel (tracking left vs right gauge), where a response can only relate to a specific input channel. This is a subtle distinction, but it implies the two channels are not truly redundant, but rather more akin to a dual-task paradigm. The $C_{STST}(t)$ form of the capacity coefficient is appropriate for this paradigm, but the resilience function is best suited for a redundant task, so its application here is limited at best.

It may be true, as Navon and Gopher suggested, that there is no "standard" secondary task due to the effect of modality on dual-task interference (Navon & Gopher, 1979), with researchers instead choosing minimally invasive secondary tasks for their chosen primary task. It is prudent for researchers to endeavour to minimise unwanted interference between tasks and to use a minimally-invasive secondary task, but our findings suggest the modality and salience of a simple stimulus may not have an appreciable impact on primary task performance. Our findings in relation to the primary task suggest SFT may be applicable to more real-world tasks than have been used in previous studies. Either the use of a continuous, temporally-sensitive data set or a transformed pseudo-RT data set may allow more sophisticated analysis of real-world tasks such as driving or flight simulations than has been achieved before. Future studies are required to apply other UCIP models to these data sets—the current study used the

$C_{STST}(t)$ model, but tasks that compare data to $C_{OR}(t)$ and $C_{AND}(t)$ benchmarks would extend the current study's findings and further scrutinise the methods presented here.

Conclusion

Cognitive models were applied to both a continuous primary tracking task and a software-based DRT to assess the differential impact of an auditory and visual DRT stimulus on user workload. The analysis of primary tracking task performance utilised SFT in a novel application, subjecting continuous tracking data and transformed pseudo-RT distributions to capacity analysis. Both measures agreed in their finding of limited capacity. A shifted-Wald model was applied to DRT data, which identified a marked decrease in the rate of evidence accumulation as task load increased, but did not find any difference in rate across DRT modality, and only small differences in response threshold and non-decision time. These findings suggest multi-modal multi-tasking had limited advantage over single-mode multi-tasking in this simple task, and that DRT modality had little effect on primary tracking task performance. Our findings also present novel methods for applying cognitive models to more real-world data sets in the future.

References

Blaha, L. (2011). *A dynamic hebbian-style model of configural learning* (Doctoral dissertation). Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2019-10-19. Retrieved from https://search-proquest-com.ezproxy.newcastle.edu.au/docview/1500409686?accountid=10499

Blaha, L. & Houpt, J. W. (2015). An extension of workload capacity space for systems with more than two channels. *Journal of Mathematical Psychology*, *66*. doi:10.1016/j.jmp.2015.01.002

Bonnel, A.-M. & Hafter, E. R. (1998). Divided attention between simultaneous auditory and visual signals. *Perception and Psychophysics*, *60*(2), 179–190. doi:10.3758/BF03206027

Brown, S. D. & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. doi:https://doi.org/10.1016/j.cogpsych.2007.12.002

Castro, S., Strayer, D. L., Matzke, D. & Heathcote, A. (2019). Cognitive workload measurement and modeling under divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(6), 826–839. doi:10.1037/xhp0000638

Causse, M., Imbert, J. P., Giraudet, L., Jouffrais, C. & Tremblay, S. (2016). The role of cognitive and perceptual loads in inattentional deafness. *Frontiers in Human Neuroscience*, *10*. doi:10.3389/fnhum.2016.00344

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Fific, M. & Little, D. R. (2017). Stretching mental processes: An overview of and guide for sft applications. In D. R. Little, N. Altieri, M. Fific & C.-T. Yang (Eds.), *Systems factorial technology: A theory driven methodology for the identification of perceptual and cognitive mechanisms*. Elsevier.

Gibney, K. D., Aligbe, E., Eggleston, B. A., Nunes, S. R., Kerkhoff, W. G., Dean, C. L. & Kwakye, L. D. (2017). Visual distractors disrupt audiovisual integration regardless of stimulus complexity. *Frontiers in Integrative Neuroscience*, *11*. doi:10.3389/fnint.2017.00001

Green, D. M. & Gierke, S. M. V. (1984). Visual and auditory choice reaction times. *Acta Psychologica*, *55*(3), 231–247. doi:https://doi.org/10.1016/0001-6918(84)90043-X

Haapalainen, E., Kim, S., Forlizzi, J. F. & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th acm international conference on ubiquitous computing* (pp. 301–310). ACM. doi:10.1145/1864349.1864395

Heath, R. (2000). *Nonlinear dynamics: Techniques and applications in psychology*. Mahwah, NJ: Lawrence Erlbaum & Associates.

