Performance on the processing portion of complex working memory span tasks is related to working memory capacity estimates

Lauren L. Richmond[1], Lois K. Burnett[1], Alexandra B. Morrison[2], B. Hunter Ball[3]

[1]Department of Psychology, Stony Brook University, Stony Brook, NY

[2]Department of Psychology, California State University, Sacramento, Sacramento, CA

[3] Department of Psychology, University of Texas at Arlington, Arlington, TX

Author Note:

Lauren L. Richmond and Lois K. Burnett, Department of Psychology, Stony Brook University. Alexandra B. Morrison, Department of Psychology, California State University, Sacramento. B. Hunter Ball, Department of Psychology, University of Texas at Arlington.

Correspondence concerning this article should be addressed to Lauren L. Richmond, Department of Psychology, Stony Brook University; Psychology B Building, Stony Brook, NY 11794; email: lauren.richmond@stonybrook.edu.

Abstract

Individual differences in working memory capacity (WMC) have long been known to relate to performance in domains outside of WM, including attentional control, long-term memory, problem solving, and fluid intelligence to name a few. Complex span WM tasks, composed of a processing component and a storage component, are often used to index WMC in these types of investigations. Capacity estimates are derived from performance on the storage component only, while processing performance is often largely ignored. Here, we explore the relationship between processing performance and WMC in a large dataset for each of three complex span tasks to better characterize how the components of these tasks might be related. We provide evidence that enforcing an 85% or better accuracy criterion for the processing portion of the task results in the removal of a disproportionate number of individuals exhibiting lower WMC estimates. We also find broad support for differences in processing task performance, characterized according to both accuracy and reaction time metrics, as a function of WMC. We suggest that researchers may want to include processing task performance measures, in addition to capacity estimates, in studies using complex span tasks to index WMC. This approach may better characterize the relationships between complex span task performance and performance in disparate domains of cognition.

Performance on the processing portion of complex working memory span tasks is related to working memory capacity estimates

Working memory (WM) is best characterized as a multi-purpose mental workspace (Baddeley & Hitch, 1974; Cowan, 1999) that varies across individuals in terms of capacity (WMC; Kane et al., 2007). Individual differences in WMC have consistently been shown to relate to performance in a wide number of other cognitive domains (e.g., dichotic listening, Conway et al., 2001; attentional control, Kane et al., 2001; stroop interference, Kane & Engle, 2003; reasoning, Kyllonen & Christal, 1990; mind wandering, McVay & Kane, 2009; modes of attentional control, Richmond et al., 2015; fluid intelligence, Unsworth, Brewer, et al., 2009). In this type of research, complex span WM tasks are a popular method of assessing WMC (e.g., Conway et al., 2005; Foster et al., 2015; Redick et al., 2012; Unsworth et al., 2005). In contrast to simple span tasks (e.g., digit span) that involve only a storage component, complex span tasks contain both a processing component that involves the presentation of stimuli requiring decisions and a storage component involving the presentation of memory items (Conway et al., 2005; Unsworth et al., 2005). An important feature of the processing task is that it is typically thought to momentarily reduce access to or block rehearsal of memory items (Unsworth et al., 2005), and performance on the processing task is emphasized in complex span task instructions in order to better capture individual differences in the storage component of the task (Conway et al., 2005).

A Brief History of Complex Span Task Development

Complex span tasks have undergone a number of design implementations since their initial instantiation by Daneman and Carpenter (1980). The initial task structure introduced by

Daneman and Carpenter (1980) was developed to address the relationship between WMC and reading comprehension that had been, to that point, only weakly observed. Previous studies used mainly simple span tasks that only taxed memory storage, and Daneman and Carpenter (1980) argued that the reason for the previously observed weak correlations was that simple span tasks did not appropriately tap the multiple components of the working memory system. In this view, simple span tasks are seen as indexing mainly short-term memory abilities. Short-term memory, in turn, may be thought of as just one sub-component of working memory, which additionally involves attentional control (see Engle et al., 1999 for an extended discussion) and other mechanisms such as a controlled search through long-term memory (Unsworth & Engle, 2007). Further, the authors argued that this complex span task would be both a better measure of WMC and a better measure of individual differences in WMC across participants (Daneman & Carpenter, 1980).

Based on the rationale outlined above, Daneman and Carpenter (1980) developed the earliest implementation of the now-classic complex span task (originally dubbed "Reading Span" and now commonly referred to as "Sentence Span", as the term "Reading Span" is now more typically used to describe the sentence verification/letter memory task; Redick et al., 2012). This task required participants to judge the veracity of sentences presented by an experimenter on index cards as the processing component, and to remember the final word of each sentence as the storage component (Daneman & Carpenter, 1980). Later, Turner and Engle (1989) replaced the sentences in Daneman and Carpenter's task with mathematical operations, thus creating the Operation Span task. Turner and Engle (1989) used this task to

demonstrate that reading comprehension could be predicted with a task that did not require reading as the processing component.

In a later publication that influenced the ubiquity of complex span tasks, Unsworth and colleagues (2005) developed an automated version of the operation span task (and later other complex span tasks) that allowed for automated computerized administration with minimal experimenter interaction. These tasks are now widely used in the literature, thanks to both the ease of administration afforded by the automated tasks and the ability of researchers to access the computerized versions of these tasks that have been made available for download by the Engle lab (see https://englelab.gatech.edu/taskdownloads). Together, these two features have contributed to cementing these tasks as mainstays in the working memory and cognitive individual differences literatures.

In recent years, shortened versions of the automated span tasks (Foster et al., 2015; Oswald et al., 2015), as well as advanced versions testing performance at larger set sizes (Draheim et al., 2018) have been introduced. These new task versions are expected to further increase the popularity and utility of complex span WM tasks. The scope of the present work will be limited to the "standard" automated versions of the Operation Span, Symmetry Span, and Reading Span tasks (Unsworth et al., 2005).

The Relationship between Processing and Storage

Although these complex span tasks necessarily contain both processing and storage, typical methods of calculating WMC estimates rely exclusively on performance in the storage component of the task. Adherence to an 85% processing accuracy criterion for inclusion in analyses has been recommended to ensure that participants sufficiently engage the processing

component (Conway et al., 2005). However, ensuring that participants' processing performance meets or exceeds this criterion is often the only consideration of this task component for scoring. As noted above, early methods for testing WMC using complex span tasks were not computerized and were completed in the presence of an experimenter. In these contexts, experimenters could ensure adequate engagement with the processing task and could help to correct participants' erroneous understanding regarding correct completion of the processing task. With the advent of more automated computerized methods, these tasks are often completed without such stringent oversight of the experimenter and therefore the 85% criterion was suggested to ensure appropriate levels of participant engagement with the processing portion of the task.

There is a relatively small body of literature focusing on investigating the relationship between processing and storage performance in complex span tasks directly. Such investigations have been conducted in healthy young adults (e.g., Engle et al., 1992; Friedman & Miyake, 2004; St Clair-Thompson, 2007a, 2007b; Towse et al., 2000; Unsworth et al., 2005; Waters & Caplan, 1996) and in typically developing children (e.g., Barrouillet & Camos, 2001; Hitch et al., 2001; St Clair-Thompson, 2007b; Towse et al., 1998). Additional work has examined the relationship between processing task accuracy and storage performance (Daneman & Tardif, 1987; Engle et al., 1992; Lépine et al., 2005; Salthouse et al., 2008; Shah & Miyake, 1996; Towse et al., 2000; Turner & Engle, 1989; Waters & Caplan, 1996). In both contexts, results have been somewhat mixed, with some reports finding evidence for a relationship between better processing performance indices (lower RT, higher accuracy) and better storage performance (e.g., St Clair-Thompson, 2007a, 2007b; Waters & Caplan, 1996) in young adult

samples. However, evidence for the opposite pattern in terms of processing RT has been observed in children (e.g., Towse et al., 1998)[1], as well as findings suggesting no relationship between processing and storage performance in younger adults (e.g., Engle et al., 1992; Shah & Miyake, 1996; Towse et al., 2000) and in children (Lépine et al., 2005). In sum, the extant literature does not yet offer a clear picture of the way(s) in which performance indices on processing and storage tasks interrelate. The present work aims to better characterize this relationship.

When introducing the automated version of the Operation Span task (OSpan), Unsworth and colleagues (2005) first tested the relationship between facets of processing time and storage performance. In the OSpan task, the average RT for processing stimulus presentation was found to be negatively related to storage accuracy (Unsworth et al., 2005). Therefore, those who were faster on the processing task exhibited higher WMC. Importantly, the sample in this study was restricted to participants exhibiting 85% processing accuracy or better, resulting in approximately 15% data loss (Unsworth et al., 2005). Following this initial examination, Unsworth, Redick, and colleagues (2009) tested the relationship between processing and storage performance without enforcing any processing accuracy criterion, as well as the relation of these variables to fluid intelligence scores. Unsworth and colleagues (2009) found that processing accuracy and processing RT factors were non-redundant with one another, and that each of these variables accounted for significant variance in fluid intelligence scores over and above that predicted by WMC storage scores. More central to the interests of

[1] Previous work has noted that relationships between processing RT and storage performance are more consistently observed in children compared to adult samples (Towse et al., 2010).

the current paper, higher processing accuracy and faster processing RTs were associated with better storage performance (i.e., WMC; Unsworth, Redick, et al., 2009). Building on this line of work, Unsworth, Fukuda, and colleagues (2014) tested the relationships between processing and storage performance and the relation of these factors to capacity, secondary memory, attentional control, and fluid intelligence. In this analysis, the negative relationship between processing time and storage accuracy that has been observed in prior work (Unsworth, Redick, et al., 2009) was replicated. Moreover, capacity, secondary memory, and attentional control were shown to fully account for the relationship between the WM indices (processing, storage) and fluid intelligence (Unsworth et al., 2014). Such findings provide initial evidence that consideration of both processing and storage together may be a worthwhile approach to characterize task performance. Importantly, more recent work using complex span tasks to characterize WMC appear to have abandoned strict adherence to an 85% processing accuracy criterion (see, for example, Ellis et al., 2020; McVay & Kane, 2009; Redick et al., 2011; Richmond et al., 2015; Unsworth et al., 2013). Đokić and colleagues (2018) recently suggested that eliminating the 85% accuracy criterion does not impact the psychometric properties of the tasks, and Unsworth, Redick, and colleagues (2009) suggested that enforcing the 85% accuracy criterion is unnecessary. However, previous research has not strongly recommended against enforcing this criterion, nor has a systematic analysis been undertaken to characterize the impact of enforcing this criterion on WMC estimates retained for inclusion in the final sample.

