Measuring verbal working memory capacity:

A reading span task for laboratory and web-based use

Jana Klaus* Herbert Schriefers

Radboud University Nijmegen

* Corresponding author:

Radboud University Nijmegen

Donders Centre for Brain, Cognition and Behavior

P.O. Box 9104, 6500 HE Nijmegen, The Netherlands

j.klaus@donders.ru.nl

We would like to thank Kristin Lemhöfer for helpful comments throughout the writing of this article, and Mirjam Feldt for support in data collection. This research was partly funded by a grant by the German Research Council awarded to JK (KL 2933/2).

Abstract

Typically, working memory (WM) capacity as a source of individual differences is assessed by complex span tasks which combine a processing and a storage task. However, there are no standardized open-source versions, and the tasks that are used are not easily comparable. We introduce a browser-based version of the reading span task, which yielded normally distributed recall performance scores. Next, we provide a within-participant comparison of this task to two other complex span tasks. Finally, we introduce a web-based version of the reading span task. WM scores were comparable to those obtained in the laboratory, but web participants were also faster and made more mistakes in the processing task. We conclude that the task introduced here is an adequate way to measure verbal WM capacity in the laboratory. In addition, it may prove to be a useful, time-efficient online tool, for instance to extract extreme groups from a larger sample.

Keywords: WORKING MEMORY CAPACITY; READING SPAN; BROWSER EXPERIMENTS

Introduction

The ability to store and manipulate information in parallel is a crucial attribute of the human cognitive system in that it allows for the relatively unhampered functioning of many complex cognitive activities. Coined “working memory” (Baddeley & Hitch, 1974), this capacity-limited system has been of interest to many experimental psychologists across many different subfields in the past decades. The strong relationship between working memory capacity (WMC) and intelligence is one of the most studied areas (e.g., Ackerman, Beier, & Boyle, 2005; Broadway & Engle, 2010; Hambrick, Kane, & Engle, 2005; Kane & Engle, 2002; Kyllonen & Christal, 1990), but WMC has also been shown to be related to language comprehension (Caplan & Waters, 1999; Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Just & Carpenter, 1992; Montgomery, 2003), dichotic listening (Conway, Cowan, & Bunting, 2001), suppressing intruding information in the anti-saccade task (Kane, Bleckley, Conway, and Engle, 2001; Roberts, Hager, & Heron, 1994), multi-tasking performance (Hambrick, Oswald, Darowski, Rench, & Brou, 2010), emotion regulation (Kleider, Parrott, & King, 2009), and a number of neurological diseases (e.g., aphasia [Caspari, Parkinson, LaPointe, & Katz, 1998], Alzheimer’s disease [Kempler, Almor, Tyler, Andersen, & MacDonald, 1998; Rosen, Bergeson, Putnam, Harwell, & Sunderland, 2002], Parkinson’s disease [Gabrieli, Singh, Stebbins, & Goetz, 1996], schizophrenia [Stone, Gabrieli, Stebbins, & Sullivan, 1998; Johnson et al., 2013; cf. Kane et al., in press]), among others.

A central assumption of these lines of research is that individuals differ in their WM capacity (WMC), and that these differences predict distinct behaviour in other cognitive domains (for overviews, see e.g., Jarrold & Towse, 2006; Kane, Conway, Hambrick, & Engle, 2007; Unsworth, in press). Not surprisingly, explaining human cognition by individual differences in WMC has gained substantial prominence in the last decades. A PubMed query requiring both “working memory” and “individual differences” to appear in the title and/or abstract of a publication provided a mere five results for the 1980s (the decade in which Alan Baddeley published his seminal monograph on the subject), which is in stark contrast to 145

results in 2016 alone. Researchers, thus, readily acknowledge the necessity to investigate cognitive processes not only by averaging performance of a test population, but also by looking at more fine-grained accounts that may – at least partly – elucidate human cognition.

