Mem Cogn (2015) 43:973–989
DOI 10.3758/s13421-015-0526-2

Working memory's workload capacity

Andrew Heathcote[1] · James R. Coleman[2] · Ami Eidels[4] · Jason M. Watson[2] · Joseph Houpt[3] · David L. Strayer[2]

Published online: 12 May 2015

Abstract We examined the role of dual-task interference in working memory using a novel dual two-back task that requires a redundant-target response (i.e., a response that neither the auditory nor the visual stimulus occurred two back versus a response that one or both occurred two back) on every trial. Comparisons with performance on single two-back trials (i.e., with only auditory or only visual stimuli) showed that dual-task demands reduced both speed and accuracy. Our task design enabled a novel application of Townsend and Nozawa's (Journal of Mathematical Psychology 39: 321–359, 1995) workload capacity measure, which revealed that the decrement in dual two-back performance was mediated by the 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 two-back task that minimized response bias, accuracy was also more strongly correlated with complex span than has been found for most other single and dual *n*-back tasks.

Keywords Working memory · *n*-Back task · Systems factorial technology · Workload capacity · Multitasking · Proactive interference

In working memory tasks, participants are required to actively maintain information and also to manipulate that information or other information. Hence, these tasks are sensitive not only to limits in storage capacity (Cowan, 2001; C. C. 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. At least two types of interference must be considered 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 was the prime motivation for the present study. To do so, we used a dual two-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 two-back task, participants must indicate whether either of two attributes of the current stimulus had 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 fourth and fifth items do not repeat their two-back predecessors. One set of attributes is auditory and the other visual. Suppose the second sequence has the attributes X–Y–X–Y–Y . . . , in which both the third and fourth items are the same as their two-back predecessors, whereas the fifth is not. In a dual two-back task, the

✉ Andrew Heathcote
andrew.heathcote@utas.edu.au

[1] Schools of Medicine and Psychology, Universitys of Tasmania and Newcastle, Sandy Bay, Tasmania 7005, Australia

[2] Department of Psychology, University of Utah, Salt Lake City, UT, USA

[3] Department of Psychology, Wright State University, Dayton, OH, USA

[4] School of Psychology, Univeristy of Newcaste, Callaghan, NSW 2308, Australia

observer must monitor both sequences and respond affirmatively if an item in *either* sequence fulfills the two-back rule (e.g., both the third and fourth items in the example), and otherwise respond negatively (e.g., for the fifth item in the example). Tasks in which 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 the 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, 2014; Donnelly, Cornes, & Menneer, 2012; Eidels, Townsend, & Algom, 2010b; Eidels, Townsend, & Pomerantz, 2008; Fitousi & Wenger, 2011; Houpt, Townsend, & Donkin, 2014b; 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, 2014a; Houpt & Townsend, 2012; Townsend & Eidels, 2011; Townsend & Honey, 2007; Townsend & Wenger, 2004). We exploited the redundant-target nature of Heathcote et al.'s (2014) task to use it as a building block for measuring the workload capacity of working memory.

In the next section, we describe Heathcote et al.'s (2014) dual two-back task in detail, providing background on its relationship to the various tasks used to measure working memory. We then report and analyze an experiment that augments Heathcote et al.'s dual two-back task with single two-back tasks, enabling a workload capacity analysis. Comparing the information-processing latencies and accuracy with two versus one source of information is the cornerstone of the workload capacity analysis. A formal definition of capacity is given later, but in brief, if the processing efficiency with two sources of information were as good as is predicted by the summed efficiency of processing each source alone, capacity would be said to be unlimited. In contrast, limited capacity would be indicated if monitoring two streams took a toll on performance relative to one stream. 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., repeating 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 the 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 on storage capacity (e.g., Burgess, Gray, Conway, & Braver, 2011).

N-back tasks Another class of tasks uses the response times (RTs) and/or accuracy for choices to infer storage capacity. The *n*-back task is popular in cognitive neuroscience, because 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 the 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 on the basis of accurate performance (e.g., Jaeggi et al., 2008). In other cases, *n* is fixed at a smaller value so that accuracy is high, and interest focuses instead on RTs (e.g., Schmiedek, Li, & Lindenberger, 2009).

As well as differing in their response measures, complex-span tasks differ from *n*-back tasks in that they require the 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 has suggested that the latent constructs derived from these two classes of tasks 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 (Jaeggi et al., 2007; Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Jaeggi, Buschkuehl, Perrig, & Meier, 2010a;

Jaeggi et al., 2003; Jaeggi, Studer-Luethi, et al., 2010b). Participants are presented with pairs of visual and auditory stimuli, with a target response being required if a stimulus in either modality is a repeat from two trials previously, and a nontarget response is required otherwise (see the illustration in Fig. 1). Stimuli are never immediately repeated, so participants cannot use the easy, familiarity-based strategies available in a one-back task (McElree, 2001), based on the high availability of items held in the focus of attention (Oberauer, 2002). Because gatekeeper is a two-back task, it needs only four items to be held in memory at any time, so it does not exceed the storage capacity limits typically ascribed to working memory (Cowan, 2001; C. C. Morey & Cowan, 2004).

Gatekeeper uses a set of only three different stimuli in each modality, so the stimuli frequently swap roles as targets (i.e., the stimuli occurring two trials back) and lures, maximizing proactive interference. In contrast to most other *n*-back tasks, in which 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, since 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 the mapping of the response associated with each stimulus varies rapidly, minimizing 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 the 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, in which separate responses are made to the stimuli in each modality (e.g., Jaeggi et al., 2007; Jaeggi et al., 2008; Jaeggi et al., 2010a, b). 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—in which one stimulus is a target (i.e., it occurred two back) and one is not (i.e., it occurred three 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.

