2015

Working Memory's Workload Capacity

Andrew Heathcote
James R. Coleman, *University of Utah*
Ami Eidels
Jason M. Watson, *University of Utah*
Joseph W. Houpt, *Wright State University - Main Campus*, et al.

Accepted, 28/April/2015 Memory & Cognition
The final publication is available at Springer via http://dx.doi.org/
DOI: 10.3758/s13421-015-0526-2

Working Memory's Workload Capacity

Andrew Heathcote[1] (andrew.heathcote@newcastle.edu.au)
James R. Coleman[2] (james.coleman@psych.utah.edu)
Ami Eidels[1] (ami.eidels@newcastle.edu.au)
Jason M. Watson[2] (jason.watson@psych.utah.edu)
Joseph Houpt[3] (joseph.houpt@wright.edu)
David L. Strayer[2] (david.strayer@utah.edu)

[1]School of Psychology, University of Newcastle, Callaghan, NSW, 2308, Australia
[2]Department of Psychology, University of Utah, Salt Lake City, UT 84112, USA
[3]Department of Psychology, Wright State University, Dayton, OH 45435, USA

Corresponding Author: Andrew Heathcote 61-2-487979794

Keywords: Working Memory; n-back task; Systems Factorial Technology; Workload Capacity; Multitasking; Proactive Interference.

Abstract

We examined the role of dual-task interference in working memory using a novel dual 2-back task that requires a redundant-target response (i.e., that neither the auditory nor visual stimulus occurred two back vs. one or both occurred two back) on every trial. Comparisons with performance on single 2-back trials (i.e., with only auditory or only visual stimuli) showed dual-task demands reduced both speed and accuracy. Our task design enabled a novel application of Townsend and Nozawa's (1995) workload-capacity measure, which revealed that the decrement in dual 2-back performance was mediated by sharing of a limited amount of processing capacity. Relative to most other single and dual n-back tasks, performance measures for our task were more reliable, due to the use of a small stimulus set that induced a high and constant level of proactive interference. For a version of our dual 2-back task that minimized response bias, accuracy was also more strongly correlated with a complex span than has been found for most other single and dual n-back tasks.

In working memory tasks, participants are required to actively maintain information and to also manipulate that information or other information. Hence, these tasks are sensitive not only to limits in storage capacity (Cowan, 2001; Morey & Cowan, 2004) but also to limits in the capacity to perform two or more tasks at the same time, each possibly interfering with the other. There are at least two types of interference to consider in working memory tasks, *dual-task interference* (Kahneman, 1973; Wickens, 1980) between maintenance and manipulation operations within a trial, and *proactive interference* (Keppel & Underwood, 1962) arising between trials. Identifying and comparing these two kinds of interference is the prime motivation for the current study. To do so, we used a dual 2-back task developed by Heathcote et al., (2014) that is analogous to the redundant-target task used by Townsend and Nozawa (1995) to measure a "workload capacity" coefficient, which provides a rigorous measure of dual-task interference.

In Heathcote et al.'s (2014) dual 2-back task participants must indicate if either of two attributes of the current stimulus appeared in a stimulus occurring two trials before. For instance, given the sequence for the first attribute A-B-A-A-B…, the third item repeats the item that appeared two trials back (i.e., the first item), whereas the forth and fifth items do not repeat their 2-back predecessors. One set of attributes is auditory and the other visual. Suppose the second sequence has attributes X-Y-X-Y-Y…, both the third and forth items are the same as their 2-back predecessors, whereas the fifth is not. In a dual 2-back task the observer must monitor both sequences and respond affirmatively if an item in *either* fulfils the 2-back rule (e.g., both third and fourth items in the example), and otherwise respond negatively (e.g., the fifth item in the example). Tasks where responses can be based on either one or another stimulus attribute have been described as *redundant-target tasks*.

Comparisons between performance in redundant-target tasks and single-target tasks (i.e., tasks in which targets are defined in terms of only one stimulus attribute) have been used extensively to measure workload capacity in perceptual paradigms (e.g., Altieri & Townsend, 2011; Donkin, Little & Houpt, in press; Donnelly, Cornes & Menneer, 2012; Eidels, Townsend & Algom, 2010; Eidels, Townsend & Pomerantz, 2008; Fitousi & Wenger, 2011; Houpt, Townsend & Donkin, 2014; Ingvalson & Wenger, 2005; Johnson, Blaha, Houpt & Townsend, 2010; Neufeld, Townsend & Jetté, 2007; Von Der Heide, Wenger, Gilmore & Elbich, 2011; Wenger & Gibson, 2004; Wenger & Townsend, 2006; Zehetleitner, Krummenacher & Müller, 2009). Workload capacity is a quantity required to perform information processing, with reduced capacity leading to slower processing. Workload-capacity limitations can slow responding when more than one process – called a *channel* in the perceptual context – must perform work (i.e., process information), because the channels must share the capacity available to perform that work (for theory, overviews and estimation methods see Burns, Houpt, Townsend & Endres, 2013; Houpt, Blaha, McIntire, Havig & Townsend, 2014; Houpt & Townsend, 2012; Townsend & Eidels, 2011; Townsend & Honey, 2007; Townsend & Wenger, 2004). We exploit the redundant-target nature of Heathcote et al.'s (2014) task to use it as a building block in measuring the workload capacity of working memory.

In the next section, we describe Heathcote et al.'s (2014) dual 2-back task in detail, providing background on its relationship to various tasks used to measure working memory. We then report and analyse an experiment that augments Heathcote et al.'s dual 2-back task with single 2-back tasks, enabling a workload-capacity analysis. Comparing information-processing latencies and accuracy with two vs. one source of information is the cornerstone of the workload-capacity analysis. A formal

definition of capacity is given later, but in brief, if processing efficiency with two sources of information is as good as predicted by the summed efficiency of processing each source alone, capacity is said to be unlimited. If monitoring two streams takes a toll on performance compared to one stream, capacity is limited. We report this analysis and its outcomes in a subsequent section.

Tasks Measuring Working Memory

Complex Span Tasks. One class of working-memory tasks focuses on the number of stored items that participants are able to report, typically averaging accuracy over a range of storage loads. This class derives from the simple-span tasks (e.g., repeat back a set of random digits in the order they were presented) and adds a requirement to manipulate information. The manipulated information can be either relevant to the recall task, such as requiring report in a backward or alphabetic order, or irrelevant to the task, such as in complex span tasks. For example, in one type of complex span task, operation-span (Engle, 2002; Turner & Engle, 1989; Unsworth, Heitz, Schrock & Engle, 2005), decisions about the veracity of mathematical equalities periodically interrupt study of items for later recall. Correlations between complex-span and measures of executive control have led to proposals that working memory depends on the effectiveness of attention control as well as storage capacity (e.g., Burgess, Gray, Conway & Braver, 2011).

N-back Tasks. Another class of tasks uses response time (RT) and/or accuracy for choices to infer storage capacity. The n-back task is popular in cognitive neuroscience, as it is suitable for event-related physical measurement, in investigations of both working memory and attention control (Owen, McMillan, Laird & Bullmore, 2005). Participants are presented with a series of stimuli with targets defined as occurring *n* trials previously. In some paradigms only target responses are

required, and in others responses are required for both targets and lures (i.e., items that occurred at some other value of *n*). Performance can be measured by averaging accuracy over a range of values of *n*, or by the value of *n* attained, where *n* is increased based on accurate performance (e.g., Jaeggi et al., 2008). In other cases *n* is fixed at a smaller value so accuracy is high and interest focuses on RT (e.g., Schmiedek, Li & Lindenberger, 2009).

As well as differing in response measures, complex span tasks differ from n-back tasks in that they require processing of information that does not need to be stored for later recall. Complex span and n-back tasks have been suggested to measure somewhat different aspects of working memory (Kane, Conway, Miura & Colflesh, 2007), although more recent research suggests the latent constructs derived from these two classes of task are difficult to distinguish (Schmiedek, Hildebrandt, Lövdén, Wilhelm & Lindenberger, 2009).

Heathcote et al. (2014) developed the "*Gatekeeper*" task, a modified version of the dual n-back verbal/spatial working-memory task that has been studied extensively by Jaeggi and colleagues (2003, 2007, 2008, 2010). Participants are presented with pairs of visual and auditory stimuli, with a target response required if a stimulus in either modality is a repeat from 2 trials previously and a non-target response required otherwise (see illustration in Figure 1). Stimuli are never immediately repeated, so participants cannot use easy familiarity-based strategies available in a 1-back task (McElree, 2001) based on the high availability of items held in the focus of attention (Oberauer, 2002). Because Gatekeeper is a 2-back task it needs only four items to be held in memory at any time, and so does not exceed the storage capacity limits typically ascribed to working memory (Cowan, 2001; Morey & Cowan, 2004).

Gatekeeper uses a set of only three different stimuli in each modality, so stimuli frequently swap roles as targets (i.e., stimuli occurring two trials back) and lures, maximizing proactive interference. In contrast to most other n-back tasks where strong proactive interference typically occurs on only a minority of trials (see Gray, Chabris & Braver, 2003), the small stimulus sets mean that proactive interference is high – and most importantly fairly constant – over trials, as stimuli that did not occur two trials back must have occurred three trials back. Heathcote et al. (2014) found that this constant level of interference (and hence less variability in interference than in other n-back tasks) led to highly reliable measurement even in a diverse online sample. Requiring a response on every trial further serves to induce proactive interference because there is a rapidly varying mapping in the response associated with each stimulus, minimising any benefits of practice-induced automaticity (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977).

Performance in the Gatekeeper task is also subject to dual-task interference, because binding processes (e.g., from stimulus representations to representations of one vs. two back positions and/or target vs. lure roles), and processes associated with stimulus encoding, must be performed in both modalities. Further interference occurs because responding in Gatekeeper differs from that required in most dual n-back tasks where separate responses are made to stimuli in each modality (e.g., Jaeggi et al., 2003, 2007, 2008, 2010a, 2010b). In the Gatekeeper task, a single response, potentially informed by both modalities, is required for each trial. Namely, participants combine the outcomes from two modalities into a single response using an OR rule: they respond affirmatively if the current visual item is the same as the visual item that appeared two trials back, or if the current auditory item is the same as the auditory item two-trials back, or if both conditions are met. Because only a single

response is made, single target trials – where one stimulus is a target (i.e., it occurred 2-back) and one is not (i.e., it occurred 3-back) – have added interference due to the conflicting individual stimulus-to-response associations. That is, a stimulus from a given modality can be associated with opposite responses on different trials depending on the context in which it occurs.