Typically, individual differences in WMC are assessed by measuring participants' performance on one or more complex span tasks. These tasks combine a processing component (i.e., judging the correctness of sentences or equations) and a storage component (i.e., memorizing a series of words or digits for subsequent recall). Performance in the recall task is then statistically related to the individual performance in other cognitive tasks. In order to measure verbal WMC as one component of WM, likely the most prominent and most widely used versions are the reading span task (Daneman & Carpenter, 1980) and the operation span task (Turner & Engle, 1989). These tasks are increasingly used as automated computer versions (Redick et al., 2012) in which the timing parameters of the processing component are either identical for all participants or are adjusted to an individual's mean reaction time in a single-task situation. Such a standardized procedure provides a more controlled testing environment, hence minimizing overly excessive recall strategies as well as undesirable variance induced by the testing behavior of the experimenter.

However, with respect to the generalization of the influence of WMC on a range of cognitive tasks, at least two problems remain: First, there is no standardized, publicly available version of the different tasks, which poses a number of caveats with regard to the interpretation of results. Second, with respect to verbal WM, different span tasks are used interchangeably, although it is not clear whether these tasks that encompass highly different cognitive processes (e.g., sentence comprehension vs. mathematical operations) ultimately measure the same underlying construct[1]. In the current study, we therefore introduce a

[1] Of course the same holds for other domains of WMC, most notably visuospatial working memory. However, the focus of the current study are verbal complex span tasks, and future studies will need to provide a comparable investigation of other kinds of these tasks.

standardized version of the reading span task and compare performance in this task to performance in—more or less—similar complex span tasks.

Why do we need standardized versions of complex span tasks?

Researchers worldwide are investigating individual differences in cognition in order to collectively gain more insight into the workings of the human mind on an individual level. Therefore it should go without saying that the tasks with which they measure individual WMC should be comparable. In reality, however, this is not really the case. Besides the unfortunate side effect that researchers then spend a lot of time generating their own test versions, this lack of homogeneity in the use of complex span tasks bears an important risk: When administering one's own version of a given task, there are so many degrees of freedom (e.g., with respect to timing parameters, presentation mode, and the nature of the stimuli) that every study looking at individual differences ultimately ends up assessing them with a slightly different task. This in turn impedes comparability across studies. When the results of WMC testing are used as a basis for dividing a sample into high- and low-WMC groups, it remains unclear whether what qualifies as, for example, a high-WMC group in one study would qualify equally as high-WMC group in a different study using a different version of a span task. Similarly, when the results of a span task are used in regression analyses, factor analyses, structural equation models, etc., we are again confronted with the problem whether the span scores mean the same across different studies using different versions of a span task.

In an attempt to overcome this problem, Randall Engle's lab has provided free versions of an impressive number of complex span tasks on their website (http://englelab.gatech.edu/tasks.html). A major drawback of this effort, however, is that all of the tasks were programmed in E-Prime, which is licensed experimental software and thus not freely accessible to everyone. This does not only limit the availability of these tasks to labs with an E-Prime subscription, but also hampers pre-screening procedures which often are necessary to recruit extreme groups from a given population. Conway et al. (2005) point out the usefulness of testing groups from the extreme ends of the (WMC) continuum in order

to investigate whether WMC affects performance in other tasks or not, as opposed to treating WMC as a continuous variable. However, unless a research group can draw from a large pool of pretested participants, implementing such a pre-screening procedure in the laboratory requires a lot of time and resources. Having the opportunity to assess the WMC of a larger sample prior to the actual study by means of a web-based test which can be done by the participants outside of the laboratory offers a fast and economic alternative to derive extreme groups from a participant population.

In a recent study, Hicks, Foster, and Engle (2016) introduced a variety of complex span tasks which were administered entirely online. However, they could only replicate a relationship between verbal WM and fluid intelligence when the to-be-remembered stimuli were more abstract entities (i.e., easy-to-verbalize pictures or Klingon characters), not when they were letters (as is often the case in verbal span tasks administered in the lab). The authors reasoned that online participants are more inclined to cheat on the recall task, so the memoranda would have to be more difficult to write down in order to measure verbal WMC accurately.