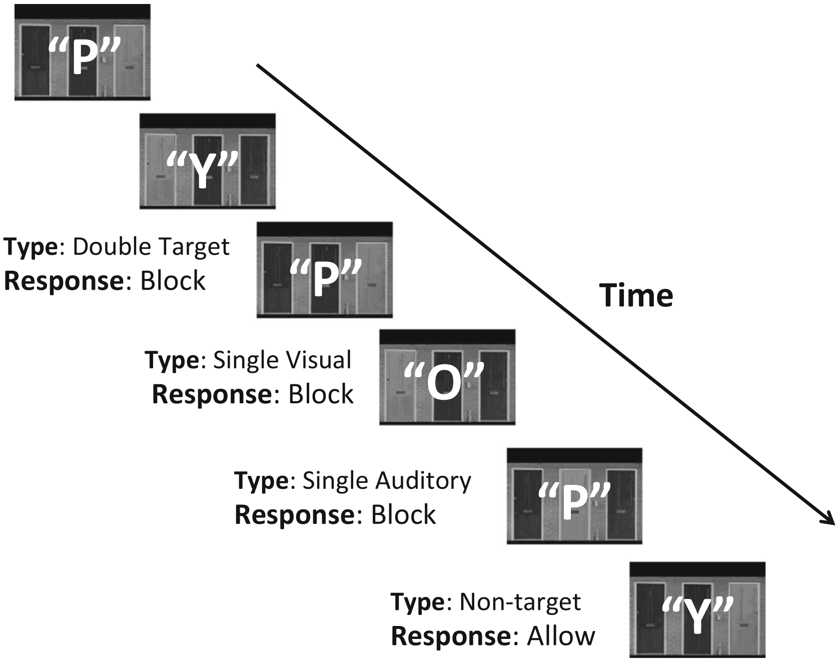

Fig. 1 Example of the first six trials in a dual-task gatekeeper block. The white letters indicate auditory stimuli (passwords), and the visual targets are the light-gray doors. Visual stimuli were presented in color, with the light-gray regions in red and the dark regions in black. No response was required for the first two trials. For each trial thereafter, the trial type and correct response are indicated. Single blocks present only the visual or only the 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*

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 that require responses to both targets and lures (e.g., Schmiedek et al., 2009b), Heathcote et al. found that RT was also a useful dependent measure.

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. The stimuli are both visuospatial (an image of three doors, in which the 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 that errors would tap memory and not perceptual errors. We also told participants that no patron was 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 emphasized by telling participants that only gatekeepers who can decide both quickly and accurately will make the grade and 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 visuospatial) and dual-task blocks. In their experiment examining practice effects, Jaeggi et al. (2010b) found better initial performance in a single than in a dual *n*-back task, and a faster improvement with practice. Hence, it seemed likely that we would observe better performance in single than in 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 the memory representations of the two-back auditory and visual stimuli, respectively, to the 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., in which participants monitor both auditory and visual streams) and single-target trials in single-task blocks (in which participants 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 the stimulus types occur with equal probabilities. 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 toward the more common "target-present" response. Because bias differences might affect the single- versus 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 the single blocks. The 50 % dual condition necessarily introduced some predictability (i.e., the next stimulus was more likely to have occurred three back), which might also lead to participant expectations that could affect performance. By running both 50 % and 75 % conditions, we sought to determine whether difference in expectations led to differences in performance that affected our 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, Unsworth, 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., 2010a, for a summary). Furthermore, 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, as 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 was possible that the high reliability of gatekeeper that Heathcote et al. (2014) found, particularly for mean RT and accuracy measures, might lead to a higher correlation. We report the results from the Spearman–Brown split-half procedure in order to quantify the reliability of the present experiment.

In summary, the purpose of the present 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 a strong and homogeneous level of proactive interference as well as multitasking demands. Our analysis first established the reliability of the performance metrics obtained with gatekeeper, and then contrasted them with another working memory capacity measure provided by the OSPAN task. Finally, systems factorial theory (Townsend & Nozawa, 1995) was used to formally model individual differences in workload capacity, identifying when performance was characterized by limited or fixed capacity and when performance was consistent with unlimited capacity.

Method

Participants

University of Utah undergraduates (372 total; 224 female, 148 male, 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 and then performed the OSPAN task, followed by the gatekeeper task. The data from 61 participants were 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 completed three practice blocks, then a simple letter span task and a block requiring the 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 three to seven 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 had been presented. The 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 3 × 4 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 2.5 s if no response was given, and a new trial would begin after a 1-s interval. As is illustrated in Fig. 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" was spoken through the computer speakers in a female voice. In single-task blocks, only the auditory or only the visual stimuli were presented. Responses were made via the keyboard using the "z" and "/ " keys to allow or block entry, with the mapping alternating 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 the 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 they could only do so after a mandatory 1-min break between blocks. At the conclusion, participants were given feedback about their overall performance.