Heathcote, A. (2004). Fitting wald and ex-wald distributions to response time data: An example using functions for the s-plus package. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 678–694. Retrieved from https://search-proquest-com.ezproxy.newcastle.edu.au/docview/204305525?accountid=10499

Heathcote, A., Coleman, J. R., Eidels, A., Watson, J. M., Houpt, J. & Strayer, D. L. (2015). Working memory's workload capacity. *Memory and Cognition*, *43*(7), 973–989. doi:10.3758/s13421-015-0526-2

Houpt, J. W., Blaha, L. M., McIntire, J. P., Havig, P. R. & Townsend, J. T. (2014). Systems factorial technology with r. *Behavior Research Methods*, *46*(2), 307–330. doi:10.3758/s13428-013-0377-3

Houpt, J. W. & Little, D. R. (2017). Statistical analyses of the resilience function. *Behavior Research Methods*, *49*(4), 1261–1277. doi:10.3758/s13428-016-0784-3

Houpt, J. W. & Townsend, J. T. (2012). Statistical measures for workload capacity analysis. *Journal of mathematical psychology*, *56*(5), 341–355. doi:10.1016/j.jmp.2012.05.004

Hsieh, L., Seaman, S. R. & Young, R. (2015). A surrogate test for cognitive demand: Tactile detection response task (tdrt). *SAE 2015 World Congress and Exhibition*, *2015-April*(April). doi:10.4271/2015-01-1385

Innes, R., Howard, Z., Eidels, A. & Brown, S. (2018). *Cognitive workload measure and analysis*. University of Newcastle.

International Organization for Standardization. (2016). Road vehicles. transport information and control systems. detection-response task (drt) for assessing attentional effects of cognitive load in driving. https://www.iso.org/standard/59887.html. (ISO Standard No. 17488:2016).

Jeffreys, H. (1961). Theory of probability (3rd edt.) oxford university press. *MR0187257*.

Jones, M. & Dzhafarov, E. (2014). Unfalsifiability and mutual translatability of major modeling schemes for choice reaction time. *Psychological Review*, *121*(1), 1–32. doi:10.1037/a0034190

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. London: Academic Press.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(3), 451–468. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-0029314974&partnerID=40&md5=5ea5f4da0ec32bc392420fc491b97f73

Lavie, N., Hirst, A., de Fockert, J. W. & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General*, *133*(3), 339–354. doi:10.1037/0096-3445.133.3.339

Lee, M. D. & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Leite, F. P. & Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple-alternative decisions. *Attention, Perception, & Psychophysics*, *72*(1), 246–273. doi:10.3758/APP.72.1.246

Little, D. R., Eidels, A., Fific, M. & Wang, T. (2015). Understanding the influence of distractors on workload capacity. *Journal of Mathematical Psychology*, *68-69*, 25–36. doi:10.1016/j.jmp.2015.08.005

Little, D. R., Eidels, A., Fific, M. & Wang, T. (2018). How do information processing systems deal with conflicting information? differential predictions for serial, parallel, and coactive models. *Computational Brain & Behavior*, *1*(1), 1–21. doi:10.1007/s42113-018-0001-9

Navon, D. & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, *3*, 214–255. doi:10.1037/0033-295X.86.3.214

Palada, H., Neal, A., Strayer, D. L., Ballard, T. & Heathcote, A. (2019). Using response time modeling to understand the sources of dual-task interference in a dynamic environment. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(10), 1331–1345. doi:10.1037/xhp0000672

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R. & Strayer, D. L. (2014). Modeling simple driving tasks with a one-boundary diffusion model. *Psychonomic Bulletin & Review*, *21*(3), 577–589. doi:10.3758/s13423-013-0541-x

Ratcliff, R. & Van Dongen, H. P. A. (2011). Diffusion model for one-choice reaction-time tasks and the cognitive effects of sleep deprivation. *Proceedings of the National Academy of Sciences*, *108*(27), 11285–11290. doi:10.1073/pnas.1100483108. eprint: https://www.pnas.org/content/108/27/11285.full.pdf

Schmiedek, F., Oberauer, K., Wilhelm, O., Su, H.-M. & Wittman, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *138*(3), 414–429.