Introducing an Automated Reading Span Task for Laboratory Settings

In Study 1, we introduce an automated reading span task. Next to WM performance, which is traditionally measured in terms of performance on the recall task, we will also report analyses on the participants' performance in the processing task as reflected by reaction times and error rates. Previous research suggests that these results should be included in the analyses for a more informative view of the performance in the complex span task performance (e.g., Daneman & Tardif, 1987; Waters & Caplan, 1996; Unsworth et al., 2009). Unsworth et al. (2009) reported moderate correlations between processing speed and accuracy in three different complex span tasks (reading span, operation span, symmetry span), but crucially, these measures accounted for independent parts of the variance in fluid intelligence. The authors thus concluded that reaction times and processing accuracy "do not reflect the same underlying construct (processing efficiency), but rather index two slightly different constructs" (p. 649).

Method

Participants. 72 participants (57 female, mean age: 23.4 years, *SD* = 5.3) from Leipzig University and Radboud University Nijmegen took part in the experiment for course credit or monetary reimbursement. No participant reported to suffer from dyslexia, and all participants were native speakers of German.

Design, materials, and procedure. The reading span task contained a processing component (judging the semantic correctness of a sentence) and a storage component (memorizing a noun for later recall). This sentence-noun compound will be referred to as a trial throughout this study. Two to six such trial combinations (i.e., set size 2 to 6) made up individual blocks, and the order of these blocks was randomized within participants to avoid predictability of the storage demands. For each set size, there were three blocks, amounting to a total of 15 blocks.

An experimental trial was structured as follows: First, the sentence was presented at the center of the screen for a maximum of 10 seconds (time-out) or until the participant provided a key response. Participants were asked to read the sentence out loud and press the right arrow button of the keyboard if it made sense, or the left arrow button if it did not make sense. After a blank screen of 500ms, the to-be-remembered word appeared for 1200ms, and participants were instructed to read this word aloud as well. Following two to six such sentence-word combinations, three question marks appeared at the center of the screen, prompting the participant to recall and overtly produce all items they could remember, regardless of serial order. Recall performance was immediately coded by the experimenter. The following trial was then initiated by the participant by pressing the enter key.

All stimuli were presented in Open Sans (20 px, black) at the center of the screen. Underneath the sentence, two icons were presented, reminding the participant that the left arrow key on the keyboard corresponded to the response “yes”, and the right arrow key corresponded to the response “no”.

For the processing task (i.e., the semantic judgment), 60 sentences with a mean length of 12 words (*SD* = 1.54), either taken from previous studies (Engle et al., 1999; Kane & Engle,

2004) or created by the authors, were used. In half of these sentences, one word was replaced with another word from the same grammatical category to create a semantically incorrect sentence (e.g., *"Die meisten Menschen sind sich einig, dass Montag der schlimmste Stab der Woche ist"* [most people would agree that Monday is the worst stick of the week]). Care was taken that these substitutions were made in different positions of the sentences, such that participants could not predict where to look for anomalies first. As word memoranda, 60 one- and disyllabic nouns were used. Half of them were imaginable (e.g., *"Gabel"* [fork]) and the other half un-imaginable (e.g., *"Zukunft"* [future]). The words were assigned such that within a given block (i.e., two to six sentence-word combinations), the words were not related phonologically, semantically, or associatively, and unrelated to the respective sentences.

The task was presented in Google Chrome using jsPsych (de Leeuw, 2015), and reaction times and key responses were recorded using XAMPP (http://www.apachefriends.org/). JsPsych is an open-source Javascript library for implementing behavioral experiments in a browser, without requiring additional experimental software. The documentation of jsPsych (available at http://docs.jspsych.org/) provides information on how to use the library, including assistance as to how to store the data. We provide the files of the reading span task used in the present study (in German, Dutch, and English) as well as R scripts to analyze the obtained output, at https://github.com/janakl4us/workingmemory/.

Scoring. The results of the recall task were scored using partial-credit unit scoring (see Conway et al., 2005; Friedman & Miyake, 2005). In this scoring method, correctly recalled items are first counted as a proportion of the respective block, and no specific weight is given to harder items (e.g., both one out of two items and three out of six items correspond to a score of .50). The scores for each block are then averaged to make up the final WM scores, which can thus range from 0 to 1.