Advances in Investigating Task Reaction Time and Accuracy

The examination of trial-level variation in RTs, rather than characterizing RTs according to their mean, has recently gained traction. One such approach involves application of the ex-

Gaussian model to RT distributions. The ex-Gaussian model convolves the Gaussian and exponential distributions together, and are described by the parameters μ, σ, and τ. The μ parameter approximates the mode of the Gaussian distribution and the σ parameter approximates the standard deviation of the Gaussian distribution, whereas the τ parameter reflects the mean and standard deviation of the exponential component of the distribution (Balota & Yap, 2011). Because the sum of μ and τ are roughly equal to the mean RT, any variable that results in an increase in τ accompanied by a decrease in μ (or vice versa) would result in a null effect at the level of mean RT, but can be easily observed with ex-Gaussian RT characterization (e.g., Ball & Brewer, 2018; Balota et al., 2008; Spieler et al., 1996). This work suggests that the application of the ex-Gaussian model to RT data can therefore reveal effects that would be masked by simply characterizing RT according to mean performance.

Similarly, characterization of task performance according to error types, rather than or in addition to overall accuracy, has gained popularity in recent years (e.g., Giovannetti et al., 2008; Scullin et al., 2012, 2020). For example, recent work by Giovannetti and colleagues in the domain of naturalistic action execution has shown that omissions in the context of a performance-based measure of everyday action are closely related to performance on tests of episodic memory, whereas commission errors are more closely associated with deficits in executive functioning (Devlin et al., 2014). This suggests that characterizing performance according to error types may provide a more detailed analysis of participant performance than simple accuracy measures alone, and this profile may map meaningfully to other domains of cognition.

The current work follows these two recent trends to provide a more detailed examination of the relationship between processing and storage performance in complex span tasks. Here, we consider four novel research questions in each of three complex span tasks. First, we address the appropriateness of the aforementioned inclusion criterion by asking (1) Does enforcing the recommended 85% processing accuracy criterion for inclusion result in the removal of more participants with low WMC estimates compared to higher WMC estimates? The next three questions investigate whether and how specific aspects of processing performance relate to WMC. (2) Do RT means and standard deviations of RTs from the processing practice portion of the task differ as a function of WMC?; (3) Do task-derived measures of processing RT, including mean RT and ex-Gaussian parameters, and processing task error profiles relate significantly to WMC?; and (4) Does modeling RT-based and error-based processing profiles together explain more variance in WMC than consideration of either alone?

Method

Operation Span (OSpan) and Symmetry Span (SymSpan) data were collected between 2011-2019 at four large state universities: Arizona State University (ASU), California State University, Sacramento (CSUS), Stony Brook University (SBU), and Temple University (TU). Reading Span (RSpan) data were collected at ASU only. All data were collected in the context of task batteries for large-scale projects. Study procedures were approved by the Institutional Review Board of each institution.

Operation Span

In OSpan, participants alternated between solving simple math problems as the processing component and remembering letters as the storage component. Participants started

out practicing each portion of the task separately - first practicing the letter memory (storage portion) task, then practicing the math (processing) portion of the task, and last practicing alternating between solving math problems and remembering letters (comparable to test trials). See Figure 1 panel A for a task schematic.

Insert Figure 1 about here

Letters were displayed for 1000 ms each in all phases of the experiment. At the recall phase, participants were shown a grid displaying 12 possible letters with a box beside each letter. Participants were told to recall the letters in the order they were presented; the chosen letters were displayed at the bottom of the screen. Participants were instructed to use the 'clear' button displayed on the screen if they made a mistake and wanted to start over. The blank button was displayed on the screen to mark the position of a forgotten letter, and participants were instructed to click the enter button displayed on the screen when they were ready to submit their response. Participants were given as long as they needed to complete the recall phase in all trials.

For the math problems, a simple arithmetic problem such as "(6*0) + 1 = ?" was displayed, and participants were told to solve the problem as quickly as possible without sacrificing accuracy. Once participants had an answer in mind, they were instructed to click to advance to the next screen. On this screen, a number is displayed at the top of the screen; displayed below this number is a box marked 'true' and another marked 'false'. If the number shown was the correct response to the math problem, the participant was instructed to choose

the 'true' button; else, they were instructed to click 'false'. During the processing practice portion, mean and standard deviation (SD) of the RTs for the problem display screen were calculated, and then an upper limit bound was set for solving math problems in the test context by taking each individual's average RT and adding 2.5 standard deviations to that number (Unsworth et al., 2005). This serves as the maximum allowable response time for clicking to advance from the problem display to the true/false screen. On trials for which participants did not click before the maximum time was reached, the trial was marked as a 'time out' error. Regardless of RT on the problem screen, participants were given unlimited time to respond on the true/false screen and the accuracy of the response was recorded.

The OSpan task consisted of 15 trials, with three trials each at set size ranging from 3-7. Set sizes were presented randomly for each participant. In total, for the test phase, participants solved 75 math problems and were shown 75 letters. The capacity score for this task was the number of letters recalled in the correct position (with 75 being the maximum possible score).

Reading Span

The RSpan task is similar to OSpan, save for differences in processing task demands. In RSpan, participants alternated between reading sentences and judging whether they made sense as the processing component and remembering letters as the storage component. The practice phase proceeded as described above in OSpan. See Figure 1 panel B for a task schematic.

The storage component of the task was exactly the same as in OSpan, described above. The RSpan processing task involved sentence verification. Here, a simple sentence ranging in length from 10 to 15 words was displayed, and participants were told to make a judgment as

quickly as possible, without sacrificing accuracy, regarding whether the sentence made sense or not. “Nonsense” sentences were created by replacing one word in the sentence (e.g. “The young pencil kept his eyes closed until he was told to look.”). Once participants had an answer in mind, they were instructed to click to advance to the next screen. On this screen, participants saw a box marked ‘true’ and another marked ‘false’. If the sentence displayed on the previous screen made sense, the participant was instructed that they should choose the ‘true’ button; else, they should click ‘false’. During the processing practice portion, mean and SD for the RTs on the sentence verification screen were calculated, and then an upper limit bound was set for making sentence judgments in the test context by taking each individual’s average RT and adding 2.5 standard deviations to that number. This served as the maximum allowable response time for clicking to advance from the sentence display to the true/false screen. On trials for which participants did not click before the maximum time was reached, the trial was terminated and counted as a time out error. Participants were given unlimited time to respond on the true/false screen.

The RSpan task consisted of 15 trials, with three trials each at set sizes ranging from 3-7. Set sizes were presented randomly for each participant. In total, for the test phase, participants completed 75 sentence judgments and were displayed 75 letters. The capacity score for this task was the number of letters recalled in the correct position (with 75 being the maximum possible score).

Symmetry Span

In SymSpan, participants alternated between making symmetry judgments as the processing component and remembering highlighted locations as the storage component. The

practice portion of the task proceeded as described above. See Figure 1 panel C for a task schematic.

Locations were shown in a 4x4 grid with one square of the grid highlighted in red. Locations were displayed for 650 ms each in all phases of the experiment. At the recall phase, participants were displayed a blank 4x4 grid. Participants were told to recall the locations in the order they were presented by clicking on each location; the chosen locations were numbered in the grid. Again, participants had access to 'clear', 'blank', and 'enter' buttons and were given as much time as needed to complete the recall phase.

For the symmetry judgments, participants were displayed an 8x8 black and white grid and were asked to make a symmetry judgment about the vertical axis. Participants were instructed to solve the symmetry problem as quickly as possible without sacrificing accuracy. Once participants had an answer in mind, they were instructed to click to advance to the next screen. On this screen, participants were instructed to respond 'true' if the grid displayed on the previous screen was symmetrical and 'false' if it was not. The processing task response deadline was computed as described above (RT mean + 2.5 SDs) for the processing screen, and unlimited time was allowed on the true/false screen. When participants did not submit a response on the processing before the maximum time was reached, the trial was terminated and counted as a time out error.

The SymSpan task consisted of 12 trials, with three trials each at set size ranging from 2-5. Set sizes were presented randomly for each participant. In total, for the test phase, participants solved 42 symmetry problems and were shown 42 to-be-remembered locations.

The capacity score for this task was the number of locations recalled in the correct order (with 42 being the maximum possible score).

Participants

Data were collected primarily from participant pools at each respective institution, consisting of undergraduate students enrolled in psychology courses who were participating in experiments for course credit and/or payment. Paid participants were also recruited through flyer advertisements and word of mouth (i.e., paid participants were not required to be registered with the university subject pool in order to participate)[2].

Although these data were collected across a number of different sites, study designs were relatively similar across sites. In all studies, participants completed sessions that were between 1 to 2 hours in length, and task batteries were completed over 1 or 2 sessions. All data reported here were collected in the context of larger studies that included a variety of other tasks in addition to the WMC measures. Participants were aged at least 18 years and provided informed consent for their participation in each study. Sample sizes and descriptive statistics for the capacity estimates in each task are displayed in Table 1, separated by site. Descriptive statistics for storage performance, processing accuracy, and processing RT are displayed in Table 2. Cronbach's alphas for each task for both processing and storage components are displayed in Table 3.

Insert Tables 1, 2, & 3 about here

[2] All analyses were also conducted without the inclusion of paid participants and the pattern of results was found to be the same as the results of the full sample reported here.