The 16 experimental blocks were divided into eight dual-task blocks and four visual and four auditory single-task blocks. The order of the dual and single blocks was chosen randomly over participants, as was the order of the visual and auditory single blocks. The auditory and visual stimuli were selected randomly and independently, with the constraint that they never repeated immediately. In the 75 %-target dual-task blocks, the available stimuli (i.e., those that did not occur on the last trial) were chosen with equal probabilities, so no-target, visual-target-only, auditory-target-only, and double-target trials occurred on average with equal frequencies. In 50 %-target dual blocks, the available stimuli were randomly selected subject to the constraint that the 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, the performance in single-target conditions is contrasted with performance in dual-target conditions, and the psychometric reliability of the different measures is examined. Correlations between the different parameters of gatekeeper are computed and compared with correlations involving the measures obtained in the OSPAN task. Finally, data from the single-target and dual-target conditions are modeled using systems factorial technology to analyze 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 (R. D. Morey & Rounder, 2012) to perform Bayes factor (BF)-based tests of correlations, *t* tests, and analysis of variance (ANOVA) using Rouder, Morey, Speckman, and Province's (2012) default prior method. BFs are not subject to the bias, in traditional frequentist approaches using a fixed-criterion *p* value, of being increasingly likely to declare effects significant as sample size increases (Raftery, 1995, Table 9). In contrast to null-hypothesis testing (see R. D. 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 the 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 were included, so were 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 added a term to or removed a term from the best model. To do so, we used BFs for the best model relative to the alternative model, which were necessarily greater than 1.

For example, BF = 10 indicates that 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, because they have a natural interpretation in terms of support for the hypotheses provided by the data, these classifications provide a useful approximate guide when summarizing the 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 3, support increases 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 for a null effect). We also provide posterior medians to illustrate the effects and use 95 % credible intervals (CIs, the 2.5th to 97.5th percentiles of the parameter's posterior distribution) to quantify the uncertainty of these estimates.

Gatekeeper and OSPAN accuracy and exclusion criteria

Participants with more than 10 % nonresponses 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 Fig. 2. There was strong evidence for greater accuracy in single (82 %) than in 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 that they did not understand or engage with the gatekeeper task. These participants, defined by a score of less than 55 % correct in either the single or the double blocks (66), were removed from further analyses. In the remaining 245 participants, we again found strong evidence of greater accuracy in single (89 %) than in double (79 %) blocks (BF = 1.9 × 10^{36}; CI = 8.7 %–11.2 % for the accuracy difference).

The right panel of Fig. 2 plots recall and math accuracy in the OSPAN task for the full sample, with participants failing the gatekeeper accuracy cutoff being plotted as triangles. Recall accuracy in the overall sample (76 %) increased only slightly (to 78 %) when participants failing the gatekeeper cutoff were removed. Unsworth et al. (2005) recommended the 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 Fig. 2 plots as triangles the gatekeeper accuracy for the 32 participants (10 % of the overall sample) with less than the 85 % math accuracy cutoff. It shows that failure of the OSPAN cutoff was not associated with failure of the gatekeeper cutoff. When the 20 participants failing the OSPAN cutoff were removed from the 245 who passed the gatekeeper cutoff, 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, for which we used only the 225 participants who passed both cutoffs.

Reliability

Table 1 displays the Spearman–Brown split-half reliabilities for the statistics derived from dual and single blocks in the gatekeeper task for data from all n = 200 trials and from 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

[1] Kass and Raftery (1995) suggested a similar scheme, but with 3–20 labeled *positive*, 20–150 *strong*, and greater than 150 *very strong*. They also discussed how a BF can be understood in terms of the relative abilities of models to predict the 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 versus model A means that model A is still remains 10 times more likely. In our application, we do not think that any such strong prior beliefs would substantially distort the conventional labeling.

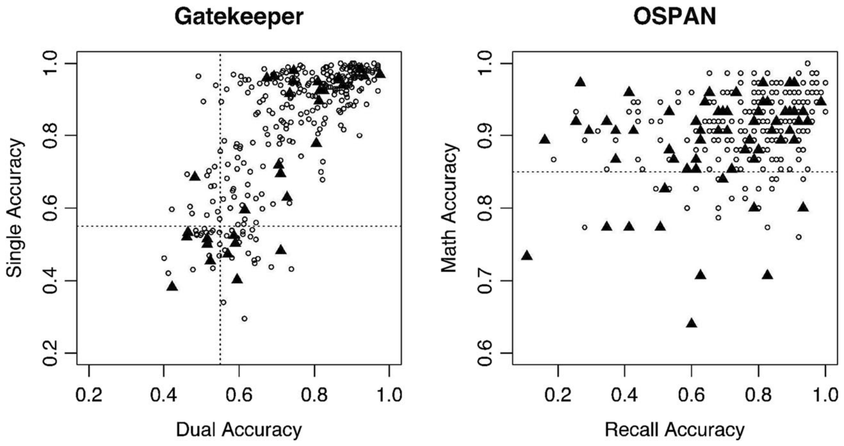

Fig. 2 The left panel shows accuracy in the single-target and double-target blocks of the gatekeeper task. Circles represent participants with greater than the 85 % cutoff for math accuracy in the OSPAN task, and triangles represent the excluded participants. The dotted lines represent the 55 % accuracy cutoffs in the gatekeeper task. The right panel shows accuracy in the OSPAN task. Circles represent participants with greater than the 55 % cutoff for single and dual accuracy in the gatekeeper task, and triangles represent the excluded participants. The dotted line represents the math accuracy cutoff in the OSPAN task

negligible. In Table 1, we also quantify the reliability of the response-choice (i.e., "block" vs. "allow entry") data, both in terms of the overall accuracy (i.e., the 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 overall percentage correct (PC), overall mean RT (MRT), and signal detection sensitivity (d')

		50 %			75 %		
		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

The subscripts indicate statistics calculated on the basis of dual-target (av), auditory (a), or visual (v) single-target trials (relative to nontarget trials, in the case of d'), nontarget (no) trials (for dual blocks), and auditory nontarget (an) and visual nontarget (vn) trials (for single blocks)

OSPAN–Gatekeeper correlations

Table 2 displays the correlations among OSPAN recall and selected gatekeeper performance measures (principally those with higher reliabilities). These correlations are based on only the results from participants with 85 % or greater accuracy in the OSPAN math task, and were calculated separately for the 50 % and 75 % groups (100 and 125 participants, respectively).