Heathcote et al. (2014) found that the high levels of interference from all of the sources just discussed result in accuracy that is typically well below ceiling even though the Gatekeeper task does not strongly tax storage capacity. As a result, accuracy is a useful dependent variable. As in other two-back tasks requiring responses to both targets and lures (e.g., Schmiedek et al., 2009), they also found that RT is a useful dependent measure.

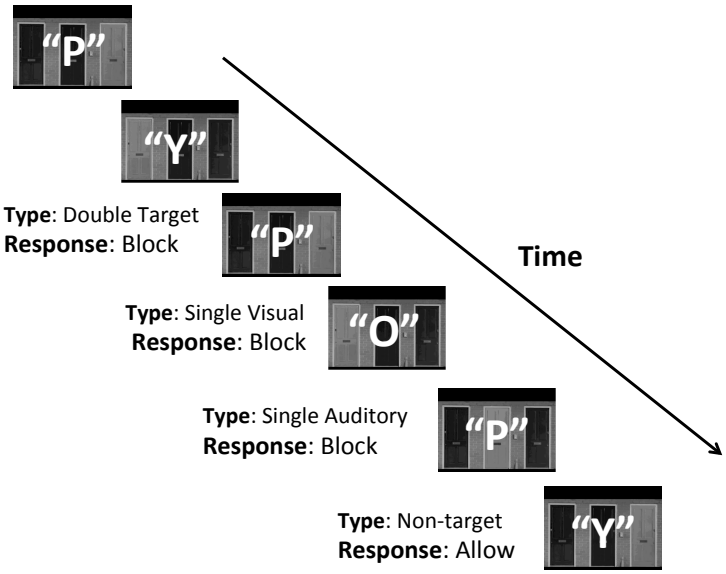

Figure 1. Example of the first 6 trials in a dual-task Gatekeeper block. White letters indicate auditory stimuli (passwords), and visual targets are the light-grey doors. Visual stimuli were presented in colour, with light-grey regions in red and dark regions in black. No response was required for the first two trials. For each trial thereafter, the trial type and correct response are indicated. Single blocks present only visual or only auditory information. For the auditory case, the correct response sequence would be *Block-Allow-Block-Allow*. For the visual case, the correct response sequence would be *Block-Block-Allow-Allow*.

Figure 1 illustrates a sequence of trials in the Gatekeeper task. We told participants they were in training to become a nightclub doorperson, and that their

task was to allow in only cool patrons. Stimuli are both visuo-spatial (an image of three doors, where door location is the critical attribute) and auditory (a spoken letter). A patron tries to gain access through one of the three doors, as indicated by that door being highlighted, and by saying one of three password letters, "P", "Y" or "O". These specific letters were selected to minimize acoustic confusion (Conrad, 1964) so errors tap memory and not perceptual errors. We also told participants that no patron is so uncool as to use the door or password from the last trial but that they do slip up by using the door and/or password from two trials back, in which case they must be blocked. Decision speed was emphasised by telling participants that only Gatekeepers who can decide both quickly and accurately make the grade and will be employed by the nightclub.

In order to directly examine the impact of the dual-task requirement we used an elaborated version of Heathcote et al.'s (2014) task with both single-task (either all auditory or all visuo-spatial) as well as dual-task blocks. In their experiment examining practice effects, Jaeggi et al. (2010a) found better initial performance in a single than dual n-back task, and a faster improvement with practice. Hence, it seems likely that we will observe better performance in the single than dual blocks. A nice feature of the design of Gatekeeper is that, because only a single response is required in our dual-task blocks, a comparison of single-target trials in single and dual-task blocks enables us to measure dual-task load with the number of targets and responses controlled. Workload-capacity measurement provides an even more refined way of quantifying dual-task interference between the two processes ("channels") that match memory representations of 2-back auditory and visual stimuli, respectively, to current auditory and visual stimuli. In particular, the workload-capacity measure is applied to memory processes, by comparing performance across various levels of processing

load. Namely, we can compare performance between double-target trials in dual-task blocks (i.e., where participants monitor both auditory and visual streams) and single-target trials in single-task blocks (where they are presented with, and monitor, only one modality at a time).

Single blocks in the Gatekeeper task have a 50% target ("block") rate and double blocks a 75% target rate when stimulus types occur with equal probability. Heathcote et al. (2014) found that in this case participants developed expectations about the probability of a target in dual blocks that biased their responding towards the more common 'target-present' response. As bias differences might affect the single vs. dual block comparison, we used dual blocks with 50% targets for some participants (by decreasing the probability of selecting the visual and auditory stimuli that occurred two-back) as well as the standard 75% target blocks for other participants. In both cases, targets occurred on 50% of trials in single blocks. The 50% dual condition necessarily introduces some predictability (i.e., the next stimulus was more likely to have occurred three back) that might also lead to participant expectations that affect performance. By running both 50% and 75% conditions we sought to determine whether difference in expectations lead to differences in performance that affected workload-capacity measurement.

In order to measure the convergent and divergent validity of measures obtained in the Gatekeeper task, participants also completed a complex span task, Heitz, Schrock and Engle's (2005) operation-span (OSPAN) task. Complex-span measures commonly have a relatively low correlation with performance in standard n-back and dual n-back tasks (see Jaeggi et al., 2010b, for a summary). Further, Kane, Conway, Miura and Colflesh (2007) found that higher proactive interference in an n-back task – as is the case for the Gatekeeper task – did not increase its correlation

with OSPAN. We also measured only a fixed level of *n=2* in the Gatekeeper task compared to a range of 3-7 in the OSPAN task, so a strong correlation would be unexpected. However, the dual version of the n-back task does make it more like complex span tasks with their dual-task requirements, so it is possible that the high reliability of Gatekeeper found by Heathcote et al. (2014), particularly for mean RT and accuracy measures, might lead to a higher correlation. We report results from the Spearman-Brown split-half procedure in order to quantify reliability for the present experiment.

In summary, the purpose of the current research is to provide a rigorous analysis of some relatively novel aspects of the concept of working memory capacity, and to investigate their relationship to individual differences in human information processing. To this end we used Heathcote et al.'s (2014) "Gatekeeper" redundant-target dual n-back task, which features both a strong and homogenous level of proactive interference as well as multi-tasking demands. Our analysis first establishes the reliability of the performance metrics obtained with Gatekeeper, and then contrasts them with another working memory capacity measure provided by the OSPAN task. Finally, Systems Factorial Theory (Townsend & Nozawa, 1995) is used to formally model individual differences in workload capacity, identifying when performance is characterized by limited or fixed capacity, and when performance is consistent with unlimited capacity.

Method

Participants

University of Utah undergraduates (372 total, 224 female, mean age 23 years) were tested in groups of up to five, and received course credit for participation. They provided informed consent and were randomly assigned to the dual block 50% target

and 75% target groups, then performed the OSPAN task followed by the Gatekeeper task. Data from 61 participants was lost due to software errors, leaving a final sample of 311, with 147 in the 50% group and 164 in the 75% group.

Procedure

OSPAN task. The task presented simple math problems requiring a "true" or "false" response (e.g. (8/2)+2=12…"False"). Following each math problem, a letter was presented for later recall. Participants first completed three practice blocks, a simple letter span task and then a block requiring speeded solution of math problems. Solution times were used to set the time allowed for responding to math problems in later blocks (mean + 2.5 standard deviations). The third practice block consisted of three sets of two trials that combined math problems and letter recall. Participants then completed three sets each of 3 to 7 math and letter pairs (75 each in total) in a random order, and were asked to perform immediate recall of the letters in the order in which they were presented. Stimuli were presented on a computer screen and responses were made with a computer mouse by clicking a true or false text box when responding to the math operations. Letter recall required participants to click the correct letters in the correct order among a 3x4 matrix of letters. The OSPAN score was the total number of letters accurately recalled in the correct order out of 75.

Gatekeeper task. Participants completed the task through a Firefox browser with auditory stimuli presented via headphones. A trial terminated with the response or after 2.5sec if no response was given, and a new trial would begin after a 1sec interval. As illustrated in Figure 1, in dual task blocks at the start of each trial one of the three doors turned red, and one of the letters "Y", "P" or "O" were spoken through the computer speakers in a female voice. In single task blocks, only the auditory or only visual stimuli were presented. Responses were made via the keyboard using the

"z" and "/" keys to allow or block entry, with the mapping alternated for each new participant. Participants were told that the initial two entries on each block of trials were the manager and the barman, who were allowed entry. Thus, they did not have to respond, but still had to remember the doors and passwords used.

Participants performed four practice blocks, starting with two 12-trial single-task blocks, one visual and one auditory. Feedback was provided at the top of the screen indicating whether responses were correct or incorrect. They then performed two practice dual-task blocks of 27 trials, the first with feedback and the second without. Practice was followed by 16 experimental blocks, each with 27 trials and no feedback. Participants were required to press the space key to move on to the next block, but could only do so after a mandatory 1-minute break between blocks. At the conclusion, participants were given feedback about their overall performance.

The 16 experimental blocks were divided into 8 dual-task blocks and 4 visual and 4 auditory single-task blocks. The order of dual and single blocks was chosen randomly over participants, as was the order of visual and auditory single blocks. Auditory and visual stimuli were selected randomly and independently with the constraint that they never repeated immediately. In 75% target dual-task blocks, available stimuli (i.e., those that did not occur on the last trial) were chosen with equal probability, so no-target, visual-target-only, auditory-target-only and double-target trials occurred on average with equal frequency. In 50% target dual-blocks, available stimuli were randomly selected subject to the constraint that double, single visual and single auditory stimuli each occurred on $16^2/_3\%$ of trials and no-target stimuli occurred on the remaining 50% of trials.

Results

Overview

Results are presented for both individual and group-level Gatekeeper and OSPAN performance. For Gatekeeper, performance in single-target conditions is contrasted with performance in dual-target conditions, and the psychometric reliability of the different measures examined. Correlations between the different parameters of Gatekeeper are computed and compared with correlations involving measures obtained on the OSPAN task. Finally, the data from single-target and dual-target conditions is modelled using Systems Factorial Technology to analyse individual differences in workload capacity. Taken together, these analyses provide a rigorous assessment of Gatekeeper as a method for understanding workload capacity in working memory.