Procedure & Statistical Approach

Significance criterion for all statistical tests was set to the $p \leq .01$ level. Statistical analyses were conducted in R (R Core Team, 2008) using the 'stats' package. Cronbach's alphas were computed using the alpha function from the 'psych' package (Revelle,2018). Ex-gaussian RT distributional components were calculated using QMPE software (Heathcote et al., 2004) and were imported into R for analysis. Where appropriate, Cohen's d effect sizes were computed using the 'lsr' package (Navarro, 2015). Plots were created in R using the 'ggplot' package (Wickham, 2016). For variables where evidence of non-normality was observed, nonparametric statistical tests were conducted. For brevity, only parametric results are reported, with differences from the non-parametric test results footnoted. Tests of normality and non-parametric test results are provided in full in the Supplemental Materials.

Importantly, popularly used automated complex span tasks (e.g., Unsworth et al., 2005) enforce a response deadline for the processing component of the task (mean RT from processing practice performance + 2.5 standard deviations from that mean). Therefore, we examine RT data from both the practice component (where no response deadline is enforced) in question 2 as well as from the task itself in questions 3 and 4. The response deadline applies only to the screen on which the processing task itself (math problem, symmetry grid, sentence reading) is presented, and not the following screen where a response is input. After participants have been alerted to the response deadline in the context of the task proper, it is possible for savvy participants to then 'game' the system by moving on from the processing task screen to the response screen and then lingering on this screen while continuing to think. Therefore, in

addressing research questions 3 and 4 we used total RTs for both the RT exhibited on the response screen in the processing component as well as RT on the processing screen itself. All analyses focusing on processing time are based on RTs derived from correct trials only.

To examine processing performance, we characterized errors according to two types. For trials on which participants failed to move on to the response screen before their individualized response deadline, these were counted as 'time out' errors (regardless of the response rendered on the response screen). Overtly incorrect responses (i.e., for the operation "2 x 3 + 5?" and a response screen displaying a value of "12," choosing the "true" box would count as an error) regardless of participant RTs were characterized as 'incorrect' errors.

Capacity estimates were examined by awarding credit for each to-be-remembered item recalled in the correct position, summed over the entire task (Unsworth et al., 2005).

Results

The results for each of our substantive research questions are reported below, separated by sub-headings.

Are WMC estimates significantly lower for individuals who do not meet the 85% processing criterion cutoff than for those that do?

This question was tested with a two-tailed Welch's two-sample t-test for unequal variances given the different sample sizes for those who missed versus met/exceeded the criterion cutoff. For OSpan, we observe a significant difference in WMC estimates by processing performance (t (206.73) = 10.74, p < .001, 95% CI [10.60, 15.37], d = 0.99). This pattern replicates in both SymSpan (t (247.82) = 10.75, p < .001, 95% CI [5.67, 8.22], d = 0.94) and RSpan (t (163.60) = 7.97, p < .001, 95% CI [9.62, 15.95], d = 0.89; see Figure 2 for visualization of

these data). Together, these results strongly suggest that enforcing a processing accuracy criterion for inclusion in the final dataset results in the removal of a greater number of individuals with low WMC compared to those who achieve high WMC estimates.

Insert Figure 2 about here

Do individuals who display higher mean RTs and/or more variable RTs (i.e., higher RT SDs) in the practice phase for the processing task also exhibit lower WMC estimates?

We tested these questions using two-tailed Pearson's correlations between (a) an individual's mean RT during the processing practice task and WMC, and (b) an individual's RT SD exhibited during the processing practice task and WMC. We observe a small relationship between mean RT during the processing practice task and WMC that was nonetheless consistent across OSpan and SymSpan. This effect was not observed for RSpan. For OSpan, there is a small but significant negative correlation between mean processing practice RT and WMC (r (1683) = -0.09, p < .001, 95% CI [-0.13, -0.04]), but not between RT SDs and WMC (r (1683) = -0.03, p = .256, 95% CI [-0.08, 0.02]). This pattern replicates in SymSpan (mean practice RT: r (1051) = -0.13, p < .001, 95% CI [-0.18, -0.07]; SD practice RT: r (1051) = -0.05, p = .142, 95% CI [-0.11, 0.02])[3]. However, deviation from this pattern is observed in RSpan, where we observe non-significant relationships between RT indices and WMC estimates (mean practice RT: r (1060) = -0.02, p = 0.436, 95% CI [-0.08, 0.04]; SD practice RT: r (1060) = -0.07, p = 0.015,

[3] The Spearman correlation for the relationship between SymSpan WMC and processing practice RT SDs reaches significance: r_s (1051) = - 0.10, p < .001.

95% CI [-0.13, -0.01]). Overall, as indicated by small correlations, evidence for differences in RTs during the processing practice phase as they relate to WMC is weak (Sawilowsky, 2009), with slightly stronger support for mean RT differences by WMC estimates compared to RT SDs from the practice portion of the task. Given the small correlations and the large sample size, however, the relationship between RTs on the practice processing task and WMC is of little practical significance.

Do task-derived measures of processing RT, including mean RT and ex-Gaussian parameters, and processing task error profiles relate significantly to WMC?

To mirror the strategy employed by Unsworth and colleagues (Unsworth et al., 2005; Unsworth, Redick, et al., 2009) in which measures of central tendency were used to characterize processing RT, a regression model with mean task-derived processing RT entered as a predictor and WMC as the outcome was built separately for each task (OSpan, SymSpan, and RSpan). Significant models predicting WMC estimates with mean RT derived from the processing task were observed consistently across all three tasks, though variance explained was small; OSpan (R^2 = .019, R^2 adjusted= .018, F (1, 1683) = 31.78, p < .001), SymSpan (R^2 = .050, R^2 adjusted= .049, F (1, 1051) = 54.82, p < .001), and RSpan (R^2 = .007, R^2 adjusted= .006, F (1, 1060) = 7.17, p = .008).

Next, we characterized the distribution of task-derived RTs using an ex-Gaussian approach and examined the significance of each predictor in these models as well as overall model fit for each task. Vincentile plots were created for each task to examine the overlap between predicted and observed values derived from the ex-Gaussian model. Vincentiles were created by rank-ordering raw RTs from fastest to slowest for each individual and calculating the

mean RT for the first 20% of RTs, the next 20%, and so on. The substantial overlap between predicted and observed values for the top third, middle third, and bottom third of participants according to WMC depicted in Figure 3 suggests that the ex-Gaussian model provided a good fit for our RT data. Vincentile plots for our entire sample and for only those participants who had 35 or more correct RTs (due to concerns over the appropriateness of the ex-Gaussian model for fitting a small number of RTs) can be found in the Supplemental Materials.

Insert Figure 3 about here

Across all three tasks, we find significant overall models which explain between 2.8% and 4.9% of the variance in WMC estimates. We also observe some inconsistency in terms of the significance of individual ex-Gaussian predictors. In OSpan, we observe an overall significant model (R^2 = .031, R^2 adjusted= .030, F (3, 1681) = 18.10, p < .001). The σ component emerged as the only significant predictor in the model (β= -0.188, t (1681) = -4.28, p < .001; μ and τ p values > .028). Similarly, an overall significant model is observed in RSpan (R^2 = .031, R^2 adjusted= .028, F (3, 1058) = 11.30, p < .001), and the σ (β= -0.188, t (1058) = -5.16, p < .001) component again emerged as the only significant predictor in the model (μ and τ p values > .10). In SymSpan, we replicate the overall significance of the model (R^2 = .052, R^2 adjusted= .049, F (3, 1049) = 19.16, p < .001), and the τ (β= -0.104, t (1049) = -3.38, p < .001) component emerged as the only significant predictor in the model (μ and σ p values > .10). In general, we find consistent support for the use of ex-Gaussian analyses to characterize task-derived RTs in

complex span tasks, but the contributions of individual components to overall predictive power varies across tasks[4].

To test whether the ex-Gaussian approach provides an advantage over characterizing RTs by only mean performance, the models described above were tested against one another. In OSpan (F (2, 1681) = 11.075, p < .001, ΔAIC = 18.06) and RSpan (F (2, 1058) = 13.28, p < .001, ΔAIC = 22.33) we observed significant improvements in model fit for the ex-Gaussian models over mean RT models[5]. There was no difference between models for the SymSpan task (F (2, 1049) =1.31, p = .270, ΔAIC = 1.37). In summary, for two of the three tasks (OSpan, RSpan), we observed improvements in model fit when using an ex-Gaussian approach compared to the more common mean RT approach. Together, these data provide support for characterizing RT data in complex span tasks using an ex-Gaussian approach.

Last, we explored whether characterizing processing errors according to their type would predict WMC in each task context by running a series of simultaneous multiple

[4] It has been suggested that ex-Gaussian approaches may produce poor model fits when the number of trials for inclusion in the model are small (Heathcote et al., 2004). In order to ensure that the inclusion of participants with few correct RT trials were not skewing our results, we assessed the significance of the ex-Gaussian model using only participants with 35 or more correct RTs. Removal of those with very few RTs changed neither the directionality nor the significance of results for OSpan nor RSpan. Following the removal of participants with few RTs in the SymSpan task, the overall model remained significant, but in this model only the μ ($\beta = -0.003, t(913) = -3.75, p < .001$) component emerged as a significant predictor. Vincentile plots for datasets including those with only 35 or more correct RT trials can be found in the Supplemental Materials.

[5] To compare the non-nested models reported here, we report ΔAIC to provide information on the best fitting model. AIC (Akaike, 1985, 1973) is a penalized likelihood model that is based on the number of estimated regression parameters. Using the relative AIC values for the candidate models being compared, one can select the best model from the set and determine whether the others provide good estimates of the observed data. Burnham and Anderson (2002) provide some general rules of thumb for estimating the level of empirical support for competing models (compared to the best-fitting model) on the basis of ΔAIC, with ΔAIC between 0 and 2 indicating substantial support, between 4 and 10 with less support, and greater than 10 indicating little to no empirical support. The authors note that these rules of thumb are generally applied to nested models and that guidelines values may be larger for non-nested models. Although there is support for the use of AIC to compare non-nested models (Burnham & Anderson, 2002), we also note that some argue against the use of AIC for selecting between non-nested models (Ripley, 2004).

regression models entering error types (time out, incorrect) as predictors and WMC as the outcome variable. Results were consistent across all three complex span tasks for both overall model fit and significance of individual predictors (time out errors and incorrect response errors). The overall model was significant for OSpan (R^2 = .139, R^2 adjusted= .138, F (2, 1682) = 135.40, p < .001), and both error types emerged as significant predictors (time out: β= -0.151, t (1682) = -6.63, p < .001; incorrect: β= -0.323, t (1682) = -14.17, p < .001) in this model. Similar results were observed for SymSpan (overall model: R^2 = .153, R^2 adjusted= .151, F (2, 1050) = 94.69, p < .001; time out: β= -0.191, t (1050) = -6.66, p < .001; incorrect: β= -0.313, t (1050) = -10.88, p < .001) and RSpan (overall model: R^2 = .119, R^2 adjusted= .118, F (2, 1059) = 71.61, p < .001; time out: β= -0.196, t (1059) = -6.75, p < .001; incorrect: β= -0.262, t (1059) = -9.04, p < .001). Across all task contexts, results indicate moderate associations between processing error profiles and WMC estimates.