Schwarz, W. (1989). A new model to explain the redundant-signals effect. *Perception and Psychophysics*, *46*, 498–500. doi:10.3758/BF03210867

Steingroever, H., Wabersich, D. & Wagenmakers, E.-J. (2018). Modeling across-trial variability in the wald drift rate parameter.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R. & Hopman, R. J. (2017). The smartphone and the driver's cognitive workload: A comparison of apple, google, and microsoft's intelligent personal assistants. *Special Issue: Everyday Attention - Part I / L'attention au quotidian - partie I*, *71*(2), 93–110. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc13&NEWS=N&AN=2017-24966-002

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J. R., Medeiros-Ward, N. & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors*, *57*(8), 1300–1324. doi:10.1177/0018720815575149

Thorpe, A., Nesbitt, K. & Eidels, A. (2019). Assessing game interface workload and usability: A cognitive science perspective. In *Proceedings of the australasian computer science week multiconference* (44:1–44:8). ACSW 2019. Sydney, NSW, Australia: ACM. doi:10.1145/3290688.3290749

Tillman, G., Strayer, D. L., Eidels, A. & Heathcote, A. (2017). Modeling cognitive load effects of conversation between a passenger and driver. *Attention, Perception, & Psychophysics*, *79*(6), 1795–1803. doi:10.3758/s13414-017-1337-2

Townsend, J. T. & Eidels, A. (2011). Workload capacity spaces: A unified methodology for response time measures of efficiency as workload is varied. *Psychonomic Bulletin and Review*, *18*(4), 659–681. doi:10.3758/s13423-011-0106-9

Townsend, J. T. & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*(4), 321–359.

Treisman, A. & Davies, A. (1973). Divided attention to eye and ear. In S. Kornblum (Ed.), *Attention and performance iv*. New York: Academic Press.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and performance viii* (pp. 239–257). Hillsdale, NJ: Lawrence Erlbaum.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomic Science*, *3*, 159–177. doi:10.1080/14639220210123806

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455. PMID: 18689052. doi:10.1518/001872008X288394. eprint: https://doi.org/10.1518/001872008X288394

Appendices

Appendix A: Pseudo-RT Simulation

As the derivation of pseudo-RT distributions described in the manuscript is novel, it is necessary to assess the reliability of the technique in deriving pseudo-RTs that reflect the underlying cognitive processes. SFT makes specific predictions about the difference in low-and high-load performance in processes with different workload capacity. However, the pseudo-RT data we derived in the manuscript only described limited-capacity processes. To investigate whether the derivation technique is capable of producing pseudo-RTs that reflect the underlying cognitive processes, we simulated data that represented these different processes and applied the derivation technique with the prior knowledge of what result the derived pseudo-RTs should produce.

Data Simulation

To simulate the initial response sub-process, we generated RT data from gamma distributions whose scale parameters were varied to represent relative performance on the low- and high-load conditions. In the context of the $C_{STST}(t)$ model used in the current study, an unlimited-capacity process will produce equivalent performance on the two conditions; in a limited-capacity process, high-load performance will be slower than in the low-load condition, and in a super-capacity process, high-load performance will be faster than low-load performance. Table 3 shows the parameters used to generate simulated data for each of these processes. 10,000 data points were simulated for each distribution, with each value multiplied by 100 to better reflect human response time

data.

Table 3

Parameters used to generate gamma distributions.

Simulated Model	Low-Load		High-Load	
	Scale	Shape	Scale	Shape
Limited-Capacity	4	1	7	1
Unlimited-Capacity	4	1	4	1
Super-Capacity	4	1	2	1

The second sub-process, the execution of the tracking task, was deterministic, as the needle could only be moved at a set rate. The time taken to complete this sub-process therefore depended on the time taken in the first sub-process, as this determined the distance between the target and needle at the beginning of the second sub-process, and the movement of the target during the second sub-process, as direction changes could change the distance the needle was required to move to finish the trial.