Bayes Factor Analysis

We used the *BayesFactor* package for the R statistical language (Morey & Rouder, 2012) to perform Bayes Factor (BF) based tests of correlations, t-tests, and ANOVA using Rouder, Morey, Speckman and Province's (2012) default prior method. Bayes Factors are not subject to the bias in traditional frequentist approaches using a fixed-criterion *p* value of being increasingly likely to declare significant effects as sample size increases (Raftery, 1995, Table 9). In contrast to null-hypothesis testing (see Morey & Rouder, 2011; Wagenmakers, 2007), they can also provide evidence for null effects relative to appropriately scaled priors. *BayesFactor* uses priors on effect sizes, and we found that our inferences were insensitive to a reasonable range of assumptions about the plausible range of effect sizes.

For ANOVA analyses we fit all possible hierarchical models, that is, all additive combinations of main effects and interactions with the restriction that when

higher-order terms are included so are all of their lower-order constituents, corresponding to a Type-II sums-of-squares approach in traditional ANOVA. We first report the best model, that is, the model with the strongest evidence as indicated by the largest BF relative to the intercept-only (grand mean) model. We then examine the strength of evidence against alternative models that either add or remove a term to the best model. To do so, we used BFs for the best model relative to the alternative model, which are necessarily greater than one.

For example, BF = 10 indicates the data increase support for the best model relative to the alternative model by a factor of 10. Jeffreys (1961, p.432) [1] described a factor of 10 or larger as indicating strong evidence, a factor from 3 to 10 as indicating positive evidence and a factor of 3 or less as providing equivocal evidence. Although we report the numerical values of BFs, as they have a natural interpretation in terms of support for hypotheses provided by the data, these classifications provide a useful approximate guide when summarizing results. For, even though a term is included in the best model it can be described as having only weak support if BF < 3. Similarly, the exclusion of a term from the best model only has weak support if BF < 3. In contrast, as the BF increases above three, there is increased support for including the term in the model (analogous to a term being significant in a frequentist analysis) or excluding the term from the model (i.e., support a null effect). We also provide posterior medians to illustrate effects and use 95% credible intervals (CIs, the 2.5th to 97.5th percentiles of the parameter's posterior distribution) to quantify uncertainty in these estimates.

[1] Kass and Raftery (1995) suggest a similar scheme, but with 3-20 labelled positive and 20-150 strong and greater than 150 very strong. They also discuss how a BF can be understood in terms of the relative abilities of models to predict observed data. It is important to note that labels can be misleading when strong prior evidence is present. For example, if model A is a-priori considered 100 times more likely than model B then a factor of 10 for model B vs. model A means model A is still remains 10 times more likely. In our application, we do not think there are any such strong prior beliefs that would substantially distort the conventional labelling.

Gatekeeper and OSPAN Accuracy and Exclusion Criteria

Participants with more than 10% non-responses on the Gatekeeper task (33) were removed. Single-target and double-target block accuracies for the remaining 311 participants are plotted in the left panel of Figure 2. There was strong evidence for greater accuracy in single (82%) than dual (74%) blocks (BF = 2.1×10^{22}, CI = 6.4% – 9.1% for the accuracy difference). Figure 2 shows that some participants responded at or below chance, indicating they did not understand or engage with the Gatekeeper task. These participants, defined by a score of less than 55% correct in either single or double blocks (66), were removed from further analyses. In the remaining 245 participants there was again strong evidence of greater accuracy in single (89%) than double (79%) blocks (BF = 1.9×10^{36}, CI = 8.7% – 11.2% for the accuracy difference).

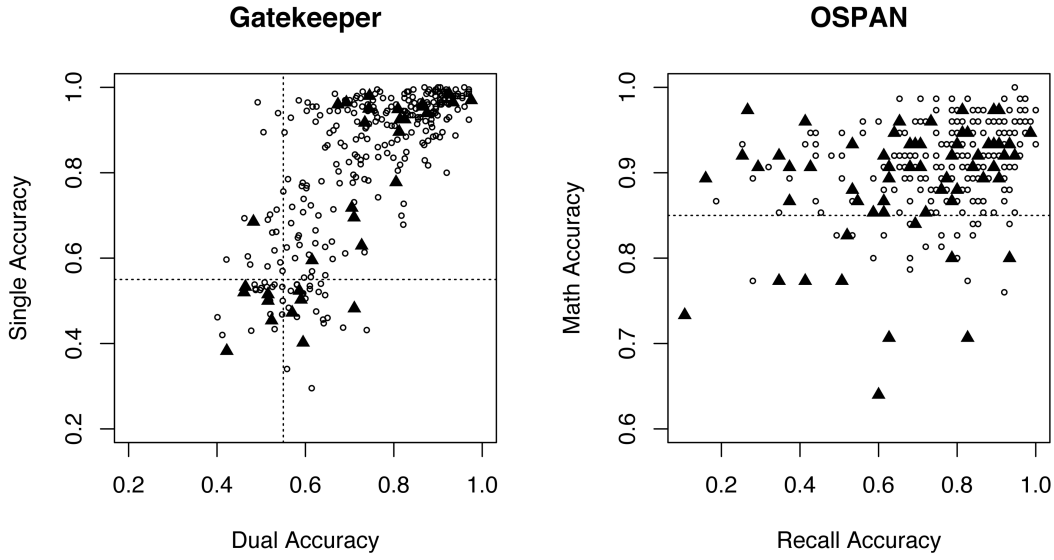

Figure 2. The left panel shows accuracy in the single-target and double-target blocks of the gatekeeper task. Circles represent participants with greater than the 85% cut-off for Math accuracy in the OSPAN task and triangles the excluded participants. Dotted lines represent the 55% accuracy cut-offs in the Gatekeeper task. The right panel shows accuracy in the OSPAN task. Circles represent participants with greater than the 55% cut-off for single and dual accuracy in the Gatekeeper task and triangles the excluded participants. The dotted line represents the Math accuracy cut-off in the OSPAN task.

The right panel of Figure 2 plots recall and math accuracy in the OSPAN task for the full sample, with participants failing the Gatekeeper accuracy cut-off plotted as triangles. Recall accuracy in the overall sample (76%) increased only slightly (to 78%) when participants failing the Gatekeeper cut-off were removed. Unsworth et al. (2005) recommended exclusion of OSPAN participants with less than 85% accuracy in the math task in case they were ignoring the math problems to boost recall. The left panel of Figure 2 plots as triangles Gatekeeper accuracy for the 32 participants (10% of the overall sample) with less than the 85% math-accuracy cut-off. It shows that failure of the OSPAN cut-off was not associated with failure of the Gatekeeper cut-off. When the 20 participants failing the OSPAN cut-off were removed from the 245 who passed the Gatekeeper cut-off, Gatekeeper accuracy was unchanged to the nearest percentage. We decided to retain the sample of 245 participants in all analyses except those directly involving OSPAN, where we used only the 225 participants passing both cut-offs.

Reliability

Table 1 displays Spearman-Brown split-half reliabilities for statistics derived from dual and single blocks in the Gatekeeper task for data from all n = 200 trials and randomly selected subsets of n = 100 and 50 trials. Reliabilities were averaged over 100 random subsets; with this number of subsets, the standard error of the mean estimate was negligible. In Table 1, we also quantify the reliability of response-choice (i.e., "block" vs. "allow entry") data both in terms of the overall accuracy (i.e., percentage of correct responses) using the normal, equal variance signal-detection theory measures (Stanislaw & Todorov, 1999).

Table 1. Average Spearman-Brown split-half reliabilities based on a design with *n* trials for: PC = overall percentage correct, MRT = overall mean RT. d' = signal detection sensitivity. Subscripts indicate statistics calculated based on dual-target (av), auditory (a) or visual (v) single-target trials (relative to non-target trials in the case of d'), non-target (no) trials (for dual blocks), and auditory non-target (an) and visual non-target (vn) trials (for single blocks).

Group		50%			75%		
n		200	100	50	200	100	50
Dual Blocks	PC	.97	.93	.90	.96	.93	.90
	d'_{av}	.78	.71	.66	.84	.81	.79
	d'_{a}	.86	.76	.67	.92	.84	.78
	d'_{v}	.85	.75	.67	.91	.83	.77
	MRT	.99	.97	.96	.99	.98	.96
	MRT_{av}	.99	.97	.96	.99	.98	.96
	MRT_{a}	.96	.93	.90	.96	.93	.90
	MRT_{v}	.73	.56	.46	.87	.77	.69
	MRT_{no}	.60	.53	.48	.68	.63	.60
Single Blocks	PC_{1}	.98	.96	.95	.99	.97	.96
	d'_{a1}	.76	.67	.92	.84	.78	.76
	d'_{v1}	.85	.75	.67	.91	.83	.77
	MRT_{1}	.98	.97	.95	.98	.96	.95
	MRT_{a1}	.96	.93	.90	.96	.93	.90
	MRT_{v1}	.73	.56	.46	.87	.77	.69
	MRT_{an1}	.60	.53	.48	.68	.63	.60
	MRT_{vn1}	.73	.58	.46	.80	.66	.56

OSPAN – Gatekeeper Correlations

Table 2. Correlations among OSPAN recall accuracy and Gatekeeper performance measures for participants with accuracy 85% for greater in the OSPAN math task. The upper triangle contains results for the 50% group and the lower triangle the 75% group. Correlation tests: single, double and triple "+" superscripts indicate 3<BF<10, 10<BF<100 and BF>100 (i.e., substantial, strong and very strong evidence that the correlation is non-zero), and a single "–" superscript indicates 0.1>BF>$^{1}/_{3}$ (i.e., substantial evidence that the correlation is zero). See Table 1 for definitions of all measures, except the workload-capacity measures (Cz, CzF and Cp), which are defined and discussed in the section "Working Memory's Workload Capacity".