Accuracy and RT in single- and dual-task blocks

Figures 3 and 4 display the accuracy and mean RT results for the 50 % and 75 % groups, broken down by the different 2 × 2 within-subjects 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 the single-target trials, focusing on the effect of dual-task load. We examined the response probabilities using signal detection theory's sensitivity and bias measures. Table 3 reports the best model selected in the ANOVA analyses. In all but one case, a model that was simpler than the most complex ANOVA model was best.

Dual-block analysis In mean correct RTs, there was equivocal evidence for slower performance in the 50 % than in the 75 % group (1,139 vs. 1,079 ms, BF = 1.7). The main effects of slower nontarget than target performance were similar for auditory and visual targets (114 and 131 ms, respectively). We observed strong evidence for an interaction between the auditory- and visual-target

Table 2 Correlations among OSPAN recall accuracy and Gatekeeper performance measures for participants with accuracy 85 % for greater in the OSPAN math task

	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	$.18^{-}$	$.24^{-}$	

The upper triangle contains the results for the 50 % group, and the lower triangle, those for 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 nonzero), 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 of the measures, except the workload capacity measures (Cz, Czf, and Cp), which are defined and discussed in the Working Memory's Workload Capacity section

versus nontarget effects (BF = 3.1 × 10^{11}). As is shown in the right panels of Fig. 3, this was due to a greater slowing for auditory nontargets versus targets when the visual stimulus was also a target (168 ms) than when it was a nontarget (60 ms).

Sensitivity (d') was greater for the 50 % group than for 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 showed 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 Fig. 4.

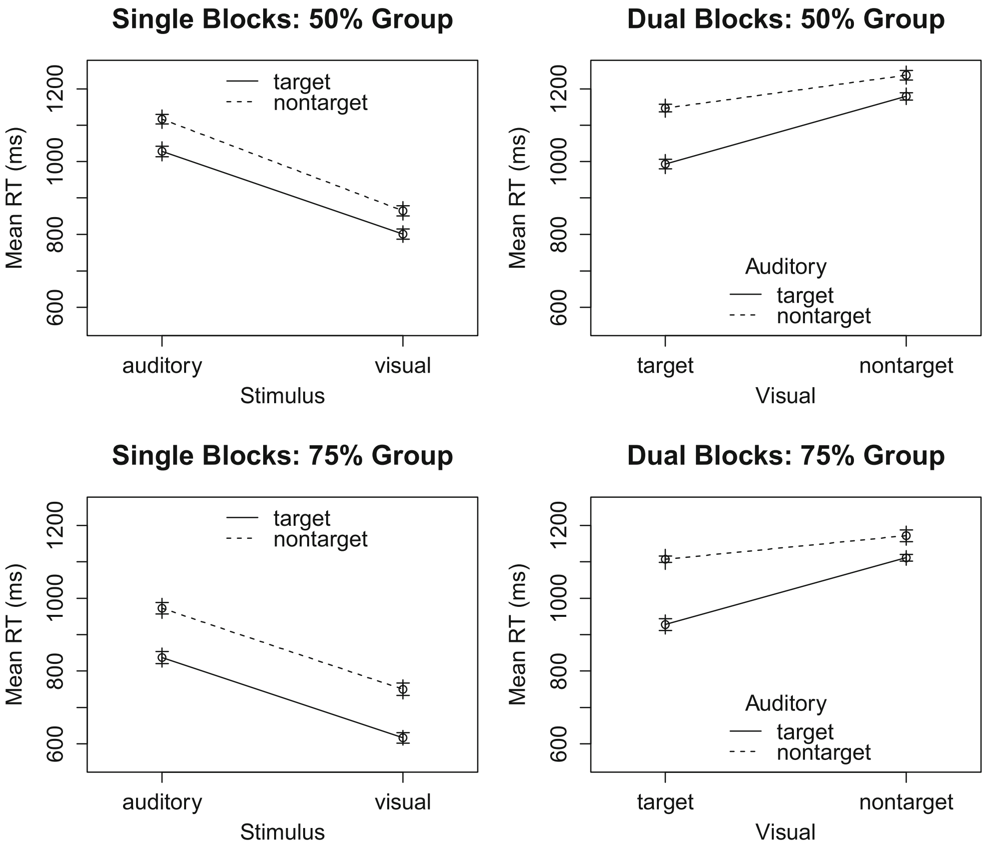

Fig. 3 Mean correct RTs, with 95 % within-subjects confidence intervals (R. D. Morey, 2008) depicted by horizontal lines and individual 95 % confidence intervals depicted by extended lines, as recommended by Baguley (2011)