Analyses. We used Bayesian inference as implemented in JASP (v0.8.0.0) to test our hypotheses. Specifically, we will report inverse Bayes factors (BF_{10}) which express the odds that an alternative hypothesis H_1 is to be preferred over the null hypothesis H_0. BF_{10} signifies

by what factor our prior belief (i.e., that H_0 is true) should be changed based on the data. For example, a BF_{10} value of 2 indicates that the data are twice as likely under H_1 compared to H_0. Speaking in discrete categories (Jeffreys, 1961), BF_{10} values between 1 and 3 are considered “anecdotal” evidence for H_1, values between 3 and 10 “moderate”, and values exceeding 10 “(very) strong”.

Results

Working memory performance. The first row of Table 1 displays descriptive statistics for the reading span task. Skewness values < 2 and kurtosis values < 4 indicate normal distribution (Kline, 1998). Cronbach’s α was calculated as a measure of internal consistency at the level of the 15 individual blocks, with the scores being computed as the correct proportion of the respective set (e.g., two out of three items recalled correctly corresponds to a proportion of .66; see Kane et al., 2004). Internal consistency for the reading span task was exceptionally high, suggesting that all blocks contributed to the same amounts to the individual scores.

TABLE 1

Processing task performance. Reaction times in the processing part of the complex span tests that deviated from a participant’s mean by more than 3 *SD*s were considered outliers and removed from further analyses (113 or 0.7% of all cases). The upper rows of Table 2 display descriptive statistics for reaction times and error rates in the processing task of the reading span task. As can be seen, performance was almost at ceiling. Further, reading span scores and error rates were negatively correlated, although this was not substantiated in the Bayesian analysis (r = -.25, BF_{10} = 1.28). However, given the polarity of the correlation (i.e., fewer errors in the processing task were associated with higher WM scores), we can assume that participants did not sacrifice the accuracy of the processing task performance in order to focus more on the recall task.

TABLE 2

Discussion

The automated reading span task for German native speakers using the open source software jsPsych resulted in normally distributed WM scores and good processing task performance. We therefore recommend it as an easy-to-implement alternative to previously introduced versions. However, based on these results we cannot give a recommendation as to whether the reading span is to be preferred over other measures of verbal WMC. As mentioned in the Introduction, verbal WMC has been assessed in a rather heterogeneous way in the literature. Particularly, variants of the operation span task—where the processing task is not a semantic judgment, but the evaluation of a mathematical equation—are often used. However, it is not clear to what extent this difference in processing demands affects performance on the recall task, and whether tasks using different processing components are actually comparable in the first place.

To explicitly investigate this issue, we had the same participants perform two other automated complex span tasks, i.e., an operation span task (judging the correctness of a mathematical equation while remembering digits) and a mixed span task (again, judging the semantic correctness of a sentence, but this time while remembering letters). For the comparison of these three tasks within participants, we kept all timing parameters identical between tasks and participants, while the stimuli and the task-specific instructions were varied (for details see below). The timing parameters concerned three different kinds of values: (1) the maximum time a to-be-processed stimulus (i.e., a sentence or a mathematical equation) was presented, (2) the time a to-be-remembered stimulus (i.e., a word, a digit, or a letter) was presented, and (3) the inter-stimulus intervals between the processing and the storage component. Previous research has shown that keeping these three factors constant across participants results in scores that are more predictive of abilities in episodic memory, executive functioning, and fluid intelligence than when they are varied (McCabe, 2010). Furthermore, with regard to the purpose of the current study, constant timing parameters

increase the likelihood that potential differences in WM performance can be attributed to the parameters we did vary: (1) the nature of the processing task, (2) the nature of the to-be-remembered stimuli, and (3) the task instruction.