	OSPAN	PC	PC_{1}	MRT	MRT_{1}	Cz	Czf	Cp
OSPAN		$.43^{+++}$	$.30^{++}$	$.01^{-}$	$-.16$	$-.09^{-}$	$-.10^{-}$	$-.21^{+}$
PC	$.21$		$.62^{+++}$	$.17^{-}$	$-.12^{-}$	$.05^{-}$	$-.03$	$.76^{+++}$
PC_{1}	$-.11^{-}$	$.62^{+++}$		$.27$	$-.17$	$-.39^{++}$	$-.34$	$.44^{+}$
MRT	$.02^{-}$	$.19^{++}$	$.30^{+++}$		$.51^{+++}$	$-.23^{+++}$	$-.26^{+++}$	$-.01$
MRT_{1}	$-.14^{-}$	$.02$	$-.07^{-}$	$.53^{+++}$		$.57^{+++}$	$.57^{+++}$	$-.11^{-}$
Cz	$-.16$	$-.02$	$-.44^{+++}$	$-.40^{+++}$	$.38$		$.97^{+++}$	$.11^{+++}$
Czf	$-.12$	$.05$	$-.36^{+++}$	$-.44^{+++}$	$.37^{++}$	$.95^{+++}$		$.14^{+++}$
Cp	$.07^{-}$	$.74^{+++}$	$.34^{+++}$	$.03^{-}$	$-.01$	$.18^{-}$	$.24^{-}$	

Table 2 displays correlations among OSPAN recall and selected Gatekeeper performance measures (principally those with higher reliabilities). These correlations

use only results from participants with 85% or greater accuracy in the OSPAN math task, and were calculated separately for 50% and 75% groups (100 and 125 participants respectively).

Accuracy and RT in Single- and Dual Task Blocks

Figures 3 and 4 display accuracy and mean RT results for the 50% and 75% groups broken down by the different 2 × 2 within-subject designs for single (visual vs. auditory × target present vs. absent) and dual (auditory target present vs. absent × visual target present vs. absent) blocks. We report three types of analyses including group as a between-subjects factor: separate analyses of the single and dual blocks, and an analysis across block types of single-target trials focusing on the effect of dual-task load. We examine response probabilities using signal-detection theory sensitivity and bias measures. Table 3 reports the best model selected in the ANOVA analyses. In all but one case a model that is simpler than the most complex ANOVA model is best.

Table 3. Bayes Factor ANOVA model selection with Bayes Factors relative to the best fitting (i.e., most complex) model for mean RT and signal-detection theory sensitivity (d') and bias (c) measures.

Measure	ANOVA	Selected Model	Bayes Factor
Mean RT	Double Blocks[1]	A + V + G + A × V	184
	Single Blocks[2]	T + M + G + T × G	275
	Single Trials[3]	B + M + G + B × M + B × G	54
d'	Double Blocks[4]	TM + G	6.3
	Single Blocks[2]	M	14
	Single Trials[3]	B + M + G + B × G	34
c	Double Blocks[5]	G	1
	Single Blocks[5]	G	57

[1] Factors: A = auditory and V = visual target vs. non-target, G = 75% vs. 50% group.
[2] Factors: T = target vs. non-target trial, M = visual vs. auditory modality, G = 75% vs. 50% group.
[3] Factors: B = single vs. double block, M = visual vs. auditory modality, G = 75% vs. 50% group.
[4] Factors: TM = target modality, visual vs. auditory vs. both, G = 75% vs. 50% group
[5] Factor: G = 75% vs. 50% group

Dual-Block Analysis. In mean correct RT there was equivocal evidence for slower performance in the 50% than 75% group (1139ms vs. 1079ms, BF = 1.7). The

main effects of slower non-target than target performance were similar for auditory and visual (114ms and 131ms respectively). There was strong evidence for an interaction between the auditory and visual target vs. non-target effects (BF = 3.1×10^{11}). As shown in the right panels of Figure 3, this was due to a larger slowing for auditory non-target vs. target when the visual stimulus was also a target (168ms) than when it was a non-target (60ms).

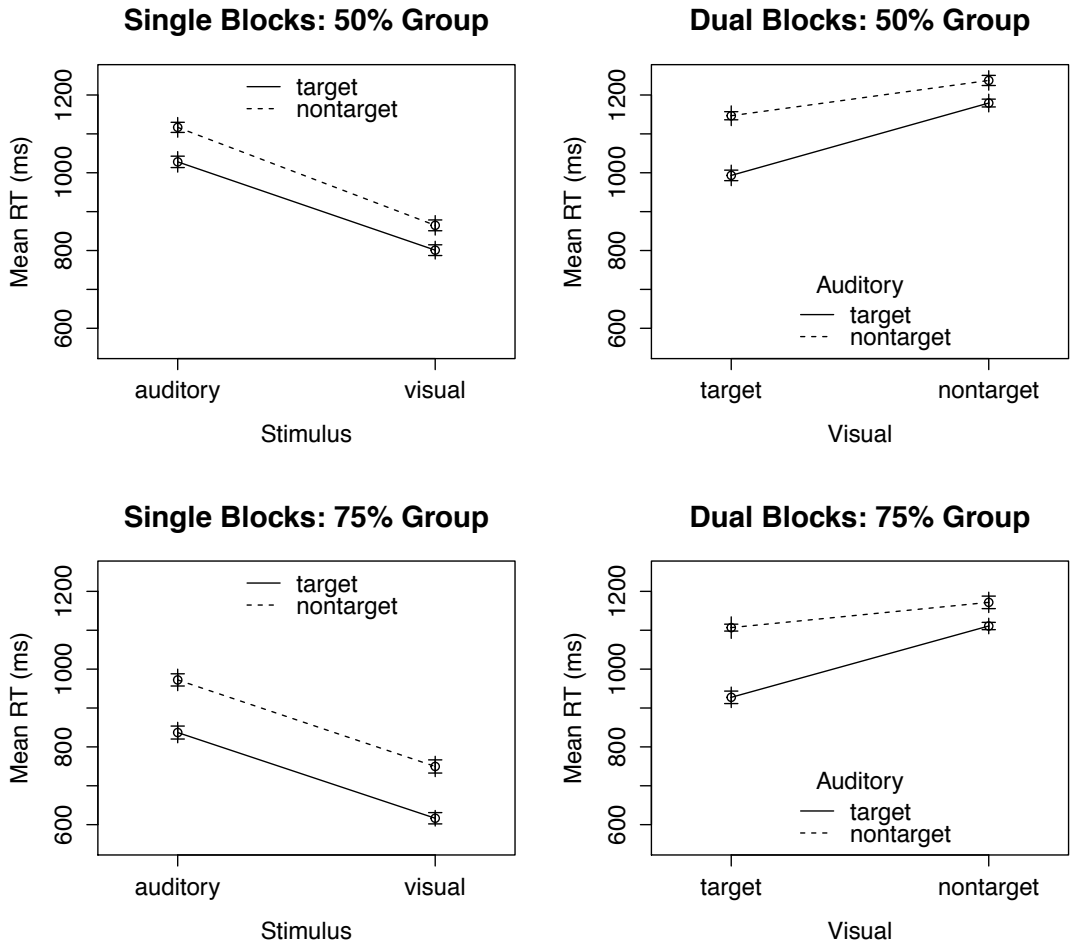

Figure 3. Mean correct RT with 95% within-subjects confidence intervals (Morey, 2008) defined by horizontal lines and individual 95% confidence intervals depicted by lines, as recommended by Baguley (2011).

Sensitivity (d') was larger for the 50% group than the 75% group (1.91 vs. 1.58, BF = 7.1) and differed between trial types (i.e., double vs. single, BF = 1.5×10^{55}), but these effects did not interact (BF = 1/6.3). A linear contrast on the trial-type main effect found positive evidence against a difference between visual-target

and auditory-target trials (1.49 vs. 1.44 respectively, BF = 1/7.8), and strong evidence for a difference between double-target (2.25) and the average of single-target trials (BF = 1.7×10^{68}), consistent with the interactions evident in the right-hand panels of Figure 4.

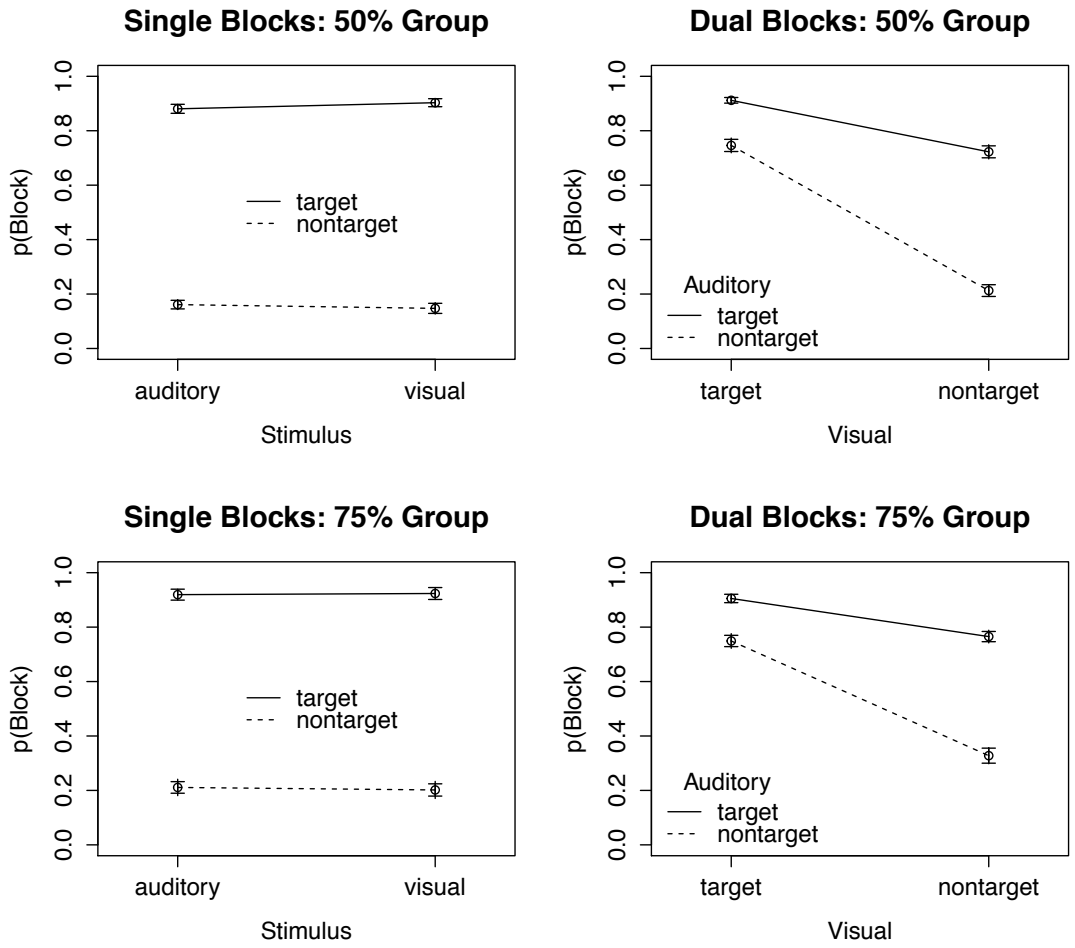

Figure 4. Average probability of detecting a target (responding "Block") with 95% within-subjects confidence intervals (Morey, 2008) defined by horizontal lines and individual 95% confidence intervals depicted by lines, as recommended by Baguley (2011).