Does modeling processing RT and processing errors together explain more variance in WMC than either alone?

RT components are derived only from correct trials, so error types add a non-overlapping piece of information about processing performance. Here, we compared model fits for regression models including RT components (μ, σ, τ) only (hereafter called the RT distribution model) and a model including error types as predictors (hereafter referred to as the error model) to a model containing both RT distributional components and error types (the RT distribution + error model). In OSpan, the RT distribution + error model explained significantly more variance in WMC estimates compared to the RT distribution model (model comparison: F (2, 1679) = 130.49, p < .001, ΔR^2 = 0.130) and the error model (model comparison: F (3, 1679) =

15.34, $p < .001$, $\Delta R^2 = 0.023$). This pattern is replicated in SymSpan (comparison to the RT distribution model: (F (2, 1047) = 95.24, $p < .001$, $\Delta R^2 = 0.146$); comparison to the error model: F (3, 1047) = 19.61, $p < .001$, $\Delta R^2 = 0.045$), and RSpan (comparison to the RT distribution model: F (2, 1056) = 59.91, $p < .001$, $\Delta R^2 = 0.099$); comparison to the error model: F (3, 1056) = 4.31, $p = .005$, $\Delta R^2 = 0.011$)[6]. While the increase in explanatory power as a result of using the RT distribution + error model varied across tasks, using the combined ex-Gaussian parameters and error types as predictors provided notable improvements in terms of variance explained. Compared to the RT distribution (only) model, the RT distribution + error model explained between 4.5% and 14.6% of the variance in WMC estimates. Compared to the error (only) model, the RT distribution + error model explained between 1% and 2.3% of the variance in WMC estimates. Across all task contexts, the full model provided a significantly improved model fit compared to models using either the ex-Gaussian RT distribution components or error profiles alone. Taken together, these data provide support for the joint use of RT-based and error-based metrics to characterize processing task performance in relation to WMC estimates.

Discussion

For our substantive research questions, we observed a fair degree of consistency across the three complex span tasks, as well as some points of departure. Below, we discuss each in terms of consistency across tasks and size of the observed effects.

First, we see strong evidence in all three tasks that WMC estimates are significantly lower for those that miss the 85% processing task accuracy cutoff than those who meet or

[6] The comparison for the quantile regression RT distribution + error model to quantile regression error-only model in RSpan does not meet our $p < .01$ significance criterion (model comparison significance: $p = .035$).

exceed the cutoff. Notably, the intention of this cutoff is to remove those who do not sufficiently engage the processing task (Unsworth et al., 2005), and was not devised to eliminate truly low performers. In the current datasets, enforcement of this criterion results in the removal of 10-17% of the full sample, and in Unsworth et al., 2005 approximately 15% data loss is reported when this criterion is enforced. Our findings suggest that enforcing this criterion will result in the removal of a disproportionate number of individuals exhibiting lower WMC estimates.

We observed some inconsistency across tasks for our second question. There were no significant associations observed between processing task RT SDs and WMC. Significant negative correlations between mean processing practice RT and WMC were observed in two of the three tasks (OSpan, SymSpan). Inconsistency by task is perhaps not surprising, given the small number of trials and the relatively simple nature of processing task practice. Moreover, these findings should be interpreted with caution as, in all cases, the strength of the correlation value was low.

Turning to our third question, in general we find significant relationships between RT (characterized by mean RT and ex-Gaussian components) and WMC. For both OSpan and RSpan, the ex-Gaussian model fit significantly better than the mean RT model. There were no differences for SymSpan. This suggests ex-Gaussian analyses are no worse than measures of central tendency for characterizing RTs, and are actually more informative for two of the three tasks. Under the ex-Gaussian analyses, the specific components that were significant differed across tasks (σ for OSpan and RSpan, τ for SymSpan when all participants were included for analysis and μ when only participants with 35 or more correct RTs were retained for modeling).

It is somewhat surprising that the τ component was only found to be a significant predictor in one of the three tasks. However, we note that trials reaching the response deadline were characterized as time-out errors, so the extent to which effects of τ could be observed may have been limited in this context. In support of this view, we saw that the significant contribution of τ to the overall model for SymSpan was eliminated when errors were modeled together with ex-Gaussian parameters, suggesting that τ and error profiles may be accounting for overlapping variance in WMC. It is possible that eliminating or extending the response deadline for each trial may lead to more consistency in terms of the contribution of the τ component of the ex-Gaussian parameters by capturing responses that exceed the cut-off to be classified as time-out errors. However, researchers may be reluctant to alter the standard RT deadline of 2.5 times the mean as it is intended to prevent participants from rehearsing items from the storage portion of the task when they should be completing the processing component (Unsworth et al., 2005).

Future work using other analytic techniques for characterizing RTs, such as diffusion modeling (Ratcliff, 1978), may further elucidate these relationships by offering a straightforward link between parameter estimates and cognitive processes. Unfortunately, concerns over an inadequate number of RTs available in the current tasks precluded the inclusion of this analysis here (Lerche et al., 2017). Researchers who wish to pursue such a characterization in future studies are encouraged to explore methods by which to increase the number of RTs available in the context of WMC tasks for the application of this model, perhaps by utilizing the advanced complex span tasks with larger set sizes (Draheim et al., 2018), increasing the number of cycles through each set size, or by loosening or removing the response

deadline for the processing task in order to obtain usable RT data for long response trials (that are currently captured instead as time-out errors under standard task conditions). Nonetheless, across all tasks, we observed predictive power for processing error profiles, with both time-out errors and incorrect errors emerging as significant predictors of WMC.

Last, in all tasks, results provide consistent evidence that the RT distribution + error models provided better fits than models containing only error information and models containing only RT distributional information. In terms of variance explained, the RT distribution + error model increased variance explained by 4.5% to 14.6% compared to the RT distribution model. Increases in variance explained compared to the error model were smaller (between ~1-2%). While the observed effects are small for some of the tasks, inclusion of both ex-Gaussian RT components and error profiles together consistently improves the explanatory power of the models. Future research may consider the specific relationships between each processing component (RT, accuracy) and other tasks sensitive to cognitive individual differences. For instance, it is possible that ex-Gaussian RT components may be more related to tasks that emphasize speeded responding whereas processing accuracy may be more related to tasks that don't require speeded responses (Unsworth, Redick, et al., 2009).

The current work dovetails with prior work regarding individual differences in WM and extends this work in some important ways. Previous work has sought to characterize individual differences in WMC and the relation of those differences to other cognitive factors of interest, including processing speed (Conway et al., 2002). WMC and processing speed are generally found to be only weakly related to one another in samples of healthy adults (see e.g., Conway et al., 2002). WMC is often observed as a better predictor of higher-order cognitive functions

such as reasoning (Kyllonen & Christal, 1990) and fluid intelligence (Conway et al., 2002) compared to processing speed measures in healthy adult samples, despite in some cases observing strong relationships between processing speed and WMC (Kyllonen & Christal, 1990, but see Conway et al., 2002). The relationship between processing speed and WMC is more strongly observed in early development (Kail, 2007) and in late life (Brown et al., 2012). Nonetheless, consideration of RT-based metrics that can be derived from complex span tasks themselves may provide a more fruitful way to characterize relationships between WMC and speed. In this regard, future work to compare the strength of the relationships observed between WMC-storage measures derived from complex span tasks and RT metrics derived from traditional processing speed tasks and from the processing portion of the complex span WM task itself would be useful.

Reflecting on the present findings, we make some recommendations for future WM research using these tasks. First, we caution against use of an 85% processing accuracy criterion as it may inadvertently bias WMC estimates against lower capacity participants, and if adherence to the cutoff is maintained researchers should be aware that this is likely to result in skewed estimates of WMC in their samples. In considering whether to forgo enforcing this criterion, researchers should assess whether their sample sizes are sufficiently powered to tolerate approximately 10-17% data loss. Researchers may instead choose to adopt a criterion closer to 50% (where performance worse than 50% likely represents misunderstanding of or insufficient engagement with the processing task) or to embed attention checks in the task in order to justify the removal of participants' data. If researchers wanted to move beyond the simple 50% accuracy criterion for processing task performance and instead include a cutoff for

ensuring that participants were included in the final were 95% or 99% likely to be above guessing probability, this could be easily achieved and data below the cut score could be discarded on either a task-wise or trial-wise basis. At the task level, to achieve 95% confidence that participants weren't guessing on a task with 42 processing trials (SymSpan) the criterion should be set to 61.9% overall processing accuracy. For a task with 75 processing steps (Ospan, RSpan), a cut score below 69% accuracy should be adopted under the 95% confidence criterion[7].

Alternatively, researchers may choose to adopt a data-driven approach to set a processing task performance threshold for inclusion in the final dataset based on their own sample. For example, participants who are found to exhibit processing task performance 2 or 2.5 SDs below the mean processing performance in that sample could be excluded from the final dataset. In cases where participants complete more than 1 complex span task, z-scored processing task performance could be computed for each participant across tasks, similar to the typical approach for combining WMC scores across tasks (see Morrison & Richmond, 2020; Redick et al., 2011; Richmond et al., 2015; Shipstead & Broadway, 2013 for examples of this approach). From here, participants exhibiting z-scores equal to or less than -2 or -2.5 could be removed from the final sample. This approach could perhaps be used in combination with a criterion for acceptable lower-bound RTs displayed on the processing task screen itself (e.g., RTs shorter than 200 ms). These approaches may be particularly useful in samples that exhibit higher average performance on complex span tasks (see for example Redick et al., 2012 Table 5 showing differences in WMC estimates by data collection site).