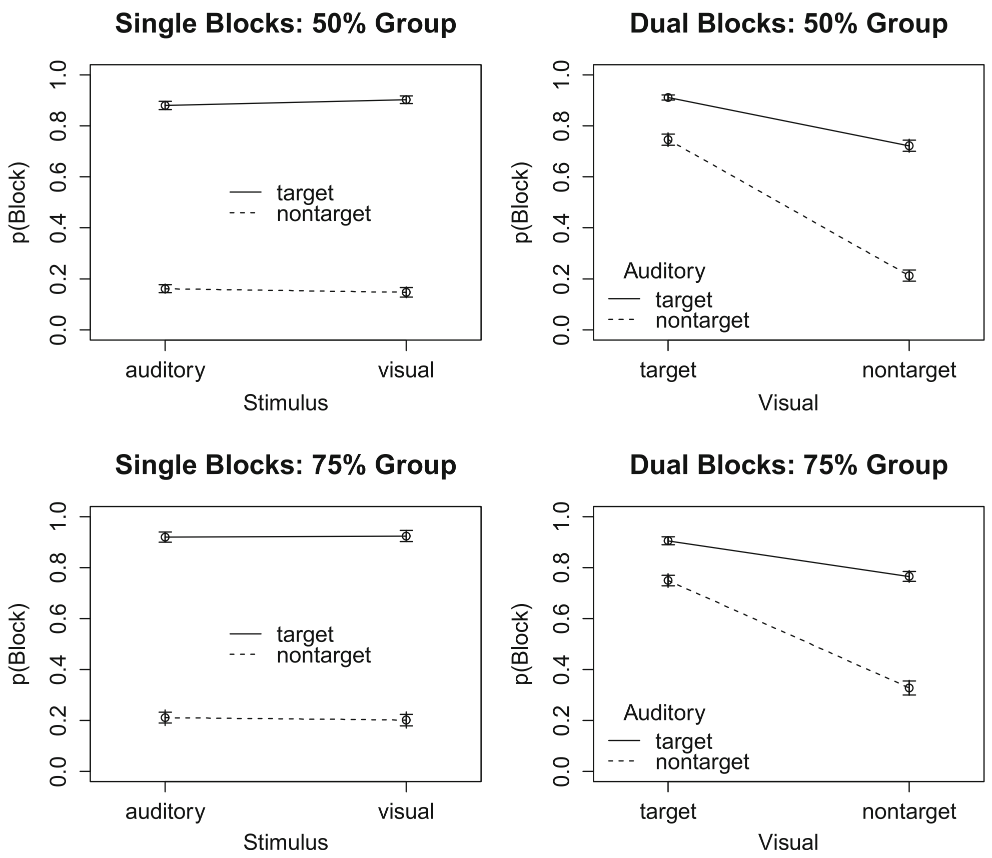

Fig. 4 Average probabilities of detecting a target (responding "Block"), with 95 % within-subjects confidence intervals (R. D. Morey, 2008) depicted by horizontal lines and individual 95 % confidence intervals depicted by vertical lines, as recommended by Baguley (2011)

We found 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).

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

Measure	ANOVA	Selected Model	Bayes Factor
Mean RT	Double blocks	A + V + G + A × V	184
	Single blocks	T + M + G + T × G	275
	Single trials	B + M + G + B × M + B × G	54
d'	Double blocks	TM + G	6.3
	Single blocks	M	14
	Single trials	B + M + G + B × G	34
c	Double blocks	G	1
	Single blocks	G	57

A = Auditory target vs. nontarget; B = single vs. double Block; G = 75 % vs. 50 % Group; M = visual vs. auditory Modality; T = Target vs. nontarget trial; TM = Target Modality, visual vs. auditory vs. both; V = Visual target vs. nontarget

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

As is shown in Table 3, strong evidence emerged for greater sensitivity in the visual than in the 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 that there was strong evidence for a difference in bias between the 75 % and 50 % groups ($c = -0.33$ 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). We also found positive evidence against the inclusion of a modality main effect (BF = 1/9.4).

Single-target trials in single versus dual blocks Mean correct RTs for single-target trials were faster in the 75 % than in the 50 % group (918 vs. 1,039 ms) and for visual than for auditory stimuli (911 vs. 1,031 ms). Critically, responding was also much faster in single than in double blocks (809 vs. 1133 ms), indicating an effect of dual-task load, and this difference was larger in the 75 % than the 50 % group (382 vs. 249 ms, BF = 2.8 × 10^8). The single- versus dual-block difference was also larger for visual than for auditory targets (426 vs. 220 ms, BF = 8.6 × 10^{20}).

An effect of dual-tasking was also supported by greater sensitivity (*d*') in single than in dual blocks (2.73 vs. 1.47). Sensitivity was greater in the 50 % than in the 75 % group (2.15 vs. 2.06), with the single-versus-dual difference being greater in the 75 % than in the 50 % group (1.46 vs. 1.02, BF = 6.1 × 10^3). The evidence was only equivocal that *d*' differed between the 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-present" responses. First, overall responding was faster in the 75 % than in the 50 % group. Single blocks (either exclusively auditory or exclusively visual) also differed from dual blocks (in which both auditory and visual streams required simultaneous monitoring), in that responses were faster. This was particularly so for visual single blocks, and was more evident for the 50 % than for the 75 % group. An overall tendency for faster "target-present" than nontarget responses was observed, in accordance with other redundant-target studies (e.g., Eidels, Townsend, Hughes, & Perry, 2015), and these trends were similar in single and dual blocks. However, in contrast to dual blocks, in single blocks the relative disadvantage for nontargets 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 greater in the 75 % than in the 50 % group.

Comparison of the single-target trials from the single and dual blocks confirmed that the slower responding in the 50 % than in the 75 % group, and the slower responding to auditory than to visual targets, was greatest in single blocks. Sensitivity, measured by *d*', was greater in the 50 % group than in 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 the auditory responses.