There was positive evidence for unbiased responding in the 50% group (Signal Detection Theory's bias measure, c = -0.01, BF = 6.5, CI = -0.07 to 0.05) and strong evidence for target-biased responding in the 75% group (c = -0.23, BF = 4.1×10^{5}, CI = -0.3 to -0.15), and for the two being different (BF = 691 for the best model in Table 3, with a group difference relative to a model with no group difference).

Single-Block Analysis. For mean correct RT, auditory was slower than visual (979ms vs. 749ms, BF = 4.8×10^{111}). Non-target was also slower than target (918ms vs. 809ms) and the 50% group was slower than the 75% group (953ms vs. 794ms), with the two effects interacting because the slowing for non-targets was smaller in the 50% than 75% group (77ms vs. 134ms, BF = 32.3). The interaction is evident in the left-hand panels of Figure 3 as a smaller gap between solid and dashed lines for the 50% group than the 75% group.

As shown in Table 3, there was strong evidence for greater sensitivity in the visual than auditory modality (*d'* = 2.83 vs. 2.63) and positive evidence against a main effect of group (BF = 1/3.1). Table 3 also shows there was strong evidence for a difference in bias between the 75% and 50% group (*c* = -0.33 vs. -0.11), and in both cases there was strong evidence of the bias being toward target responses (BF = 2.9×10^{26}, CI = -0.38 to -0.28, and BF = 5.1×10^{4}, CI = -0.15 to -0.07, respectively). There was also positive evidence against the inclusion of a modality main effect (BF = 1/9.4).

Single-target trials in Single vs. Dual Blocks. Mean correct RT for single target-trials was faster in the 75% than 50% group (918ms vs. 1039ms) and for visual compared to auditory (911ms vs. 1031ms). Critically, responding was also much faster in single than double blocks (809ms vs. 1133ms), indicating an effect of dual-task load, and this difference was larger in the 75% than 50% group (382ms vs. 249ms, BF = 2.8×10^{8}). The single vs. dual block difference was also larger for visual than auditory targets (426ms vs. 220ms, BF = 8.6×10^{20}).

An effect of dual-tasking was also supported by greater sensitivity (*d'*) in single than dual blocks (2.73 vs. 1.47). Sensitivity was greater in the 50% than 75% group (2.15 vs. 2.06), with the single vs. dual difference greater in the 75% than 50%

group (1.46 vs. 1.02, BF = 6.1×10^3). There was only equivocal evidence that *d'* differed between visual and auditory modalities (2.16 vs. 2.04, BF = 2.6).

Discussion

The detailed pattern of results in the Gatekeeper task suggested a speed-accuracy trade off in the 75% target group, due to a bias to make fast target responses. First, overall responding was faster in the 75% than 50% group. Single blocks (either exclusively auditory or exclusively visual) also differed from dual blocks (both auditory and visual streams require simultaneous monitoring) in that responses were faster. This was particularly so for visual single blocks, and was more evident for the 50% than 75% group. An overall tendency for faster target-present than non-target responses was observed, in accordance with other redundant-target studies (e.g., Eidels, Townsend, Hughes & Perry, 2015), and was similar in single and dual blocks. However, in contrast to dual blocks, in single blocks the relative disadvantage for non-targets in the 75% group was larger, even though that group was faster overall. This finding is consistent with fast target-biased responses in the 75% group. Supporting this conclusion, the signal-detection theory measure of target response bias was larger in the 75% group than the 50% group.

Comparison of single-target trials from single and dual blocks confirmed that the slower responding in the 50% vs. 75% group, and slower responding to auditory vs. visual targets, was greatest in single blocks. Sensitivity, measured by d', was greater in the 50% group than the 75% group, suggesting a speed-accuracy trade-off due to faster and less accurate responses in the 75% group. However, a speed-accuracy trade-off was not indicated for the faster visual responses, which if anything were more accurate than auditory responses.

Of importance for questions about multi-tasking and capacity, responding to single-target trials was much faster in single than double blocks, even though in both there was only one target and only one response was required. Sensitivity was also much greater in single than dual blocks, ruling out a speed-accuracy trade-off, and suggesting strong dual-task demands on the capacity available for information processing. Performance differences across single and dual blocks are commensurate with basic principles of information theory, where performance depends not only on the stimulus currently presented but also on other stimuli in the set that could have been presented, although may not be displayed on that particular trial (e.g., Garner, 1974). We investigate this issue in more detail in the next section.

Heathcote et al.'s (2014) finding that some Gatekeeper performance measures have better reliability than traditional n-back measures was replicated and was somewhat stronger, perhaps due to the greater homogeneity of the undergraduate participant sample in the present experiment compared to their online sample, where ages ranged over seven decades. In particular, for both 50% and 75% groups, and down to as few as 50 trials, the reliability of dual-block accuracy was 0.9 or greater, and single-block accuracy, as well as mean RT for dual and single blocks, had 0.95 or better reliability. Reliability was similar for the 50% and 75% groups and for analogous measures in single and dual blocks.

Results for the 75% group were consistent with our expectation of little correlation between OSPAN recall and Gatekeeper accuracy. However, in the 50% group there was strong evidence for a correlation of 0.43 with dual-block accuracy and 0.3 with single-block accuracy. The high reliability of the Gatekeeper accuracy score might be one reason for these strong correlations, but it cannot be the only factor, as the high correlation is specific to the 50% group and dual-block accuracy,

whereas reliability was equally high for both groups and single-block accuracy. To be specific, there was strong evidence for a greater correlation in the 50% than in the 75% group between dual-block (BF = 655) and single-block (BF = 47) accuracy. Furthermore, when both single- and dual-block accuracy (which are themselves highly correlated) were entered into a regression on OSPAN recall, a model with only dual-block accuracy was selected (BF = 862) and there was positive evidence against including both predictors (BF = 1/5.4). We discuss the relationship with OSPAN further in the General Discussion, but we first turn to the measurement of workload capacity.

Working Memory's Workload Capacity

We have argued that the Gatekeeper task is strongly affected by two types of interference acting between trials within a single task (proactive interference) and within trials from two simultaneous tasks (dual-task interference). In our analysis comparing single-target trials, we found single-target responses in dual-task blocks were both slower and less accurate than responses in single-task blocks. In order to better understand the role of dual-task interference we used Townsend and Nozawa's (1995) Systems-Factorial Technology, which provides a rigorous measure of the level of dual-task interference.

In Systems Factorial Technology, the speed of a double-target condition relative to single-target conditions has been used to ascertain whether processing in two perceptual processes or "channels" share a limited pool of capacity. We use double-target responses from dual-task blocks and single-target responses from single-task blocks to ask the same question about the processes matching stimuli on the current trial to the contents of working memory. We did not use in this calculation single-target trials in the dual-task blocks, as they still require processing in both

channels, which could potentially cause some capacity sharing and require interference control, and so would address a somewhat different definition of workload capacity.

Most applications of Systems Factorial Technology have defined workload capacity using RT, by comparing the distribution of RT for double and single target conditions. RT distribution can also be characterized in terms of a hazard-rate function, h(t), the instantaneous probability that a response occurs at time t given that it has not already occurred. In particular, workload capacity at time t is defined as

$$C(t) = H_{AV}(t) / (H_A(t) + H_V(t)), \qquad (1)$$

where H(t) is the integral of h(t) from zero to t, the subscript AV indicates the double-target (auditory and visual) condition and the subscripts A and V the single auditory and visual target conditions, respectively. If processing occurs in parallel and is statistically independent for the auditory and visual channels, and at time t channels do not share capacity (i.e., in the double-target condition processing in the auditory channel does not affect the speed of the visual channel and vice versa), $C(t) = 1$.

The unlimited-capacity independent parallel model acts as a baseline against which to compare other cases. For example, if capacity is limited in the sense that processing is serial (i.e., only one channel is active at any given time), $C(t) = ½$. Similarly, if processing is parallel but there is a fixed capacity that is shared among active channels, so that processing in each channel is slowed in the double-target condition relative to the single-target conditions, $C(t) = ½$. Partial sharing, or a decrease in overall capacity available to be shared as more channels become active, can result in other values of $C(t) < 1$. Supercapacity – where processing in each channel is faster in the double than single target conditions – occurs when $C(t) > 1$,

and is associated with positive interactions between channels, such can arise from gestalt phenomena (e.g., Eidels, Townsend & Pomerantz, 2008).

Systems factorial technology has usually been applied to high-accuracy paradigms and so has focused on RT (but see Townsend & Altieri, 2012; Donkin, Little & Houpt, in press). Given Gatekeeper performance is error prone we also examined a measure of workload capacity based on error rates for targets, Cp, which we define below. Townsend and Altieri (2012) present another approach, based on measures of workload capacity they called assessment functions, which simultaneously take into account both RT and accuracy. However, these measures are a function of time, and cannot be readily subjected to regression analysis. For C(t), Houpt and Townsend (2012) derived a convenient summary statistic that can be used to calculate correlation with other measures, such as OSPAN. Hence, we preferred to use Houpt and Townsend's measure along with the Cp summary statistic for capacity based on accuracy, although we acknowledge that future research might seek to exploit the extra information contained in the time course of C(t).

Like the RT based measure, the accuracy-based workload capacity measure compares single and double target performance and again uses the unlimited-capacity parallel independent model as a baseline. Assuming statistical independence, and that activity in one channel does not affect accuracy of processing in another channel:

$$p(\text{miss} \mid \text{double}) = p(\text{miss} \mid \text{single visual}) \times p(\text{miss} \mid \text{single auditory}) \quad (2)$$

For example, if there was a 10% error rate in each of the single conditions then (2) predicts only a 1% error rate in the double condition. We can then define a capacity measure in terms of error probabilities that has a baseline value of zero and is positive for super-capacity and negative for limited capacity:

$$Cp = p(\text{miss} \mid \text{single visual}) \times p(\text{miss} \mid \text{single auditory}) - p(\text{miss} \mid \text{double}) \quad (3)$$

The capacity measure for RT, C(t), is a continuous function over time. As indicated before, for statistical inference it is convenient to use a single-number summary of capacity. Houpt and Townsend (2012) defined such a summary, the measurement-error weighted average of C(t) over each time point, which we calculated using the *sft* package for the R statistical language (Houpt, et al. 2014). We call this measure Cz as it has a standard normal distribution if the baseline model holds. Like Cp, Cz has a baseline value of zero when capacity is unlimited, with positive values indicating super-capacity and negative values indicating limited capacity. We also calculated a version of Cz where the baseline value of zero equated to fixed capacity (i.e., C(t) = ½), which we called Czf. By using both Cz and Czf we were able to investigate individual differences in workload capacity in an absolute sense. That is, we can ask the question, do any of our participants display evidence of greater than fixed capacity or perhaps even supercapacity?