[7] Thank you to an anonymous reviewer for suggesting this approach.

At the same time, we acknowledge that simply reducing or abandoning a criterion for inclusion on the processing task is not expected to fully eliminate issues regarding inclusion of problematic data. Instead, this approach is expected to improve the retention of data for engaged participants at the lower end of the WMC spectrum. In other words, it is possible that enforcing a less stringent processing accuracy criterion level could allow for the inclusion of a small number of participants who fail to adequately engage with the processing task (i.e., faux lows) in the final dataset. More importantly, forgoing strict adherence to the 85% accuracy criterion is expected to allow for the inclusion of an interesting and important sub-set of individuals with lower working memory capacity estimates. Approaches such as those described above (enforcing an accuracy criterion at chance levels, using a data driven approach to setting a criterion for inclusion, setting a lower-bound of acceptable processing RTs for inclusion) are expected to minimize the number of participants included who truly failed to engage with the processing task, but also maximize the number of participants retained for analysis who were engaged in the task and simply struggled with both processing and storage components of the task at hand. Moreover, we note that the traditional criterion for inclusion only applies to the processing portion of the task, and there is no lower bound accuracy metric enforced for performance on the storage portion. To separate out truly disengaged participants from engaged participants with lower WMC, setting a criterion for acceptable storage scores (e.g., at least one correct trial at the lowest set size), in addition to adopting some or all of the practices we recommend above, could prove useful in future studies.

Another important avenue for future research will be to examine if the current findings extend to different complex span task configurations and design choices. For example, in the

complex span tasks used in the current research, the processing task itself was designed to disrupt active maintenance of information in working memory through rehearsal (Conway et al., 2005). Previous work examining complex span tasks including semantically related information for both the processing and storage components (Towse et al., 2010) suggests that the reinstatement of context at processing could serve to boost, rather than disrupt, storage performance (see e.g., Delaney & Sahakyan, 2007; Wahlheim et al., 2016, 2017; Wahlheim & Huff, 2015; Wingfield & Kahana, 2002 for further discussions of contextual reinstatement and memory performance). It is as yet unknown whether the patterns observed in the current study would extend to conditions where the processing task is designed to support, rather than disrupt, access to the to-be-remembered information. In considering other task configurations, it has recently been suggested that the order of the storage and processing components can impact estimates of WMC, with the processing-storage sequence resulting in higher estimates of WMC compared to storage-processing (Debraise et al., 2020). Future work may explore whether the relationships between processing and storage task components observed in the current work (using the standard processing-storage sequence) would extend to complex span tasks with a storage-processing sequence.

Finally, we also encourage those using complex span tasks in their own work to examine processing performance more thoroughly. The advantages of this approach are twofold: (1) more efficient use of collected data, and (2) task length could potentially be reduced while still obtaining stable estimates of WMC. Large-scale individual differences studies should consider processing RT, processing error types, and storage together in relation to other cognitive constructs (see also Unsworth, Redick, et al., 2009). Inclusion of both storage and processing

components is expected to be informative and may reveal subtle relationships between component processes embedded in complex span tasks and other cognitive domains. Overall, the present work provides strong support for careful consideration of processing performance indices, in addition to storage performance, in the context of complex span tasks.

Open Practices Statement

Analysis scripts, aggregate data files, PDF outputs, and Supplemental Materials are available at https://osf.io/85yfe/?view_only=204c6d7d4ba24b1d80ce145675960f21. [*Please note:* This link will be replaced with a publicly accessible link after acceptance.] The analyses reported in this paper were not preregistered.

References

Akaike, H. (1985). Prediction and Entropy. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 387–410). Springer. https://doi.org/10.1007/978-1-4612-1694-0_30

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings of the 2nd International Symposium on Information Theory* (pp. 267–281). Akadémiai Kiadó.

Baddeley, A., & Hitch, G. (1974). Working memory. In G. Bower (Ed.), *Recent advances in learning and motivation* (pp. 47–89). Academic Press.

Ball, B. H., & Brewer, G. A. (2018). Proactive control processes in event-based prospective memory: Evidence from intraindividual variability and ex-Gaussian analyses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(5), 793–811. https://doi.org/10.1037/xlm0000489

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry: The power of response time distributional analyses. *Current Directions in Psychological Science*, *20*(3), 160–166. https://doi.org/10.1177/0963721411408885

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language*, *59*(4), 495–523. https://doi.org/10.1016/j.jml.2007.10.004

Barrouillet, P., & Camos, V. (2001). Developmental increase in working memory span: Resource sharing or temporal decay? *Journal of Memory and Language*, *45*(1), 1–20. https://doi.org/10.1006/jmla.2001.2767

Brown, L. A., Brockmole, J. R., Gow, A. J., & Deary, I. J. (2012). Processing speed and visuospatial executive function predict visual working memory ability in older adults. *Experimental Aging Research*, *38*(1), 1–19. https://doi.org/10.1080/0361073X.2012.636722

Burnham, K. P., & Anderson, D. R. (2002). A practical information-theoretic approach. *Model Selection and Multimodel Inference*, *2*.

Conway, A., Cowan, N., & Bunting, M. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, *8*, 331–335.

Conway, A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183.

Conway, A., Kane, M. J., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

Cowan, N. (1999). An embedded process model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge University Press.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Verbal Learning and Verbal Behavior*, *19*(4), 450–466.

Daneman, M., & Tardif, T. (1987). Working memory and reading skill re-examined. In *Attention and performance 12: The psychology of reading.* (pp. 491–508). Lawrence Erlbaum Associates, Inc.

Delaney, P., & Sahakyan, L. (2007). Unexpected costs of high working memory capacity following directed forgetting and contextual change manipulations. *Memory & Cognition*, *35*(5), 1074–1082.

Devlin, K. N., Giovannetti, T., Kessler, R. K., & Fanning, M. J. (2014). Commissions and omissions are dissociable aspects of everyday action impairment in schizophrenia. *Journal of the International Neuropsychological Society*, *20*(08), 812–821. https://doi.org/10.1017/S1355617714000654

Đokić, R., Koso-Drljević, M., & Đapo, N. (2018). Working memory span tasks: Group administration and omitting accuracy criterion do not change metric characteristics. *PLOS ONE*, *13*(10), e0205169. https://doi.org/10.1371/journal.pone.0205169

Draheim, C., Harrison, T. L., Embretson, S. E., & Engle, R. W. (2018). What item response theory can tell us about the complex span tasks. *Psychological Assessment*, *30*(1), 116–129. https://doi.org/10.1037/pas0000444

Ellis, D. M., Ball, B. H., Kimpton, N., & Brewer, G. A. (2020). The role of working memory capacity in analytic and multiply-constrained problem-solving in demanding situations. *Quarterly Journal of Experimental Psychology*, *73*(6), 920–928. https://doi.org/10.1177/1747021820909703

Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(5), 972–992. https://doi.org/10.1037/0278-7393.18.5.972

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309.

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226–236. https://doi.org/10.3758/s13421-014-0461-7

Friedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory and Language*, *51*(1), 136–158. https://doi.org/10.1016/j.jml.2004.03.008

Giovannetti, T., Bettcher, B. M., Brennan, L., Libron, D. J., Kessler, R. K., & Duey, K. (2008). Coffee with jelly or unbuttered toast: Commissions and omissions are dissociable aspects of everyday action impairment in Alzheimer's disease. *Neuropsychology*, *22*(2), 235–245. https://doi.org/10.1037/0894-4105.22.2.235

Heathcote, A., Brown, S., & Cousineau, D. (2004). QMPE: Estimating Lognormal, Wald, and Weibull RT distributions with a parameter-dependent lower bound. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 277–290. https://doi.org/10.3758/BF03195574

Hitch, G. J., Towse, J. N., & Hutton, U. (2001). What limits children's working memory span? Theoretical accounts and applications for scholastic development. *Journal of Experimental Psychology: General*, *130*(2), 184–198. https://doi.org/10.1037//0096-3445.130.2.184

Kail, R. V. (2007). Longitudinal evidence that increases in processing speed and working memory enhance children's reasoning. *Psychological Science*, *18*(4), 312.

Kane, M. J., Bleckley, M., Conway, A., & Engle, R. (2001). A controlled-attention view of working memory capacity. *Journal of Experimental Psychology: General*, *130*(2), 169–183.

Kane, M. J., Conway, A., Hambrick, D., & Engle, R. (2007). Variation in working memory capacity as variation in executive attention and control. In C. Jarrold & A. Conway (Eds.), *Variation in Working Memory* (pp. 21–48). Oxford University Press.

Kane, M. J., & Engle, R. (2003). Working-memory capacity and the control of attention: The contributions of goal maintenance, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*, 47–70.

Kyllonen, P., & Christal, R. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*, 383–433.

Lépine, R. lle, Parrouillet, P., & Camos, V. (2005). What makes working memory spans so predictive of high-level cognition? *Psychonomic Bulletin & Review*, *12*(1), 165–170. https://doi.org/10.3758/BF03196363

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods*, *49*(2), 513–537. https://doi.org/10.3758/s13428-016-0740-2

McVay, J., & Kane, M. J. (2009). Conducting the train of thought: Working memory capacity, goal neglect and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *35*(1), 196–294.

Morrison, A. B., & Richmond, L. L. (2020). Offloading items from memory: Individual differences in cognitive offloading in a short-term memory task. *Cognitive Research: Principles and Implications*, *5*(1). https://doi.org/10.1186/s41235-019-0201-4

Navarro, D. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.5)*. University of Adelaide. http://ua.edu.au/ccs/teaching/lsr

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, *47*(4), 1343–1355. https://doi.org/10.3758/s13428-014-0543-2

R Core Team. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*(3), 164–171.