Of importance for questions about multitasking and capacity, responding to single-target trials was much faster in single than in double blocks, even though in both block types there was only one target and only one response was required. Sensitivity was also much greater in single than in 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 were commensurate with the basic principles of information theory, in which performance depends not only on the stimulus currently presented, but also on the other stimuli in the set that could have been presented, although they 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 than in their online sample, whose ages ranged over seven decades. In particular, for both the 50 % and 75 % groups, and down to as few as 50 trials, the reliability of dual-block accuracy was .9 or greater, and single-block accuracy, as well as the mean RTs for dual and single blocks, had .95 or better reliability. The reliabilities were similar for the 50 % and 75 % groups and for analogous measures in the single and dual blocks.

The results for the 75 % group were consistent with our expectation of little correlation between OSPAN recall and gatekeeper accuracy. However, in the 50 % group we found strong evidence for a correlation of .43 with dual-block accuracy and .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, since the high correlation is specific to the 50 % group and to dual-block accuracy, whereas the reliabilities were equally high for both groups and single-block accuracy. To be specific, strong evidence emerged 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 will discuss the relationship with OSPAN further in the General Discussion, but first we 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 the 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 that 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 fechnology, 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 is used to ascertain whether the processing related to two perceptual processes or "channels" shares a limited pool of capacity. We used double-target responses from dual-task blocks and single-target responses from single-task blocks to ask the same question about the processes for matching the 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, because they still required 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 distributions of RTs for double- and single-target conditions. RT distributions can also be characterized in terms of a hazard-rate function, $h(t)$, the instantaneous probability that a response will occur at time t, given that it has not already occurred. In particular, the workload capacity at time t is defined as

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

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 if 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 a fixed capacity is shared among active channels, so that the 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 the overall capacity that is available to be shared as more channels become active, can result in other values of $C(t) < 1$. Supercapacity—in which the processing in each channel is faster in double- than in single-target conditions—occurs when $C(t) > 1$, and is associated with positive interactions between the channels, such as can arise from gestalt phenomena (e.g., Eidels et al., 2008).

Systems factorial technology has usually been applied to high-accuracy paradigms, and so has focused on RTs (but see Donkin et al., 2014; Townsend & Altieri, 2012). Given that gatekeeper performance is error prone, we also examined a measure of workload capacity based on the error rates for targets, Cp, which we define below. Townsend and Altieri (2012) presented another approach, based on measures of workload capacity that 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 correlations 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 the accuracy of processing in another channel:

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

For example, if there were a 10 % error rate in each of the single conditions, Eq. 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 0 and is positive for supercapacity and negative for limited capacity:

$$Cp = p(\text{miss}|\text{single visual}) \times p(\text{miss}|\text{single auditory}) - p(\text{miss}|\text{double}). \tag{3}$$

The capacity measure for RT, $C(t)$, is a continuous function over time. As we 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., 2014a). We call this measure Cz, because it has a standard normal distribution if the baseline model holds. Like Cp, Cz has a baseline value of 0 when capacity is unlimited, with positive values indicating supercapacity and negative values indicating limited capacity. We also calculated a version of Cz in which the baseline value of 0 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 could ask the question, did any of our participants display evidence of greater-than-fixed capacity, or perhaps even supercapacity?

Figure 5 plots the individual workload capacity estimates. Given that 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 .05 level correspond to values with an absolute magnitude greater than 1.96 (indicated by the dotted lines in Fig. 5). Shapiro–Wilk tests could not reject a normal model for the distribution of Czf over participants for the 50 % (W = .99, p = .62) and 75 % (W = .997, p = .99) groups, and for Cz for the 50 % group (W = .98, p = .11), but they did reject it for the 75 % group (W = .972, p = .007). However, the latter result was due to a single positive outlier (see Fig. 5); when it was removed, the normal model was not rejected (W = .989, p = .35). In contrast, Cp, which is not predicted to have a normal distribution, was strongly left-skewed with a large mode just below 0 for both the 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 the Spearman–Brown split-half reliability estimates for the three workload capacity statistics. Reliability was lower for the accuracy-based estimate, but was relatively good for the RT-based estimates when they were 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 the 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 the interpretation of the RT-based workload capacity measures (Cz and Czf) than on the accuracy-based measure (Cp).

Table 2 shows that Cz and Czf are very highly correlated, as would be expected, given that they are measured from the