For the processing task of the operation span, we used 60 mathematical equations (e.g., (10 ÷ 2) - 3 = 2) including all four basic arithmetic operations. Half of these equations had been changed to result in a wrong outcome (e.g., (10 ÷ 10) - 1 = 2). Participants were asked to judge the correctness of the equation by pressing the same buttons as for the semantic judgment in the reading span task. As memoranda, digits from 1 to 9 instead of words were used. They were assigned to a specific block randomly, with the only constraint that no digit appeared more than once in a given block. Unlike in the reading span task, participants were not instructed to read out the equation or the digit. The mixed span task was completely identical to the reading span task, except that we used 18 consonants instead of nouns as memoranda. These were assigned to the blocks so that the items did not repeat within a given block, did not share the same onset, and did not rhyme (e.g., we avoided combinations such as F and S or B and T). All experimental trials for both tasks were structured and timed exactly as in the reading span task, and performance was scored in the same way.

WM and processing task performance are displayed in the lower parts of Tables 1 and 2, respectively. In terms of WM performance, the scores in the operation span task were non-normally distributed. Therefore, all scores for the complex span tasks were transformed by raising them to the power of four for further analyses to approach a normal distribution. Multivariate outliers were assessed using Mahanalobis distance (Conway et al., 2005). One participant was located as being outside of the χ^2 distribution ($p < .001$). However, the overall results did not differ regardless of whether this case was excluded from the analyses or not, so we kept the data in the analysis. Both the operation span task and the mixed span task elicited higher WM scores than the reading span task (reading span task vs. operation span task: $BF_{10} = 4.74 \times 10^9$; reading span task vs. mixed span task: $BF_{10} = 4.68 \times 10^6$). Notably, the operation span task was performed almost at ceiling: 16 out of the 72 participants recalled all items correctly, and 75% of the participants scored higher than .90, which is

higher than the mean value of the two other tasks. To further test the similarity of the three tasks, we computed Pearson correlation coefficients for all contrasts. Unlike previous research, not all of the tasks correlated with each other, despite the fact that all tasks were tested within participants. The highest correlation was found between the reading span task and the mixed span task (r = .57 , BF_{10} = 89,532.7), which both had the same processing component (i.e., verifying the correctness of a sentence). Further, the operation span task and the mixed span task, which both had a similar recall component (i.e., memorizing single-unit entities [digits and letters, respectively] which were drawn from a finite response set), correlated highly with each other (r = .42 , BF_{10} = 125.6). Notably, there was no evidence for a correlation between the reading span task and the operation span task (r = .24 , BF_{10} = 1.0), which can be considered the two most distinct tasks with regard to processing component (semantic vs. mathematical judgment), recall component (words vs. digits), and instructions (loud vs. silent reading of the processing stimuli).

In terms of processing task performance, the operation span task was performed faster, but at the cost of accuracy compared to the reading span task (for reaction times, BF_{10} = 3.9 × 10^{29}; for error rates, BF_{10} = 868,322.0). Not surprisingly, there was no evidence for differences in reaction times and error rates between the reading span and the mixed span task, because they contained the same processing component (for reaction times, BF_{10} = 0.3; for error rates, BF_{10} = 0.2).

In order to investigate the relationship between verbal WMC and processing efficiency further, we calculated two composite scores (CS) for each task. These scores took into account not only the WM scores, but also the z-transformed mean reaction times (RT) and accuracy (acc) of the respective processing task:

$$\mathrm{CS}_{\mathrm{RT}} = \frac{z(\mathrm{RT})}{z(\mathrm{WM})} \text{ and } \mathrm{CS}_{\mathrm{acc}} = \frac{z(\mathrm{acc})}{z(\mathrm{WM})}$$

The correlations of the individual scores are displayed in Table 3. Not surprisingly, again the two complex span tasks sharing the same processing component (i.e., the reading span task and the mixed span task) correlated highly, both when incorporating reaction times and

accuracy of the processing task into the combined score. Importantly, and in contrast to the simple score correlations, the speed-adjusted score for the operation span task also correlated with the speed-adjusted scores of both the reading span task and the mixed span task. There was no similar relationship for the accuracy-adjusted scores.