Redick, T. S., Calvo, A., Gay, C. E., & Engle, R. W. (2011). Working memory capacity and go/no-go task performance: Selective effects of updating, maintenance, and inhibition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*(2), 308–324.

Revelle, W. (2018). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University. https://CRAN.R-project.org/package=psych

Richmond, L. L., Redick, T. S., & Braver, T. S. (2015). Remembering to prepare: The benefits (and costs) of high working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1764–1777. https://doi.org/10.1037/xlm0000122

Ripley, B. D. (2004). Selecting amongst large classes of models. In *Methods and Models in Statistics: In Honour of Professor John Nelder, FRS* (pp. 155–170). World Scientific.

Salthouse, T. A., Pink, J. E., & Tucker-Drob, E. M. (2008). Contextual analysis of fluid intelligence. *Intelligence*, *36*(5), 464–486. https://doi.org/10.1016/j.intell.2007.10.003

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 26.

Scullin, M. K., Ball, B. H., & Bugg, J. M. (2020). Structural correlates of commission errors in prospective memory. *Cortex*, *124*, 44–53. https://doi.org/10.1016/j.cortex.2019.10.013

Scullin, M. K., Bugg, J. M., & McDaniel, M. A. (2012). Whoops, I did it again: Commission errors in prospective memory. *Psychology and Aging*, *27*(1), 46–53. https://doi.org/10.1037/a0026112

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, *125*(1), 4–27.

Shipstead, Z., & Broadway, J. M. (2013). Individual differences in working memory capacity and the Stroop effect: Do high spans block the words? *Learning and Individual Differences*, *26*, 191–195. https://doi.org/10.1016/j.lindif.2012.04.003

Spieler, D. H., Balota, D. A., & Faust, M. E. (1996). Stroop performance in healthy younger and older adults and in individuals with dementia of the Alzheimer's type. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(2), 461–479. https://doi.org/10.1037/0096-1523.22.2.461

St Clair-Thompson, H. L. (2007a). The influence of strategies on relationships between working memory and cognitive skills. *Memory*, *15*(4), 353–365. https://doi.org/10.1080/09658210701261845

St Clair-Thompson, H. L. (2007b). The effects of cognitive demand upon relationships between working memory and cognitive skills. *Quarterly Journal of Experimental Psychology*, *60*(10), 1378–1388. https://doi.org/10.1080/17470210601025505

Towse, J. N., Hitch, G. J., Horton, N., & Harvey, K. (2010). Synergies between processing and memory in children's reading span: Synergies between processing and memory in reading span. *Developmental Science*, *13*(5), 779–789. https://doi.org/10.1111/j.1467-7687.2009.00929.x

Towse, J. N., Hitch, G. J., & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of Memory and Language*, *39*(2), 195–217. https://doi.org/10.1006/jmla.1998.2574

Towse, J. N., Hitch, G. J., & Hutton, U. (2000). On the interpretation of working memory span in adults. *Memory & Cognition*, *28*(3), 341–348. https://doi.org/10.3758/BF03198549

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

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2009). There's more to the working memory capacity—Fluid intelligence relationship than just secondary memory. *Psychonomic Bulletin & Review*, *16*(5), 931–937. https://doi.org/10.3758/PBR.16.5.931

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Working memory capacity and retrieval from long-term memory: The role of controlled search. *Memory & Cognition*, *41*(2), 242–254. https://doi.org/10.3758/s13421-012-0261-x

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*, 1–26. https://doi.org/10.1016/j.cogpsych.2014.01.003

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

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory*, *17*(6), 635–654. https://doi.org/10.1080/09658210902998047

Wahlheim, C. N., Ball, B. H., & Richmond, L. L. (2017). Adult age differences in production and monitoring in dual-list free recall. *Psychology and Aging*, *32*(4), 338–353. https://doi.org/10.1037/pag0000165

Wahlheim, C. N., & Huff, M. J. (2015). Age differences in the focus of retrieval: Evidence from dual-list free recall. *Psychology and Aging*, *30*(4), 768–780. https://doi.org/10.1037/pag0000049

Wahlheim, C. N., Richmond, L. L., Huff, M. J., & Dobbins, I. G. (2016). Characterizing adult age differences in the initiation and organization of retrieval: A further investigation of retrieval dynamics in dual-list free recall. *Psychology and Aging*, *31*(7), 786–797. https://doi.org/10.1037/pag0000128

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology Section A*, *49*(1), 51–79. https://doi.org/10.1080/713755607

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wingfield, A., & Kahana, M. J. (2002). The dynamics of memory retrieval in older adulthood. *Canadian Journal of Experimental Psychology*, *56*(3), 187–199.

Table 1

Sample Sizes and Descriptive Statistics for WMC estimates from each Task and Site

	Overall		ASU		CSUS/SBU		TU		Site difference
	M	SD	M	SD	M	SD	M	SD	
OSpan	56.80	13.72	58.19	12.77	50.07	16.73	54.45	14.72	**
	n = 1685		n = 1180		n = 121		n = 384		
SymSpan	28.61	7.77	29.63	7.79	29.40	7.14	26.56	7.75	**
	n = 1053		n = 585		n = 121		n = 347		
RSpan	--		52.52	14.81	--		--		N/A
			n = 1062						

Note. ** indicates significant site differences observed at the $p \leq .001$ level.

Table 2

Descriptive statistics for storage performance, processing accuracy, and processing RT measures

Span Task		M	SD	Skew	Kurtosis	Norm Violation?
OSpan						
	Proc Acc	0.91	0.07	-2.71	11.81	Y
	Mean Proc Prac RT	3243.03	1385.79	1.51	3.68	N
	SD Proc Prac RT	1768.66	1306.06	2.41	9.23	Y
	RT Dist, μ	2546.80	761.69	1.63	5.40	Y
	RT Dist, σ	489.49	360.66	2.41	10.03	Y
	RT Dist, τ	1253.81	741.13	1.97	7.08	Y
	Time-Out Err	1.43	1.89	3.37	21.70	Y
	Incorrect Err	5.17	4.41	3.28	17.43	Y
SymSpan						
	Proc Acc	0.91	0.10	-2.66	10.42	Y
	Mean Proc Prac RT	2141.52	1041.46	1.58	4.29	Y
	SD Proc Prac RT	1190.12	831.42	2.86	14.30	Y
	RT Dist, μ	1721.80	676.88	1.74	3.85	Y
	RT Dist, σ	309.22	303.46	1.84	3.84	Y
	RT Dist, τ	746.69	467.15	1.82	6.67	Y
	Time-Out Err	0.80	1.28	2.56	9.58	Y
	Incorrect Err	3.18	3.74	3.04	13.02	Y
RSpan						
	Proc Acc	0.90	0.11	-3.43	15.44	Y
	Mean Proc Prac RT	3947.43	1276.23	1.35	3.85	N
	SD Proc Prac RT	1456.58	749.58	2.45	10.37	Y
	RT Dist, μ	3436.06	1054.71	0.97	4.77	Y

RT Dist, σ	669.88	415.34	3.16	20.40	Y
RT Dist, τ	991.19	587.05	2.11	11.16	Y
Time-Out Err	1.61	2.28	8.03	125.49	Y
Incorrect Err	5.62	7.40	3.73	17.83	Y

Note. WMC: capacity estimate; Proc Acc: processing task accuracy; Mean Proc Prac RT: mean practice processing task RT; SD Proc Prac RT: standard deviation practice processing task RT; RT Dist, μ: mu component of the ex-Gaussian analysis for task-relevant RTs; RT Dist, σ: sigma component of the ex-Gaussian analysis for task-relevant RTs; RT Dist, τ: tau component of the ex-Gaussian analysis for task-relevant RTs; Time-Out Err: processing task errors due slow responding; Incorrect Err: processing task errors due to incorrect responding. "Y" under Norm Violation? column indicates that the assumption of normality was violated, defined as skew > |2| and/or kurtosis >|4|, whereas "N" indicates that skew and kurtosis values were found to be in the acceptable range. For rows marked "Y", non-parametric statistical tests can be found in the Supplemental Materials, and in cases where non-parametric and parametric findings differed these are footnoted throughout the manuscript.

Table 3

Cronbach's alpha scores for processing and storage components of each task

	OSpan	SymSpan	RSpan
Processing Accuracy	.77	.81	.91
Processing RT	.93	.89	.91
Storage Accuracy	.92	.83	.92