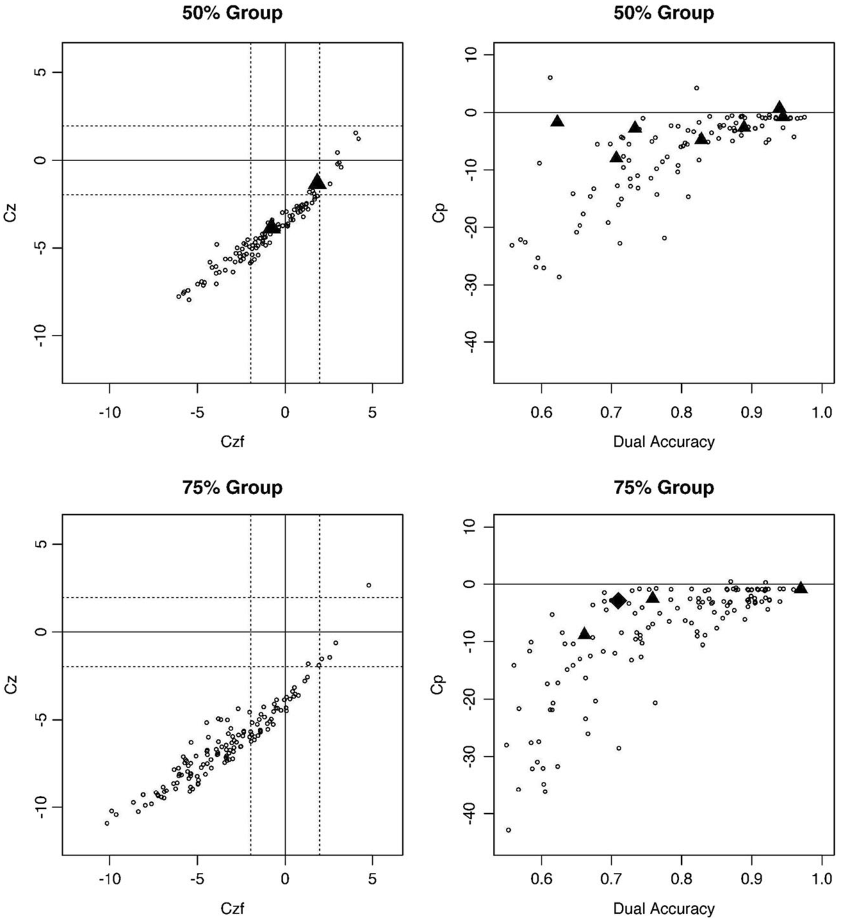

Fig. 5 Scatterplots of workload capacity with a zero baseline for unlimited capacity (Cz) and fixed capacity (Czf) (left panels), and accuracy-based capacity with a zero baseline for unlimited capacity (Cp) against accuracy in dual blocks (right panels). The solid lines indicate baselines, and the dotted lines are 1.96 standard units on either side of the baselines. The large triangle symbols in the Cz-versus-Czf plots are the two participants with large Cp values (4.2 and 6, respectively). The large triangles in the Cp-versus-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)

	50 %			75 %		
	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

same data, although the correlation is not perfect due to the ways in which the weighting function across time interacts with the different Cz and Czf baselines. Figure 5 shows that they provide highly consistent classifications of the individual participants. Table 2 also shows that Cp and dual-block accuracy are highly correlated, with Fig. 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 cutoffs in Fig. 5 indicate that seven participants in the 50 % group and four in the 75 % group could be classified as having significantly greater than fixed capacity, and one in the latter group significantly greater than unlimited capacity (i.e., supercapacity). 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.

Consistent with the individual results, we found strong evidence that the population means were less than zero for all three measures and for both groups (all BFs > 2,000), indicating severely limited capacity that is less than fixed. There was also strong evidence for mean estimates being lower in the 75 % group than in 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 for Cp (–8.2 vs. –6.9, BF = 3.61). It seems likely that the group differences are due to the strong dual-block “target-present” 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 the 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., RTs or accuracy) did not do so on the other, suggesting that they did not represent cases of genuine supercapacity (i.e., facilitation of performance in the dual-task setting). As is shown in Table 4, the 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 visuospatial 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 the associations to target and nontarget responses promotes proactive interference and requires constant updating of the 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 et al., 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 sections, we discuss the main results that 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 enabled us to measure dual-task interference with the numbers of both targets and responses controlled. Average performance in terms of both accuracy and speed was clearly better in the single- than in 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 the performance for double targets in the dual-task blocks to the 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 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 that was 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). The latter participants might perhaps correspond to Watson and Strayer's (2010) *supertaskers*—individuals with extraordinary multitasking 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 the 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 the speed–accuracy trade-offs 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 trade-offs are ubiquitous and potentially confound inferences about psychological processes based on RTs while ignoring accuracy, or vise versa. Eidels, Donkin, Brown, and Heathcote (2010a) extended Brown and Heathcote's linear ballistic accumulator (LBA) model to account for choices relying on the 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 for gatekeeper accuracy in the 50 %-target conditions: .43 in dual blocks and .3 in single blocks. In contrast, most correlations between *n*-back sensitivity and complex-span measures have been in the range between .1 and .24. Jaeggi et al. (2010a) noted that some exceptions—with magnitudes similar to our dual-block finding—might be attributable to improved reliability, obtained by combining either zero- to three-back scores (Shelton, Elliott, Hill, Calamia, & Gouvier, 2009) or several complex-span measures (Shamosh, De Young, Green, Reis, Johnson, Conway, … & Gray, 2008). Given the high reliability of our gatekeeper accuracy scores, a similar factor might 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 the stimuli for the two tasks occur sequentially rather than simultaneously, as in gatekeeper task.

However, the specificity of the high correlations to the 50 % condition—in the 75 % condition, correlations were at best .21—suggests that 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, whereas others take longer), and so deflate correlations. Consistent with this possibility, the standard deviations of bias estimates were substantially greater in the 75 % group than in the 50 % group for both dual (.46 vs. .32) and single (.39 vs. .23) blocks.

Given the surprising nature of the correlation with OSPAN, and its basis in a relatively small sample (100 participants), more research will be needed. A structural equation modeling approach would likely be advantageous, in order to identify the latent factors that underpin any shared variance. For example, one potential avenue would be to explore whether the 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., 2010b).

Future directions

Such future research and wider uses of the gatekeeper task are encouraged by the excellent reliability displayed by both the accuracy and mean-RT measures. It is likely that reliability was good because the small stimulus sets in the gatekeeper task—hence, lures with a homogeneous 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. (2009b) 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 avoided these combinatorial constraints.

A further reason for our 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 responses on all trials enabled the 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 nontarget trials). If a choice is required between these measures, accuracy would likely be preferable, even though it was slightly less reliable, since it showed greater sensitivity to individual differences than did mean RT. Our results examining reliability as a function of numbers of trials suggest that the gatekeeper accuracy and mean-RT measures could be deployed with as few as 50–100 trials in applied settings in which the larger number used in the present experiment would be impractical. However, we 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 the gatekeeper task could use dual blocks with either 75 % or 50 % targets. In the 75 % version, we found fast target-biased responses and greater inconsistency in individual bias settings than in 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 that occurred two trials back.