Figure 5 plots individual workload-capacity estimates. Given Cz and Czf have standard normal distributions for each participant assuming the unlimited capacity and fixed capacity models respectively, significant deviations from these models at the two-tailed 0.05 level correspond to values with an absolute magnitude greater than 1.96 (indicated by dotted lines in Figure 5). Shapiro-Wilk tests could not reject a normal model for the distribution of Czf over subjects for 50% (W = .99, p = 0.62) and 75% (W = .997, p = 0.99) groups, and for Cz for the 50% group (W = .98, p = 0.11), but did reject it for the 75% group (W = .972, p = 0.007). However, the latter result was due to a single positive outlier (see Figure 5); when it was removed the normal model was not rejected (W = .989, p = 0.35). In contrast Cp, which is not predicted to have a normal distribution, was strongly left-skewed with a large mode just below zero for both 50% (W = .854, p < .001) and 75% (W = .77, p < .001)

groups, and this was not changed when the two positive outliers for the 50% group were removed (W=.825, $p < .001$).

Table 4 gives Spearman-Brown split-half reliability estimates for the three workload-capacity statistics. Reliability was lower for the accuracy-based estimate but relatively good for the RT based estimates when based on all of the available data. Given that perceptual applications of workload capacity have generally been based on more trials per participant than the present experiment, the reliabilities for RT-based measures in Table 4 are quite encouraging. This performance can be attributed to the relatively high efficiency in the way Houpt and Townsend (2012) capacity estimates take a weighted combination of data across time. Given these results, we place more emphasis on interpretation of the RT-based workload capacity measures (Cz and Czf) than the accuracy-based measure (Cp).

Table 2 shows that Cz and Czf are very highly correlated as is expected given that they are measured from the same data, although the correlation is not perfect due to the way in which the weighting function across time interacts with different Cz and Czf baselines. Figure 5 shows they provide highly consistent classifications of individual participants. Table 2 also shows that Cp and dual-block accuracy are highly correlated, with Figure 5 showing that this association is limited to lower levels of accuracy due to the large mode in the Cp distribution just below zero. For the RT-based measures, Cz and Czf, the 1.96 standard unit cut-offs in Figure 5 indicate that 7 participants in the 50% group and 4 in the 75% can be classified as having significantly greater than fixed capacity and one in the latter group significantly greater than unlimited capacity (i.e., super). However, a much larger number of participants in the 50% and 75% groups were classified as having less than unlimited (89% and 96% respectively) and less than fixed (33% and 69% respectively) capacity.

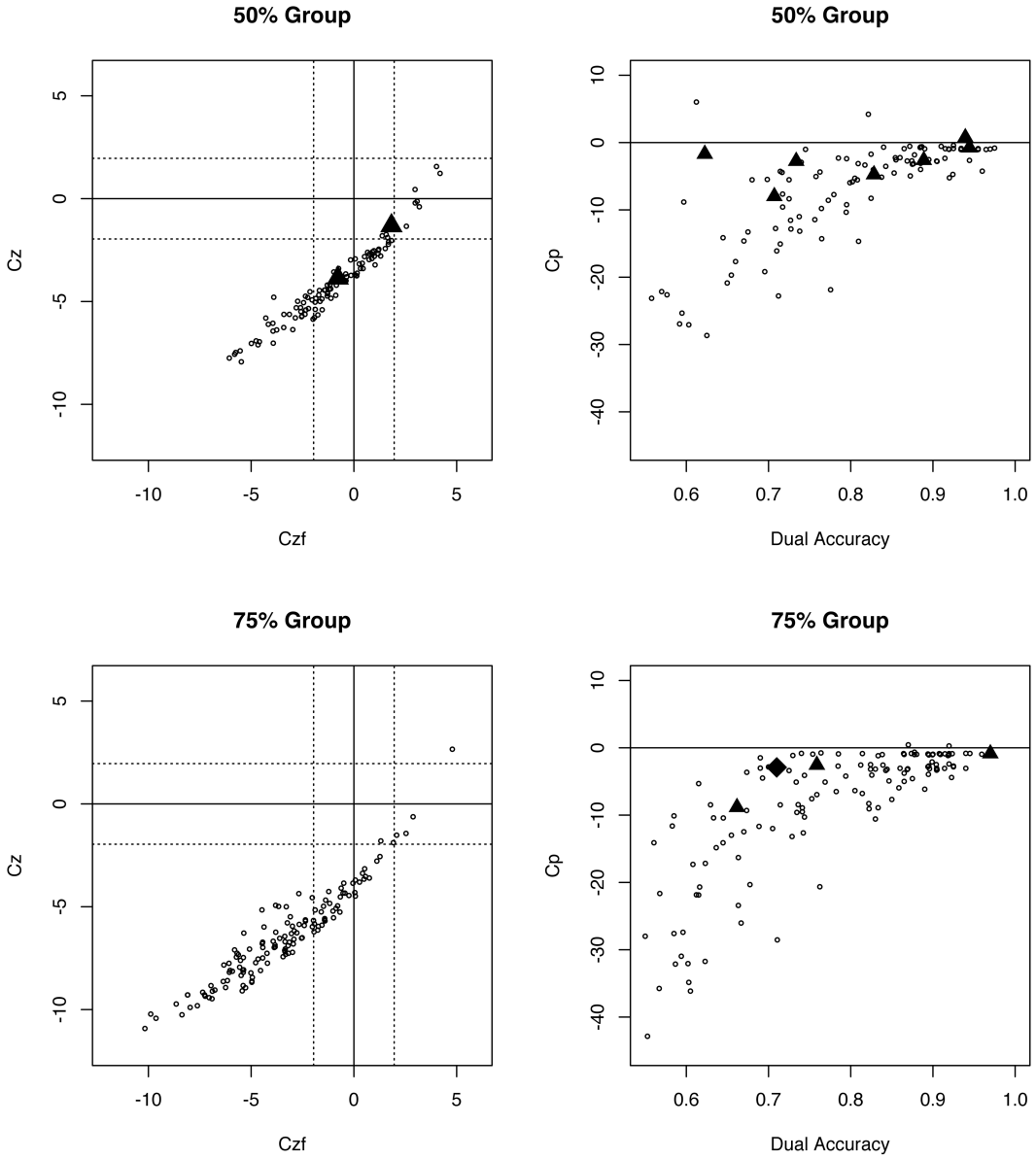

Figure 5. Scatterplots of workload capacity with a zero baseline of unlimited capacity (Cz) and fixed capacity (Czf), and accuracy-based capacity with a zero baseline of unlimited capacity (Cp) against accuracy in dual blocks. Solid lines indicate baselines and dotted lines are 1.96 standard units on either side of the baselines. Large triangle symbols in the Cz vs. Czf plots are the 2 participants with large Cp values (4.2 and 6 respectively). Large triangles in the Cp vs. Dual Accuracy plots are participants with Czf > 1.96 and the large diamond is a participant with Cz > 1.96.

Table 4. Average Spearman-Brown split-half reliabilities of workload capacity for a design with *n* trials, based on accuracy with a zero baseline for unlimited capacity (Cp), and based on RT with a zero baseline for unlimited capacity (Cz) and fixed capacity (Czf).

Group	50%			75%		
n	400	200	100	400	200	100
Cz	.86	.76	.67	.91	.84	.78
Czf	.85	.75	.66	.91	.83	.77
Cp	.76	.59	.49	.84	.72	.62

Consistent with the individual results, there was strong evidence that the population mean was less than zero for all three measures and for both groups (all BFs > 2000), indicating severely limited capacity that is less than fixed. There was also strong evidence for mean estimates being lower in the 75% group than the 50% group for Cz (-6.4 vs. -4.1, BF = 2.6×10^{12}) and Czf (-3.4 vs. -1.0, BF = 5.9×10^{8}) but not Cp (-8.2 vs. -6.9, BF = 3.61). It seems likely that the group differences are due to the strong dual-block target response bias displayed by the 75% group.

In summary, the population mean results indicated less than fixed capacity. The same held for about 50% of participants individually. Most of the remaining participants had performance that was not appreciably different from fixed capacity, with a small minority being closer to unlimited capacity (i.e., no dual task interference). The only cases that clearly exceeded unlimited capacity on one type of measure (i.e., RT or accuracy based) did not do so on the other, suggesting that they did not represent cases of genuine super-capacity (i.e., facilitation of performance in the dual-task setting). As shown in Table 4 capacity estimates were fairly reliable given measurement over the 400 trials used in our experiment.

General Discussion

The Gatekeeper task is a version of a dual n-back task with *n* fixed at two and using minimal sets of three auditory and three visuo-spatial stimuli developed by Heathcote et al. (2014). Binary speeded responses, which are required on every trial, indicate whether one or both of the stimuli are targets (i.e., match the stimulus from two trials back). The small stimulus sets and the constant remapping of associations to target and non-target responses promotes proactive interference and requires constant updating of bindings between representations, making Gatekeeper trials much more attention-demanding than the majority of trials in traditional n-back or dual n-back

tasks (Gray, Chabris & Braver, 2003). Gatekeeper also minimizes the effects of memory-capacity limitations that affect complex-span tasks, and so more directly measures individual differences in interference control in working memory.

In the following we discuss the main results we obtained from our analysis of performance in the Gatekeeper task. We first address the role of dual-task demands and the way in which we quantified them by applying the capacity measure developed by Townsend and Nozawa (1995). We then discuss the relationship between Gatekeeper performance and the widely used operation-span measure of working-memory capacity (Unsworth et al., 2005). We then discuss further potential applications and extensions of the Gatekeeper task.

Dual-Task Demands

Two sources of evidence suggested the presence of strong dual-task demands on the capacity available for information processing in the Gatekeeper task. First, we compared single-target trials in single and dual-task blocks, which enables us to measure dual-task interference with the number of targets and responses controlled. Average performance in terms of both accuracy and speed was clearly better in the single-block setting than the dual-block setting, supporting the presence of dual-task interference.