Fifteen blocks of empirical target movement data were used for the simulation exercise. When the target changed direction, a response time was sampled with replacement from the distributions described above. During this time, the simulated needle did not change direction. After the sampled time had passed, the simulated needle moved towards the target's position until it was within the boundaries of the target object. This process was repeated for all target time series. Pseudo-RT distributions were derived using the same process as described in the manuscript. The top panels of Figure 15 show the ECDFs generated from the parameters in Table 3, while the bottom panels of Figure 15 show the ECDFs of derived pseudo-RTs. The relationships between high- and low-load conditions are consistent, showing the derivation technique's ability to recover the underlying pattern of responses.

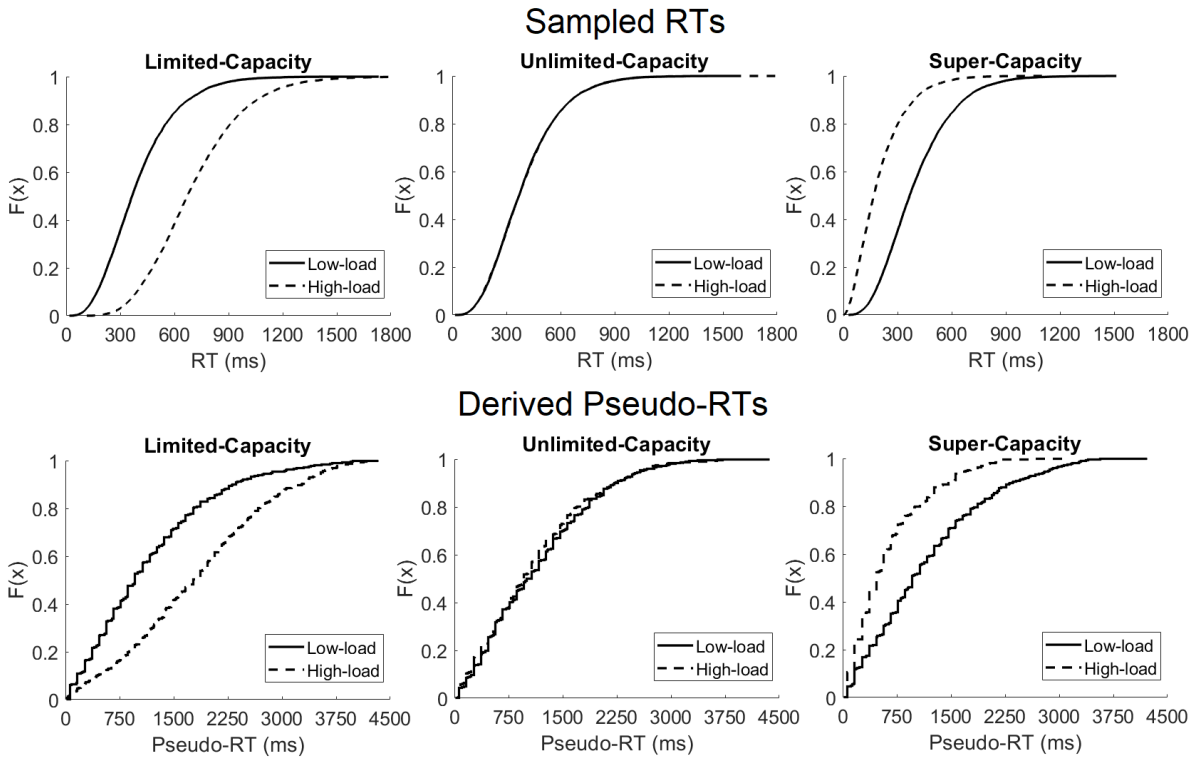

Figure 15. ECDFs for sampled RT distributions (top) and derived pseudo-RT distributions (bottom) for limited-capacity (left), unlimited-capacity (middle), and super-capacity (right) processes.

Capacity Analysis

Capacity coefficients were estimated for each of the simulated models. As Figure 16 shows, the capacity coefficients estimated from simulated data correspond to the workload capacity used in each sampling distribution. For example, the left panel shows a limited-capacity coefficient ($C(t) < 0$), the data for which was simulated using the limited-capacity sampling distributions depicted in the top-left panel of Figure 15. This indicates that the pseudo-RT derivation technique and subsequent capacity analysis correctly identify the underlying simulated cognitive processes.