TABLE 3

In sum, then, comparing the reading span task introduced above with more or less similar complex span tasks showed that—given the current task configurations—they are not easily comparable in terms of WM performance: The operation span task was performed almost at ceiling, while the reading span task and the mixed span task yielded substantially lower scores. Crucially, when comparing the reading span task and the operation span task, which are used interchangeably in the literature to assess verbal WMC, we found no evidence for a correlation between the WM scores. Only by computing composite scores that incorporated individual's reaction times of the processing task, thus accounting for the fact that processing task performance was significantly faster in the operation span task, the correlation became more pronounced.

This finding—which is, to our knowledge, new in the literature—can be caused by at least two factors or their interplay. The two major differences between the tasks as we used them in the current study are (a) the respective processing demands (i.e., judging the correctness of a sentence vs. an equation) and (b) the instructions (i.e., participants were asked to read the sentences, but not the equations out loud). In a similar experiment comparing different complex span tasks, Unsworth et al. (2009) reported mean reaction times of 3451ms for an operation span task and 4036ms for a reading span task, and a high correlation between the two tasks (r = .77). Crucially, in this study, both tasks did not require reading aloud, suggesting that solving mathematical operations is indeed faster than reading and judging a sentence, even if the latter task is also performed quietly. By inference, the faster reaction times for the operation span task in the current study cannot solely have been caused by the

differing instructions, but also at least partly by the different processing demands, which in turn facilitates storing the memoranda. A follow-up study should investigate if the WM scores yielded by both tasks could be approximated if any part of the operation span task was made more difficult (e.g., by increasing the difficulty of the equations, the set size, or the memoranda).

Importantly, all three tasks were performed in highly controlled laboratory settings in which the experimenter could provide feedback on the correct execution of the tasks if necessary. As mentioned in the Introduction, many future studies might benefit from a simpler screening process, for instance if the respective complex span task could be performed outside the laboratory. To test this, we developed a web-based version of the reading span task in Study 2.

Introducing a Web-Based Reading Span Task

In Study 2, we tested German native speakers using a web-based variant of the reading span task. If the scores obtained with this method turned out to be comparable to those obtained in a controlled laboratory setting (in terms of means and distribution of the recall task), this would support the conclusion that this task was suited for online testing, despite less strict monitoring of the participants' behavior throughout the testing procedure.

Method

Participants. 128 participants were recruited via the research participation system of Radboud University and received 5 € or course credit for their participation. 127 were aged between 18 and 30, and 101 were female. Eight participants were removed from the analysis because they made more than 15% errors in the processing task or did not give any keyboard responses at all.

Materials and procedure. The reading span task was identical to the one used in Experiment 1, with the following differences: Participants first saw a welcome screen which provided a brief summary of the task and informed them that they were free to cancel their participation at any time. Then they were asked to provide demographic information (i.e., age, gender, and whether the test language was their first language). On the next screen,

they could provide their email address if they wished to receive an evaluation of their performance. Finally, the test started. Participants received the same written instruction as in Experiment 1, with the only exception that the to-be-remembered words had to be written in six empty text fields at the end of a block. Overall, the experiment lasted for a maximum of about 15 minutes. Only complete datasets (i.e., those of participants who took the whole test) were stored on a secured server of the Radboud University and subsequently analyzed. German, Dutch, and English versions of the web-based test as well as corresponding analysis R scripts can be found at https://github.com/janakl4us/workingmemory/.

Results

The data were analyzed as for the laboratory version of the task reported above. 58 reaction times in the processing task were classified as outliers and removed from further analyses. The descriptive results of the WM score and processing task performance are displayed in Table 4. Web-based WM scores displayed a similar pattern as those obtained in the laboratory with respect to mean, range, and distribution. However, online participants were faster, while also making more mistakes. To statistically compare the laboratory and the web-based version of the reading span task, we performed Bayesian independent *t* tests with the grouping variable test situation (laboratory vs. web-based) for the WM scores as well as reaction times and error rates in the processing task. There was no evidence for a difference between the two groups regarding the WM scores ($BF_{10} = 0.30$), but strong evidence for a difference in reaction times ($BF_{10} = 4.64 \times 10^{8}$) and error rates ($BF_{10} = 1.21 \times 10^{7}$)[2].