The second type of evidence came from our novel application to memory processes of the Systems Factorial Technology workload-capacity measure (Houpt & Townsend, 2012; Townsend & Nozawa, 1995). The Gatekeeper task is a version of a redundant target design widely used to investigate perceptual workload capacity, except that the definition of a target changes on every trial. Workload capacity is measured by comparing performance for double targets in the dual-task blocks to performance for single targets in the single-task blocks. Our results indicated severe

dual-task interference, with performance averaged over participants being clearly less than fixed capacity. That is, the interference was more than would be expected from sharing of a fixed amount of capacity between visual and auditory processes, or if visual and auditory processes were carried out sequentially.

Houpt and Townsend's (2012) measure also allowed us to look at performance at the individual participant level. About half of the participants displayed dual-block performance degraded below that of a fixed-capacity system. The remaining participants displayed performance consistent with fixed capacity, with very few approaching the level of performance associated with an unlimited capacity system (i.e., having no dual-task disadvantage). These latter participants might perhaps correspond to Watson and Strayer's (2010) *supertaskers* – individuals with extraordinary multi-tasking ability (see also Medeiros-Ward, Watson & Strayer, 2014) – in which case Houpt and Townsend's analysis of Gatekeeper performance might provide an efficient method of screening for such individuals. However, some caution is warranted given that we found some inconsistencies between accuracy- and RT-based performance measures.

Future research might seek to resolve inconsistencies between accuracy and RT based measures using evidence accumulation modeling (e.g., Brown & Heathcote, 2008; Ratcliff & Smith, 2004). Such models account for speed-accuracy tradeoffs observed in choice tasks in terms of the latent variables quantifying the rate of evidence accumulation and the amount of accumulated evidence required to trigger a decision. Such tradeoffs are ubiquitous and potentially confound inferences about psychological processes based on RT while ignoring accuracy or vise versa. Eidels, Donkin, Brown and Heathcote (2010) extended the Brown and Heathcote's *Linear Ballistic Accumulator* (LBA) model to account for choices relying on logical "OR"

and "AND" contingencies among multiple stimuli, and successfully applied the model to data from a perceptual redundant-target paradigm. Hence, the same extension would be appropriate for the Gatekeeper task, and would represent a potentially informative new cognitive-process-model variation on the latent variable modeling commonly used in working-memory research.

Gatekeeper and Operation Span

We also explored the relationship between the OSPAN complex-span measure and performance in the Gatekeeper task, and observed a surprisingly high correlation between for Gatekeeper accuracy in the 50% target condition: 0.43 in dual-blocks and 0.3 in single blocks. In contrast, most correlations between n-back sensitivity and complex span measures have been in the range between 0.1 – 0.24. Jaeggi et al., (2010b) noted that some exceptions – with magnitudes similar to our dual-block finding – might be attributable to improved reliability, obtained either by combining 0 to 3 back scores (Shelton, Elliott, Hill, Calamia & Gouvier, 2009) or several complex-span measures (Shamosh et al., 2008). Given the high reliability of our Gatekeeper accuracy scores, a similar factor may be in play here. The stronger result for dual blocks suggests that another important component of the high correlation with OSPAN recall is dual-task load, consistent with OSPAN also using two tasks, although stimuli for the two tasks occur sequentially rather than simultaneously as in Gatekeeper.

However, the specificity of the high correlations to the 50% condition – in the 75% condition it was at best 0.21 – suggests other factors are also important. One aspect that differentiates dual-block responding in the 50% group from that in the 75% group is that it was unbiased. Adopting target-biased responding likely requires participants to notice and act upon the predominance of targets in the 75% group,

which may inflate individual differences (i.e., some participants may be quick to learn the built-in contingencies while others take longer) and so deflate correlations. Consistent with this possibility, the standard deviation of bias estimates was substantially greater in the 75% group than the 50% group for both dual (0.46 vs. 0.32) and single (0.39 vs. 0.23) blocks.

Given the surprising nature of the correlation with OSPAN, and its basis in a relatively small sample (100 participants), more research is needed. A structural equation modeling approach will likely be advantageous in order to identify latent factors that underpin any shared variance. For example, one potential avenue is to explore whether Gatekeeper 50% target dual-task accuracy and complex span explain different components of variance in fluid intelligence, as has been found to be the case for n-back performance (Jaeggi et al., 2010a).

Future Directions

Such future research, and wider uses of the Gatekeeper task, is encouraged by the excellent reliability displayed by both accuracy and mean RT measures. It is likely that reliability was good because the small stimulus sets in the Gatekeeper task – and hence lures with a homogenous level of proactive interference – promote a constant level of difficulty for all trials, whereas in traditional n-back tasks with larger stimulus sets, in contrast, proactive interference, and hence difficulty, can fluctuate more widely. Schmiedek et al. (2009) noted: "the importance of carefully controlling the occurrence of lures in applications of n-back tasks [and that] … such control is possible to a considerable but not unlimited degree, due to combinatorial constraints." (p.207). Our use of small stimulus sets avoids these combinatorial constraints.

A further reason for higher reliability is that responses are collected on all trials in the Gatekeeper task, whereas in some other versions of n-back tasks

responses are required on only a subset of trials. Requiring response on all trials enabled collection of what proved to be the two most reliable Gatekeeper measures, overall accuracy and mean RT (i.e., accuracy and mean RT averaged over both target and non-target trials). If a choice is required between these measures, accuracy is likely preferable, even though it was slightly less reliable, as it showed greater sensitivity to individual differences than mean RT. Our results examining reliability as a function of number of trials suggest that the Gatekeeper accuracy and mean RT measures could be deployed with as few as 50 to 100 trials in applied settings where the larger number used in the present experiment are impractical. However, we would caution that a reasonable number of practice trials should always be given in order to make sure that participants understand the demands of the task.

Future applications of Gatekeeper could use either the dual blocks with 75% targets or 50% targets. In the 75% version we found fast target-biased responses and greater inconsistency in individual bias settings compared to the 50% version. These results suggest that the 50% version of Gatekeeper is preferable. However, the 50% version does introduce some predictability about the nature of the upcoming stimulus within each modality, because stimuli that occurred three trials back must be selected with greater probability than those the occurred two trials back.

Another way of achieving a 50% target probability, but without introducing predictability, is to use an exclusive-or ("XOR") response rule. That is, access to the club in the Gatekeeper task is blocked only if the stimulus in one modality occurred two back and the other did not. In this way, the XOR rule may be ideal because it makes equally probable the two possible stimuli that can occur in each modality and the two possible responses required by the combined stimuli.

A further potential advantage of an XOR version of Gatekeeper is that it increases dual-task interference and makes it more homogenous over trials. In the original Gatekeeper, due to the nature of redundant-target tasks, participants need only fully process one modality in order to make an accurate "block" response. That is, they need only detect that a stimulus in one or other modality occurred two back, so could stop processing both modalities as soon as detection occurs in one or the other. In contrast, they must fully process both modalities to make an accurate "allow" response, so the potential level of dual-task interference differs between these two trial types. The XOR version always requires both modalities to be fully processed because the correct response is defined by the relationship between the stimuli in each modality.

High and consistent levels of dual-task interference in the XOR version of Gatekeeper, combined with the high and consistent levels of proactive interference attending the use of small stimulus sets, has the potential to create an even more challenging and reliable task. We are presently replicating the experiment reported here using the XOR Gatekeeper in order to explore this potential. If it fulfills its promise we then plan to attempt to develop a unified account of both speed and accuracy in the XOR Gatekeeper task by extending the LBA-based methods developed in Eidels et al. (2010) to model an XOR logical contingency.

Conclusions

The number of successfully retrieved items often defines working memory capacity. In perceptual tasks another type of capacity has been discussed, workload capacity. Workload capacity underpins the ability to process information as processing load increases through an increase in the number of signals to be processed. We developed a novel task and analyses that allow assessment of workload

capacity in working memory. The task, Gatekeeper, requires maintenance of information in working memory about either one or two types of attributes for the last two items studied. By comparing performance in single- vs dual-attribute versions, Gatekeeper provides reliable measures of working memory's workload capacity. These measures, in turn, enable the understanding of individual differences, indicating where dual-task performance is better characterized by unlimited capacity and where it is better characterized as fixed or limited capacity. We found limited capacity to be the predominant case here when processing both visual and auditory attributes. Taken together, the new task and measurement approach help to sharpen our theoretical understanding of working memory capacity and multitasking ability.

Acknowledgements

We acknowledge support from ARC Professorial Fellowship DP110100234 to AH and an ARC Discovery Project DP120102907 to AE and AH.

References

Altieri, N. & Townsend, J. T. (2011). An assessment of behavioral dynamic information processing measures in audiovisual speech perception. *Frontiers in Psychology, 2*, 1–15.

Barrett, Tugade, M. M. & Engle, R. W. (2004). Individual Differences in Working Memory Capacity and Dual-Process Theories of the Mind. *Psychological Bulletin*, *130*(4), 553–573.

Baguley, T. (2011). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods*, *44*(1), 158–175.

Brown, S. D. & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Burgess, G. C., Gray, J. R., Conway, A. R. A. & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and memory span. *Journal of Experimental Psychology. General*, 140, 674-692.

Burns, D., Houpt, J.W., Townsend, J.T. & Endres, M.J. (2013). Functional principal components analysis of workload capacity functions. *Behavior Research Methods, 45*, 1048-1057.

Chuderski, A. (2013). When are fluid intelligence and working memory isomorphic and when are they not? *Intelligence, 41*, 244-262.

Clapp, W. C. & Gazzaley, A. (2012). Distinct mechanisms for the impact of distraction and interruption on working memory in aging. *Neurobiology of Aging*, *33*, 134–148.

Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology, 55*, 75-84.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O. & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769-786.

Conway, A. R., Kane, M. J. & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*(12), 547–552.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity, *Behavioral & Brain Sciences*, *24*, 87-185.

Dobbs, A. R. & Rule, B. G. (1989). Adult age differences in working memory. *Psychology and Aging*, *4*(4), 500.

Donkin, C., Little, D. & Houpt, J. W. (in press). Assessing the speed-accuracy trade-off effect on the capacity of information processing. *Journal of Experimental Psychology: Human Perception and Performanc*e.

Donnelly, N., Cornes, K. & Menneer, T. (2012). An examination of the processing capacity of features in the Thatcher illusion. *Attention, Perception & Psychophysics, 74*, 1475–1487.