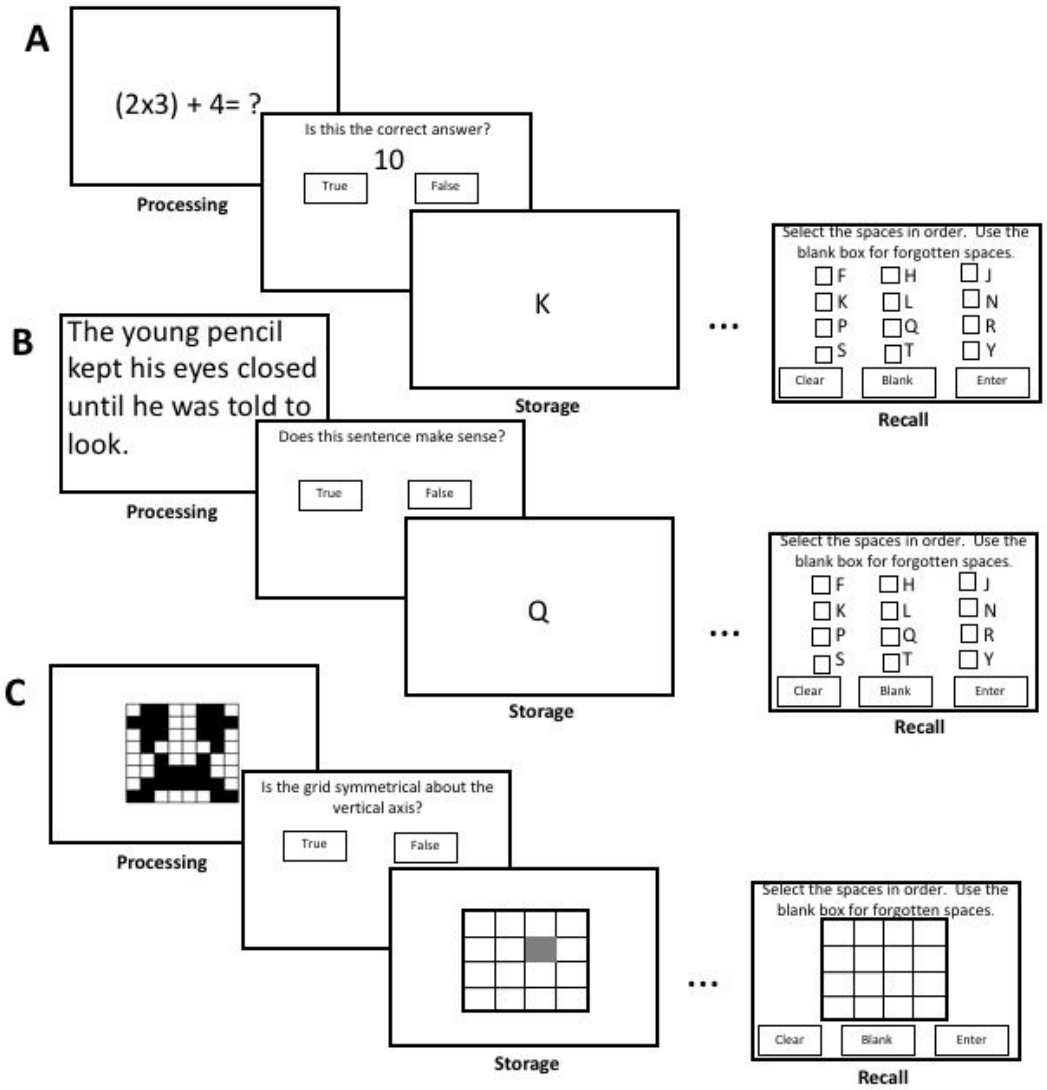


Figure 1. Complex Span Working Memory task schematics depicting Operation Span (panel A), Reading Span (panel B), and Symmetry Span (panel C).

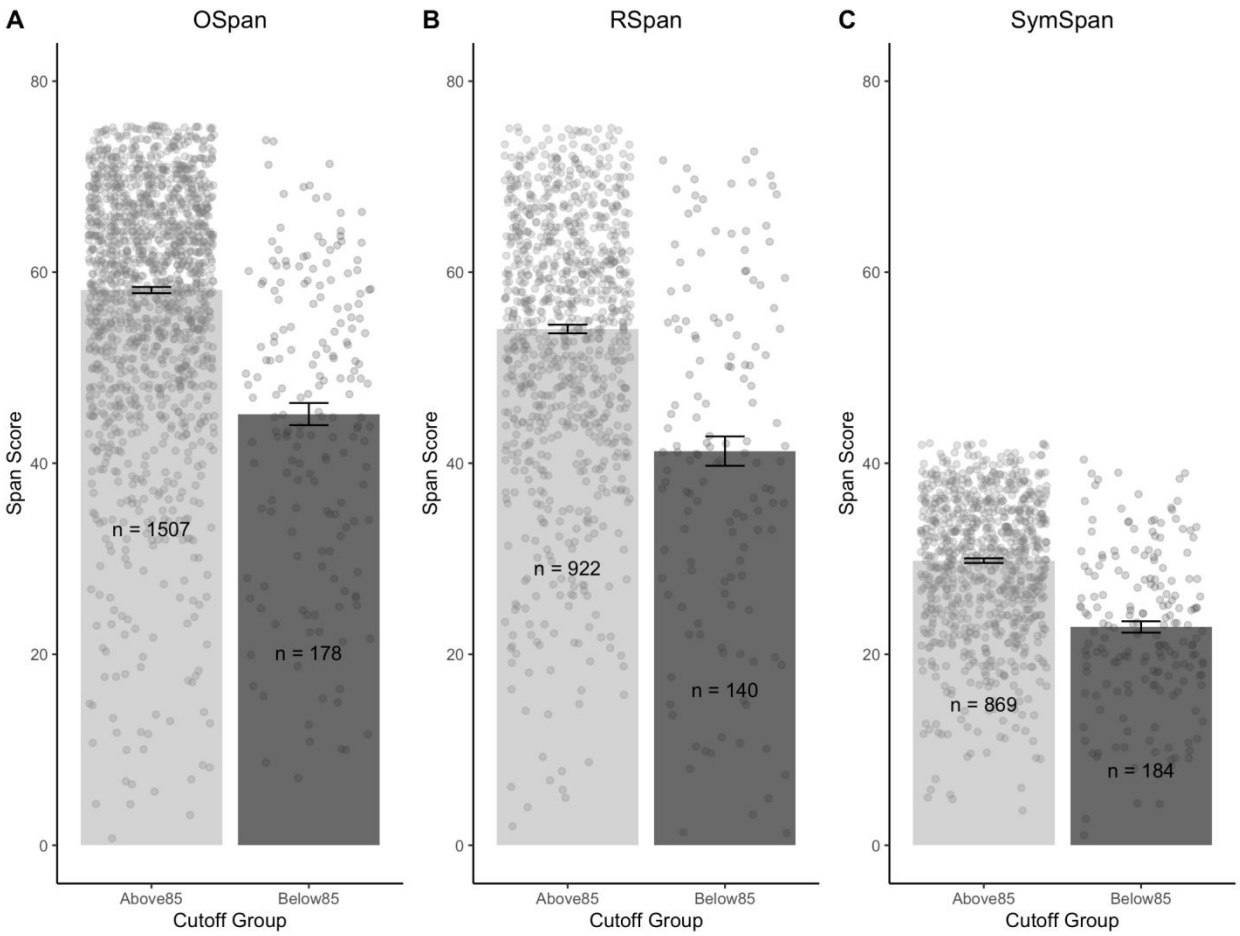


Figure 2. Participants who meet or exceed the 85% processing accuracy cutoff have significantly higher span scores across tasks compared to those who miss the cutoff. Individual data points are depicted by the circles and the error bar represents standard error.

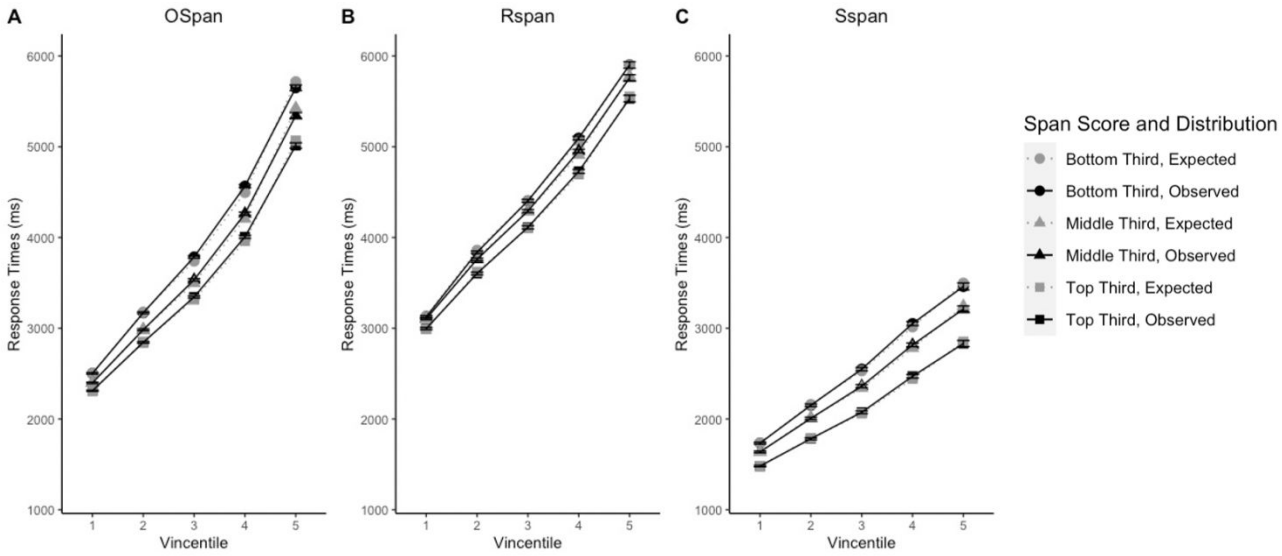


Figure 3. Vincentile plots by task (panel A: Operation Span, panel B: Reading Span, panel C: Symmetry Span) and by WMC partial span score (top third, middle third, and bottom third). This depiction is for illustrative purposes only; WMC was included in our models as a continuous variable. Expected values for each group are denoted by the grey dotted lines and the observed values are denoted by the black solid lines. The substantial overlap between predicted and observed values indicates that these data are fit well by the exGaussian function.

Supplemental Materials

Non-Parametric Analyses: OSpan

The following nonparametric analyses were conducted in the OSpan dataset due to the observation of violations of normality for the factors included in the tests:

For our research question regarding differences in WMC as a function of processing accuracy (question 1), we conducted a Wilcoxon Rank Sum Test. Consistent with the Welch's two sample t-test for unequal variances, results here indicate a significant difference in WMC estimates by processing performance (W = 203059, p < .001, 95% CI [11.00, 15.00], r = -0.35) for those who met/exceeded the 85% criterion cutoff, consistent with our parametric analysis.

For the research question regarding the extent to which the SD of practice RTs are related to WMC (question 2b), a Spearman's Rho correlation revealed a non-significant correlation (ρ (1683) = -0.03, p = 0.175) between WMC and SDs of RTs during the practice portion of the processing task. This finding is consistent with the result of the parametric Pearson's Product Moment correlation.

For our question regarding the extent to which ex-Gaussian characterization of task-derived RTs predict WMC (question 3b), we entered the components μ, σ, and τ as simultaneous predictors into a quantile regression where WMC estimates served as the outcome variable. Consistent with the results of the linear regression, the σ component emerged as the only significant predictor in the model (β= -0.21, t (1681) = -3.85, p < .001; τ and μ p values > .14).