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

A further potential advantage of an XOR version of gatekeeper would be that it would increase dual-task interference and make it more homogeneous over trials. In the original gatekeeper task, 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 the other modality occurred two back, and so could stop processing both modalities as soon as detection had occurred in one or the other. In contrast, in the XOR version they must fully process both modalities to make an accurate "allow" response, so the potential levels of dual-task interference would differ 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 the two modalities.

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, have 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. (2010a) 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 allowed an 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 allowing comparisons of performance in single- and 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 characterized better by unlimited capacity and where it is characterized better by 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.

Author note We acknowledge support from ARC Professorial Fellowship DP110100234 to A.H. and an ARC Discovery Project DP120102907 to A.E. and A.H.

References

Altieri, N., & Townsend, J. T. (2011). An assessment of behavioral dynamic information processing measures in audiovisual speech perception. *Frontiers in Psychology, 2*(238), 1–15. doi:10.3389/fpsyg.2011.00238

Baguley, T. (2011). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods, 44,* 158–175. doi:10.3758/s13428-011-0123-7

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

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. M., Houpt, J. W., Townsend, J. T., & Endres, M. J. (2013). Functional principal components analysis of workload capacity functions. *Behavior Research Methods, 45,* 1048–1057. doi:10.3758/s13428-013-0333-2

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

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain*

Sciences, 24, 87–114, disc. 114–185. doi:10.1017/S0140525X01003922

Donkin, C., Little, D. R., & Houpt, J. W. (2014). Assessing the speed–accuracy trade-off effect on the capacity of information processing. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 1183–1202. doi:10.1037/a0035947

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. doi:10.3758/s13414-012-0330-z

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,* 1441–1463.

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010a). Converging measures of workload capacity. *Psychonomic Bulletin & Review, 17,* 763–771. doi:10.3758/PBR.17.6.763

Eidels, A., Townsend, J. T., & Algom, D. (2010b). 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. doi:10.3758/s13414-014-0788-y

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11,* 19–23. doi:10.1111/1467-8721.00160

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: Erlbaum.

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

Heathcote, A., Eidels, A., Houpt, J., Colman, J., Watson, J., & Strayer, D. (2014). Multi-tasking in working memory. In B. H. Ross (Ed.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 601–606). Austin: Cognitive Science Society.

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

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

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

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

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., 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. doi:10.3758/CABN.7.2.75

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, 105,* 6829–6833. doi:10.1073/pnas.0801268105

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

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

Jeffreys, H. (1961). *The theory of probability* (3rd ed.). Oxford: Oxford University Press.

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. doi:10.1016/j.jmp.2009.06.006

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

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. doi:10.1037/0278-7393.33.3.615

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. doi:10.1080/01621459.1995.10476572

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.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 817–835. doi:10.1037/0278-7393.27.3.817

Medeiros-Ward, N., Watson, J. M., & Strayer, D. L. (2014). On supertaskers and the neural basis of efficient multitasking. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-014-0713-3

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

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. doi:10.3758/BF03196573

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16,* 406–419. doi:10.1037/a0024377

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

Neufeld, R. W. J., Townsend, J. T., & Jetté, J. (2007). Quantitative response time technology for measuring cognitive-processing capacity in clinical studies. In R. W. J. Neufeld (Ed.), *Advances in clinical cognitive science: Formal modeling and assessment of processes and symptoms* (pp. 207–238). Washington, DC: 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. doi:10.1037/0278-7393.28.3.411

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. doi:10.1002/hbm.20131

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 111–164). Cambridge: Blackwell.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111,* 333–367. doi:10.1037/0033-295X.111.2.333

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. doi:10.1016/j.jmp.2012.08.001

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009a). 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. doi:10.1037/a0015730

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

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1–66. doi:10.1037/0033-295X.84.1.1

Shamosh, N. A., DeYoung, C. G., Green, A. E., Reis, D. L., Johnson, M. R., Conway, A. R. A., … 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). A comparison of laboratory and clinical working memory tests and their prediction of fluid intelligence. *Intelligence, 37,* 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. doi:10.1037/0033-295X.84.2.127

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31,* 137–149. doi:10.3758/BF03207704

Townsend, J. T., & Altieri, N. (2012). An accuracy–response time capacity assessment function that measures performance against standard parallel predictions. *Psychological Review, 119,* 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. doi:10.1006/jmps.1995.1033

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. doi:10.1037/0033-295X.111.4.1003

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28,* 127–154. doi:10.1016/0749-596X(89)90040-

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37,* 498–505. doi:10.3758/BF03192720

Von Der Heide, R. J., Wenger, M. J., Gilmore, R. O., & Elbich, D. B. (2011). Developmental changes in encoding and the capacity to process face information. *Journal of Vision, 11*(11), 450. doi:10.1167/11.11.450

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. doi:10.3758/BF03194105

Watson, J. M., & Strayer, D. L. (2010). Supertaskers: Profiles in extraordinary multitasking ability. *Psychonomic Bulletin & Review, 17,* 479–485. doi:10.3758/PBR.17.4.479

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. doi:10.1037/0096-1523.30.4.708

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. doi:10.1037/0096-1523.32.3.755

Wickens, C. D. (1980). The structure of attentional resources. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 239–257). Hillsdale: 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 or interactive race architectures. *Attention, Perception, & Psychophysics, 71,* 1739–1759. doi:10.3758/APP.71.8.1739