Summary

We simulated tracking-task time series data based on the RT distributions predicted in limited-, unlimited- and super-capacity processes. We then applied the pseudo-RT derivation technique described in the manuscript and carried out a capacity

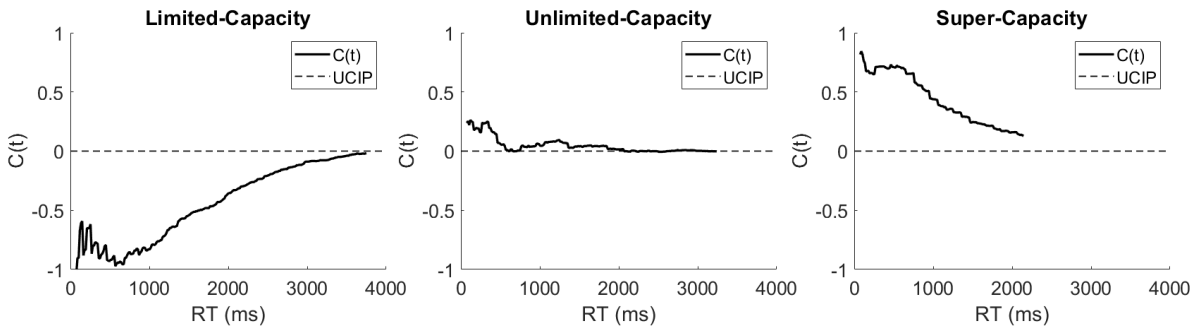

Figure 16. Capacity coefficients estimated from pseudo-RTs for limited-capacity (left), unlimited-capacity (middle), and super-capacity (right) processes.

analysis. In each case, our analysis correctly identified the processing efficiency of the underlying cognitive processes. This suggests our technique is capable of discriminating between processes with different workload capacity.

Appendix B: Capacity Coefficient Estimation with Smaller Sample Sizes

As Figure 14 illustrates, estimates of workload capacity changed when the threshold used to calculate pseudo-RTs was changed. This may not be due to truncating the distributions as discussed above, but rather due to the estimates themselves becoming unreliable due to smaller sample sizes when the threshold was more lenient. To address whether sample sizes changed with different threshold values, we calculated the average number of observations per condition per participant for each threshold in Figure 14. On average, a threshold of 0.2°produced 654 pseudo-RTs, 0.3°produced 567 pseudo-RTs, 0.4°produced 480 pseudo-RTs, 0.5°produced 393 pseudo-RTs, 0.6°produced 305 pseudo-RTs, 0.7°produced 229 pseudo-RTs, 0.8°produced 168 pseudo-RTs, 0.9°produced 125 pseudo-RTs, and 1.0°produced 95 pseudo-RTs per condition. This trend illustrates the issue raised above—as the threshold became more lenient, the number of observations decreased. The potential effect of this decreasing number of observations on the outcome of a capacity analysis was assessed by generating data sets based on the sample sizes listed above and comparing the accuracy of capacity estimates from each of these data sets.

Data Resampling

An initial large data set was generated by resampling with replacement from simulated pseudo-RT distributions generated using the same method outlined in Appendix A. The resulting distributions were based on an unlimited-capacity process, with 767 observations in the low-load distribution, and 774 observations in the high-load distribution. We resampled with replacement from these distributions to generate nine smaller data sets—the number of observations resampled in each of these data sets was equal to the average number of observations for each threshold listed above, so the sizes of these data sets decreased in the same way our empirical data did under more lenient threshold values. Capacity analysis was then carried out on each data set. Bootstrapped resampling was used to calculate the average capacity estimate from each data sets—data was resampled from each data set and capacity coefficients estimated 1,000 times and these coefficients were averaged to generate a single averaged capacity coefficient. The resulting nine averaged capacity coefficients are presented in Figure 17.

Estimates of capacity were stable across data sets, with each estimate correctly estimating unlimited-capacity for the majority of the time series. The data set with 125 data points (bottom-centre in Figure 17 over-estimated capacity in the first 1,000ms of the time series, while several other data sets over-estimated capacity for very fast RTs.