[2] We also performed the online reading span task with 154 Dutch participants. Materials were an exact translation of the German version, and participants were also recruited via the Radboud University participation system. An independent Bayesian *t* test with the grouping variable test language (Dutch vs. German) provided no evidence for a difference in the WM scores between the two samples ($BF_{10} = 0.30$). An additional independent Bayesian t test with the grouping variable test situation (German laboratory vs. Dutch online) provided only anecdotal evidence for a difference in the WM scores between the two samples ($BF_{10} = 2.75$).

TABLE 4

Discussion

Experiment 2 introduced a web-based version of the reading span task that removed experimenter control over the compliance of the participants and potentially increased participants' susceptibility to distraction. Testing a large sample of university students, we found strong evidence that online participants performed the processing task faster, but made more mistakes, while WM scores (i.e., the variable of interest) were comparable across test situations.

The fact that online participants are faster while at the same time making more mistakes can almost certainly be attributed to the testing situation: If performed outside of the laboratory, participants do not receive direct feedback on how they perform or if they correctly stick to the instructions, and technically, they have no incentive to do so. For instance, it is conceivable that many participants in fact did not read the sentences out loud. Recall that the mean reaction times in the reading span task and mixed span task in Experiment 1 were well above 5000ms, while in the online dataset, about half of the participants responded at 4000ms and lower—a reaction time which is virtually impossible if they had really read the sentences out loud and responded only afterwards. Similarly, participants might be less focused or more easily distracted by external influences, which may explain the higher error rates in the processing task.

Interestingly, however, this different processing task performance does not seem to affect performance on the storage/recall component. Presumably, as long as participants allocate some of their attention to the processing task—and given that presentation times for the to-be-remembered items were identical for the online and the laboratory version—it does not matter whether the test is performed under supervision or not. In fact, if sticking to the instructions—as reflected by processing speed and/or accuracy on the processing task—indeed had a systematic influence on WM performance, WM scores on the web-based test

should be higher than in the laboratory settings because participants were able to neglect performance in the processing task while shifting more attention to the storage component. However, because this was not the case (i.e., participants did not obtain higher WM scores in the web-based test), we conclude that the trade-off in processing task performance has no profound impact on WM performance.

General Discussion

The current study introduced a browser-based, easy-to-implement version of the reading span task, compared performance on this task to two other complex span tasks, and implemented the reading span task as a web-based screening tool of verbal WMC. For the reading span task tested in the laboratory, we found normally distributed WM scores while WM performance on an operation span task was at ceiling. Based on these findings, we advise researchers to preferentially use the reading span task as a measure of verbal WMC. Of course, it is feasible that an operation span task with different parameters (i.e., larger set sizes or more difficult equations and/or memoranda) can be successful in approaching the difficulty levels of the reading span task as it is introduced here, but ideally, this should be demonstrated in a within-participant study first.

Measuring the reading span with a web-based test yielded comparable WM scores compared to those obtained in the laboratory. However, online participants were also faster and made more mistakes in the processing task, which can be attributed to the lack of experimenter control. Yet, given the comparability of the WM scores, which usually are the measure of interest when investigating individual differences, we believe it may be a useful, time-efficient online pre-screening tool, for instance to extract extreme groups from a larger sample. This ease of obtaining data from web-based experiments is also reflected in the larger sample size reported in Study 2. In order to exclude the possibility that the reported results (i.e., no difference between the laboratory and web-based WM scores) are contaminated by the different *N*s, we conducted a post-hoc analysis in which we drew 72 random participants from the web-based sample and compared their scores to those obtained in the laboratory version. After iterating this analysis 20 times, there was no

indication that the WM scores differed between the two samples, even though the sample sizes were now identical (range BF_{10}: 0.18 – 1.98; see Figure 1). This lends further support to the initially reported result that both measurements yield comparable WM scores.