We also conducted a quantile regression for our question regarding the relation between processing errors and WMC (question 3c). Consistent with the linear regression model, both time out errors (β= -0.17, t (1682) = -5.66, p < .001) and incorrect errors (β= -0.35, t (1682) = -9.24, p < .001) emerged as significant predictors in the model.

Turning to our last research question regarding improvements in model fit for RT distribution + error predictor models over error alone and RT distributional components alone (question 4), we tested the quantile regression models described above (RT distribution only, error only) against a model that added error type (time-out errors, incorrect errors) to the ex-Gaussian components to predict WMC estimates. Consistent with the comparison of linear regression models, the RT distribution + error model explained significantly more variance in WMC estimates compared to the RT-distribution only model (model comparison: F (2, 1679) = 71.90, p < .001) and the error only model (model comparison: F (3, 1679) = 9.18, p < .001).

Non-Parametric Analyses: SymSpan

The following nonparametric analyses were conducted in the SymSpan dataset due to the observation of violations of normality for the factors included in the tests:

For our research question regarding differences in WMC as a function of processing accuracy (question 1), we conducted a Wilcoxon Rank Sum Test. Consistent with the Welch's two sample t-test for unequal variances, results here indicate a significant difference in WMC estimates by processing performance (W = 118992, $p < .001$, 95% CI [6.00, 9.00], $r = -0.32$) for those who met/exceeded the 85% criterion cutoff, consistent with our parametric analysis.

For the research questions regarding the extent to which mean and SD of practice RTs are related to WMC (questions 2a and 2b), we conducted Spearman's Rho correlations. The result for question 2a (mean practice RT) revealed a significant correlation (ρ (1051) = -0.16, $p < .001$) between WMC and mean RT during the practice portion of the processing task. This finding is consistent with the result of the parametric Pearson's Product Moment correlation. Inconsistent with the result of the parametric test for question 2b (SD of practice RT), we observed a significant Spearman's Rho correlation between RT SD during the practice portion of the processing task and WMC (ρ (1051) = -0.10, $p < .001$).

For our question regarding the extent to which ex-Gaussian characterization of task-derived RTs predict WMC (question 3b), components μ, σ, and τ were entered as simultaneous predictors into a quantile regression where WMC estimates served as the outcome variable. Consistent with the results of the linear regression, the τ component emerged as a significant predictor in the model (β= -0.13, t (1049) = -2.85 p = .004; μ and σ p values > .04).

We also conducted a quantile regression for our question regarding the relation between processing errors and WMC (question 3c). Consistent with the linear regression model, both time out errors (β= -0.25, t (1050) = -4.57, $p < .001$) and incorrect errors (β= -0.38, t (1050) = -7.930, $p < .001$) emerged as significant predictors in the model.

Turning to our last research question regarding improvements in model fit for RT distribution + error predictor models over error alone and RT distributional components alone (question 4), we tested the quantile regression models described above (RT distribution only, error only) against a model contained both elements (RT distribution + error). Consistent with the comparison of linear regression models, the RT distribution + error model explained significantly more variance in WMC estimates compared to the RT-distribution only model (model comparison: F (2, 1047) = 48.56, $p < .001$) and the error only model (model comparison: F (3, 1047) = 27.84, $p < .001$).

Non-Parametric Analyses: RSpan

The following nonparametric analyses were conducted in the RSpan dataset due to the observation of violations of normality for the factors included in the tests:

For our research question regarding differences in WMC as a function of processing accuracy (question 1), we conducted a Wilcoxon Rank Sum Test. Consistent with the Welch's two sample t-test for unequal variances, results here indicate a significant difference in WMC estimates by processing performance (W =91722, p < .001, 95% CI [10.00 , 16.00], r = -0.25) for those who met/exceeded the 85% criterion cutoff, consistent with the findings from our parametric analysis.

For the research question regarding the extent to which the SD of practice RTs are related to WMC (question 2b), we conducted a Spearman's Rho correlation which revealed a non-significant correlation (ρ (1060) = -0.07, p = .021), consistent with the findings from our parametric analysis.

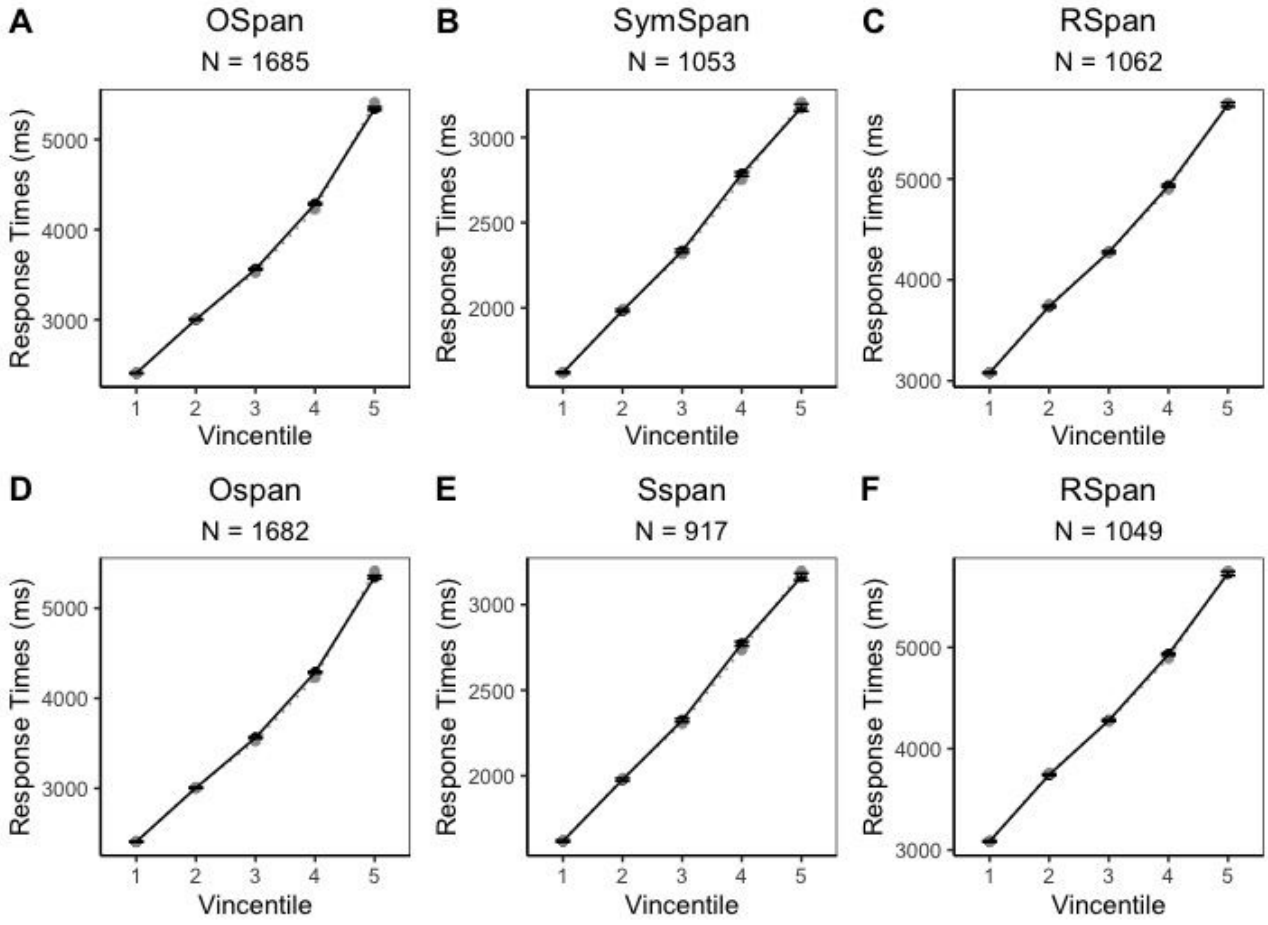
For our question regarding the extent to which ex-Gaussian characterization of task-derived RTs predict WMC (question 3b), components μ, σ, and τ were entered as simultaneous predictors into a quantile regression where WMC estimates served as the outcome variable. Consistent with the parametric analysis, the σ component emerged as the only significant predictor in the model (β= -0.007, *t* (1058) = -2.85, *p* = .004; μ and τ *p* values > .71).

We also conducted a quantile regression for our question regarding the relation between processing errors and WMC (question 3c). Consistent with the linear regression model, both time out errors (β= -0.21, *t* (1059) = -4.33, *p* < .001) and incorrect errors (β= -0.35, *t* (1059) = -6.39, *p* < .001) emerged as significant predictors in the model.

Turning to our last research question regarding improvements in model fit for RT distribution + error predictor models over error alone and RT distributional components alone (question 4), we tested the quantile regression models described above (RT distribution only, error only) against a model that included both RT distribution and error type as predictors. Consistent with the comparison of linear regression models, the RT distribution + error model explained significantly more variance in WMC estimates compared to the RT-distribution only model (model comparison: *F* (2, 1056) = 38.75, *p* < .001). Inconsistent with the comparison of linear regression models, the RT distribution + error model did not explain significantly more variance in WMC estimates compared to the error only model at the level of p < .01 (model comparison: *F* (3, 1056) = 2.88, *p* = .035).

Ex-Gaussian Model Fits

Below we present vincentile plots depicting model fits for the Ex-Gaussian characterization of RTs. Vincentile plots allow for examination of the raw RT distribution across conditions without making assumptions about the underlying shape of the distribution (Andrews & Heathcote, 2001). For each of the three span tasks, vincentiles were computed for by rank ordering raw RTs from shortest to longest for each participant, and calculating the mean of the first 20%, the second 20%, and so on. These plots present the best-fitting predicted vincentiles (dotted lines) superimposed on the observed vincentiles (solid lines). Across all tasks, the minimal divergence of the expected line from the observed line indicates that the data was well fit by the Ex-Gaussian function. Across all three complex span tasks, we present plots including all participants as well as plots excluding participants with fewer than 35 correct RTs.



Supplementary Figure 1. Panels A, B, and C depict observed and expected vincentiles for the complete samples. Panels D, E, and F depict observed and expected vincentiles when including only participants who contributed 35 or more correct RTs.

References

Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 514–544. https://doi.org/10.1037/0278-7393.27.2.514