Summary

Increasing the threshold of acceptable performance resulted in fewer observations in empirical pseudo-RT distributions. To assess whether this caused capacity estimates to change, we generated data sets with smaller sample sizes by resampling from a single large data set and estimated capacity coefficients from these smaller sets. We found that capacity estimates were relatively stable in their correct estimation of unlimited-capacity, though there were some cases of over-estimation for fast RTs. This suggests that accurate capacity coefficients may be estimated from smaller data sets.

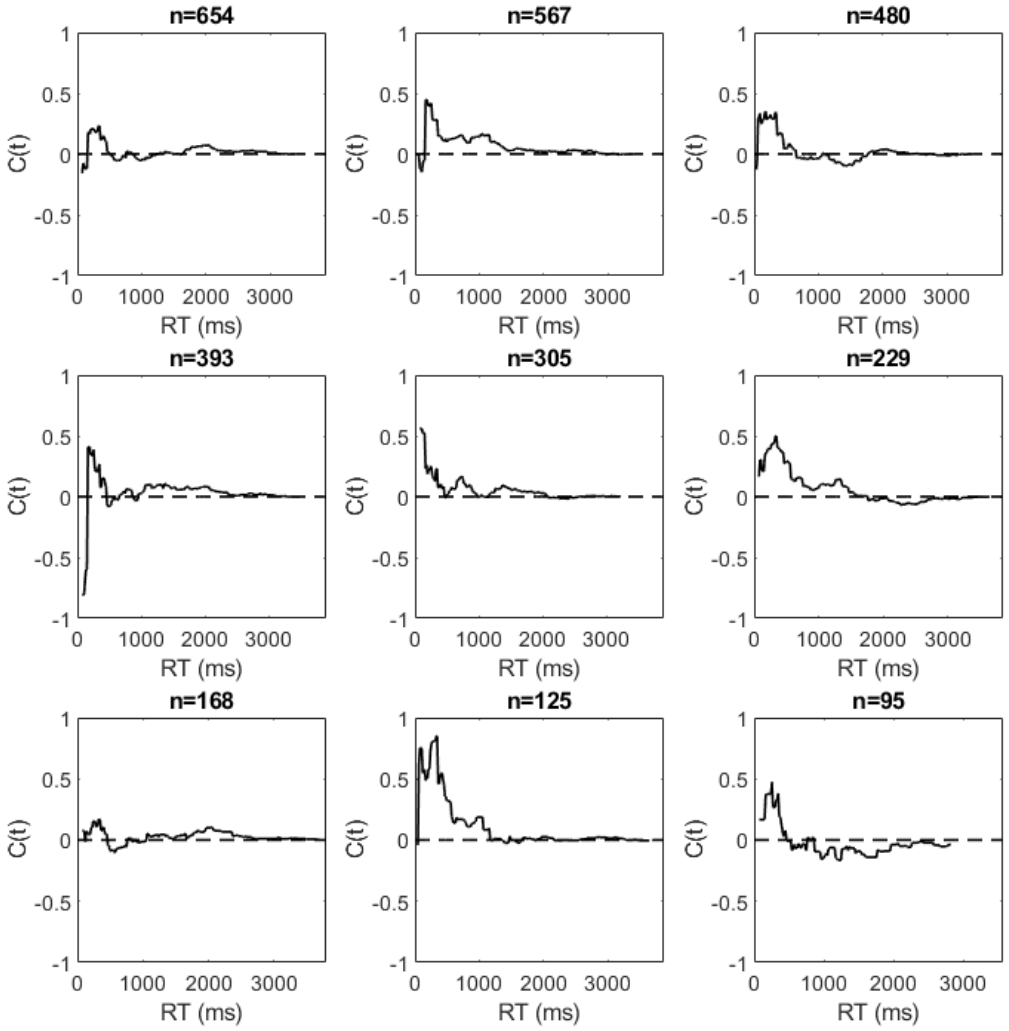

Figure 17. Capacity coefficients estimated from resampled pseudo-RT distributions with different sample sizes, listed above each graph. Solid lines are capacity coefficients; dashed lines are UCIP benchmarks.