Duff, S. J. & Hampson, E. (2001). A Sex Difference on a Novel Spatial Working Memory Task in Humans. *Brain and Cognition*, *47*(3), 470–493.

Eidels, A., Donkin, C., Brown, S. D. & Heathcote, A. (2010). Converging measures of workload capacity. Psychonomic Bulletin & Review, 17(6), 763–771.

Eidels, A., Townsend, J.T. & Algom, D. (2010). Comparing perception of Stroop stimuli in focused versus divided attention paradigms: Evidence for dramatic processing differences. *Cognition, 114*, 129-150.

Eidels A., Townsend J. T., Hughes H. C., & Perry L. A. (2015). Evaluating perceptual integration: uniting response-time- and accuracy-based methodologies. *Attention, Perception, & Psychophysics, 77* 659-680.

Eidels, A., Townsend, J. T. & Pomerantz, J. R. (2008). Where similarity beats redundancy: the importance of context, higher order similarity, and response assignment. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1441–1463.

Engle, R. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19-23.

Fitousi, D. & Wenger, M.J. (2011). Processing capacity under perceptual and cognitive load: A closer look at load theory. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 781–798.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbum.

Gray, J. R., Chabris, C. F. & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, *6*, 316–322.

Halford, G. S., Wilson, W. H. & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803–864.

Heathcote, A., Eidels, A., Houpt, J., Colman, J., Watson, J. & Strayer, D. (2014). Multi-tasking in working memory. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Houpt, J.W., Blaha, L.M., McIntire, J.P., Havig, P.R. & Townsend, J.T. (2014). Systems Factorial Technology with R. *Behavior Research Methods.*

Houpt, J. W. & Townsend, J. T. (2012). Journal of Mathematical Psychology. *Journal of Mathematical Psychology*, *56*(5), 341–355.

Houpt, J.W., Townsend, J.T. & Donkin, C. (2014). A new perspective on visual word processing efficiency. *Acta Psychologica, 145*, 118-127.

Hyde, J. S. (2005). The Gender Similarities Hypothesis. American Psychologist, 60(6), 581–592.

Ingvalson, E. M. & Wenger,M. J. (2005). A strong test of the dual-mode hypothesis. *Perception & Psychophysics, 67*, 14–35.

Jaeggi, S. M., Buschkuehl, M., Etienne, A., Ozdoba, C., Perrig, W. J. & Nirkko, A. C. (2007). On how high performers keep cool brains in situations of cognitive overload. *Cognitive, Affective & Behavioral Neuroscience*, 7, 75–89.

Jaeggi, S. M., Seewer, R., Nirkko, A. C., Eckstein, D., Schroth, G., Groner, R. & Gutbrod, K. (2003). Does excessive memory load attenuate activation in the prefrontal cortex? Load-dependent processing in single and dual tasks: functional magnetic resonance imaging study. *NeuroImage*, 19, 210–225.

Jaeggi, S. M., Buschkuehl, M., Jonides, J. & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 6829–6833.

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y.-F., Jonides, J. & Perrig, W. J. (2010a). The relationship between n-back performance and matrix reasoning — implications for training and transfer. *Intelligence*, *38*(6), 625–635.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J. & Meier, B. (2010b). The concurrent validity of the N-back task as a working memory measure. Memory, 18(4), 394–412.

Jeffreys, H. (1961). *The Theory of Probability* (3 ed.). Oxford University Press: Oxford.

Johnson, S. A., Blaha, L. M., Houpt, J. W. & Townsend, J. T. (2010). Systems factorial technology provides new insights on global–local information processing in autism spectrum disorders. *Journal of Mathematical Psychology, 54*, 53–72.

Kahneman, D. (1973). *Attention and effort*. New York: Prentice Hall.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kane, M. J., Conway, A. R. A., Miura, T. K. & Colflesh, G. J. H. (2007). Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 615–622.

Keppel, G. & Underwood, B.J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning and Verbal Behavior*, 1, 153-161.

Loring-Meier, S. & Halpern, D. F. (1999). Sex differences in visuospatial working memory: Components of cognitive processing. *Psychonomic Bulletin & Review*, *6*(3), 464–471.

Mantyla, T. (2013). Gender Differences in Multitasking Reflect Spatial Ability. *Psychological Science*, *24*(4), 514–520.

Mantyla, T. & Todorov, I. (2013). Questioning anecdotal beliefs and scientific Findings: A reply to Strayer, Medeiros-Ward, and Watson (2013). *Psychological Science*, *24*(5), 811–812.

McElree, B. (2001). Working memory and the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 817–835.

Medeiros-Ward, N., Watson, J. M., & Strayer, D. L. (2014). On Supertaskers and the Neural Basis of Efficient Multitasking. *Psychonomic Bulletin & Review.*

Morey, C. C. & Cowan, N. (2004). When visual and verbal memories compete: Evidence of cross-domain limits in working memory. *Psychonomic Bulletin & Review*, *11*, 296-301.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*, 61–64.

Morey, R. D. & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4), 406–419.

Morey, R. D. & Rounder, J. (2012). *BayesFactor: An R package for computing Bayes factors in common research designs*. http://bayesfactorpcl.r-forge.r-project.org/

Neufeld, R. W., Townsend, J. T. & Jetté, J. (2007). Quantitative response time technology for measuring cognitive-processing capacity in clinical studies. In R.W. Neufeld (Ed.), *Advances in clinical cognitive science: Formal modeling and assessment of processes and symptoms* (pp. 207–238). Washington, D. C: American Psychological Association.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 411–421.

Oberauer, K. (2005). Binding and Inhibition in Working Memory: Individual and Age Differences in Short-Term Recognition. *Journal of experimental psychology: General*, 134, 368–387.

Oberauer, K., Süβ, H.-M., Wilhelm, O. & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence*, *36*(6), 641–652.

Owen, A.M., McMillan, K.M., Laird, A.R. & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, *25*, 46-59.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–164.

Ratcliff, R. & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. *Psychological Review*, *111*(2), 333–367.

Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374.

Salthouse, T. A. (1994). The aging of working memory. *Neuropsychology*, *8*, 535–543.

Schmidt, H., Jogia, J., Fast, K., Christodoulou, T., Haldane, M., Kumari, V. & Frangou, S. (2009). No gender differences in brain activation during the N-back task: An fMRI study in healthy individuals. *Human Brain Mapping*, *30*(11), 3609–3615

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O. & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*, 1089-1096.

Schmiedek, F., Li, S.-C. & Lindenberger, U. (2009). Interference and facilitation in spatial working memory: Age-associated differences in lure effects in the n-back paradigm. *Psychology and Aging*, *24*(1), 203–210.

Schneider, W. & Shiffrin, R. M. (1977). Controlled and automatic human information processing. I. Detection, search, and attention. *Psychological Review*, *84*, 1–66.

Shamosh, N. A., DeYoung, C. G., Green, A. E., Reis, D. L., Johnson, M. R., Conway, A. R. A., Engle, R.W., Braver, T.S. & Gray, J.R. (2008). Individual Differences in Delay Discounting: Relation to Intelligence, Working Memory, and Anterior Prefrontal Cortex. *Psychological Science*, *19*, 904–911.

Shelton, J. T., Elliott, E. M., Hill, B. D., Calamia, M. R. & Gouvier, W. D. (2009). Intelligence. *Intelligence*, *37*(3), 283–293.

Shiffrin, R. M. & Schneider, W. (1977). Controlled and automatic human information processing. II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*, 127–190.

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers*, *31*(1), 137–149.

Strayer, D.L., Medeiros-Ward, N. & Watson, J.M. (2013). Gender invariance in multitasking: A comment on Mantyla (2013). *Psychological Science*, *24,* 809-810.

Watson, J. M. & Strayer, D. L. (2010). Supertaskers: Profiles in extraordinary multitasking ability. *Psychonomic Bulletin & Review*, *17*(4), 479–485.

Strayer, D. L. & Watson, J. M. (2012). Supertaskers and the multitasking brain. *Scientific American Mind*, *23*, 22-29.

Townsend, J. T. & Altieri, N. (2012). An accuracy–response time capacity assessment function that measures performance against standard parallel predictions. *Psychological Review*, *119*(3), 500–516.

Townsend, J. T. & Eidels, A. (2011). Workload capacity spaces: A unified methodology for response time measures of efficiency as workload is varied. *Psychonomic Bulletin & Review, 18*, 659–681.

Townsend, J. T. & Honey, C. J. (2007). Consequences of base time for redundant signals experiments. *Journal of Mathematical Psychology, 51*, 242–265.

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*, 321–359.

Townsend, J. T. & Wenger, M. J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review, 111*, 1003–1035.

Turner, M. L. & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154.

Unsworth, N., Heitz, R. P., Schrock, J. C. & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505.

Von Der Heide, R. J.,Wenger,M. J., Gilmore, R. O. & Elbich, D. (2011). Developmental changes in encoding and the capacity to process face information. *Journal of Vision, 11*, 450.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Watson, J. M. & Strayer, D. L. (2010). Supertaskers: Profiles in extraordinary multitasking ability. *Psychonomic Bulletin & Review*, *17*, 479–485.

Watson, J.M., Lambert, A.E., Miller, A.E. & Strayer, D.L. (2011). The magical letters P, F, C, and sometimes U: The rise and fall of executive attention with the development of prefrontal cortex. In K. Fingerman, C. Berg, J. Smith & T. Antonucci (Eds.), *Handbook of Lifespan Psychology* (pp. 407-436). New York: Springer.

Weiss, E., Kemmler, G., Deisenhammer, E. A., Fleischhacker, W. W. & Delazer, M. (2003). Sex differences in cognitive functions. *Personality and Individual Differences*, *35*, 863–875.

Wenger, M.J. & Gibson, B.S. (2004). Using hazard functions to assess changes in processing capacity in an attentional cuing paradigm. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 708–719.

Wenger, M.J. & Townsend, J.T. (2006). On the costs and benefits of faces and words: Process characteristics of feature search in highly meaningful stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 755–779.

Wickens, C. D. (1980). The structure of attentional resources. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 239-257). Hills- dale, NJ: Erlbaum.

Zehetleitner, M., Krummenacher, J. & Müller, H. J. (2009). The detection of feature singletons defined in two dimensions is based on salience summation, rather than on serial exhaustive architectures. *Attention, Perception & Psychophysics, 71*, 1739–1759.