In sum, we provide evidence that the reading span task introduced in the current study can be used to measure verbal WM capacity both in and outside the laboratory. Nonetheless, researchers should be aware that online studies bear additional risks which do not apply for traditional research within a controlled setting. For instance, there is no reasonable way to examine whether participants cheated during the test (see also Hicks et al., 2016). It is therefore advisable to validate results of, for example, a general screening to form extreme groups of low versus high WMC gained online with an additional confirmation of the resulting groups in the laboratory with the mixed span task introduced in the Discussion of Study 1. If WM performance is highly correlated for each participant, it is safe to assume that the online measurement was successful.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131,* 30-60. doi:10.1037/0033-2909.131.1.30

Broadway, J.M., & Engle, R.W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods, 42,* 563-570. doi:10.3758/BRM.42.2.563

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*, 77-126.

Caspari, I., Parkinson, S. R., LaPointe, L. L., & Katz, R. C. (1998). Working memory and aphasia. *Brain and Cognition, 37,* 205-223. doi:10.1006/brcg.1997.0970

Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review*, *8*, 331-335. doi:10.3758/BF03196169

Conway, A.R.A., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O., & Engle, R.W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12,* 769-786. doi:10.3758/BF03196772

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450-466. doi:10.1016/S0022-5371(80)90312-6

Daneman, M., & Merikle, P. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422-433. doi:10.3758/BF03214546

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods***,** *47***,** 1-12. doi:10.3758/s13428-014-0458-y

Engle, R.W., Tuholski, S.W., Laughlin, J.E., & Conway, A.R.A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128,* 309-331. doi:10.1037/0096-3445.128.3.309

Friedman, N.P., & Miyake, A. **(**2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods, 37,* 581-590. doi:10.3758/BF03192728

Gabrieli, J. D. E., Singh, J., Stebbins, G. T., & Goetz, C. G. (1996). Reduced working memory span in Parkinson's disease: evidence of the role of a frontostriatal system in working and strategic memory. *Neuropsychology, 10,* 322-332. doi:10.1037/0894-4105.10.3.321

Hambrick, D.Z., Kane, M.J., & Engle, R.W. (2005). *The role of working memory in higher-level cognition: Domain-specific versus domain-general perspectives.* In R. Sternberg & J.E. Pretz (Eds.), Cognition and Intelligence: Identifying the Mechanisms of the Mind (pp. 104-121). New York: Cambridge University Press.

Hambrick, D.Z., Oswald, F.L., Darowski, E.S., Rench, T.A., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied Cognitive Psychology, 24,* 1149-1167. doi:10.1002/acp.1624

Hicks, K. L., Foster, J. L., & Engle, R. W. (2016). Measuring Working Memory Capacity on the Web With the Online Working Memory Lab (the OWL). *Journal of Applied Research in Memory and Cognition*. doi:10.1016/j.jarmac.2016.07.010

Jarrold, C., & Towse, J.N. (2006). Individual differences in working memory. *Neuroscience, 139,* 39-50. doi:10.1016/j.neuroscience.2005.07.002

JASP Team (2016). JASP (Version 0.8.0.0) [Computer software].

Jeffreys, H. (1961). Theory of probability. Oxford, England: Oxford University Press.

Johnson, M. K., McMahon, R. P., Robinson, B. M., Harvey, A. N., Hahn, B., Leonard, C. J., Luck, S. J., & Gold, J. M. (2013). The relationship between working memory capacity and broad measures of cognitive ability in healthy adults and people with schizophrenia. *Neuropsychology*, *27*, 220-229. doi:10.1037/a0032060

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: individual differences in working memory. *Psychological Review*, *99*, 122-149. doi:10.1037/0033-295X.99.1.122

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working memory capacity. *Journal of Experimental Psychology: General*, *130*, 169-183. doi:10.1037/0096-3445.130.2.169

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: an individual differences perspective. *Psychonomic Bulletin & Review, 9,* 637-671. doi:10.3758/BF03196323

Kane, M.J., Hambrick, D.Z., Tuholski, S.W., Wilhelm, O., Payne, T.W., & Engle, R.W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189-217. doi:10.1037/0096-3445.133.2.189

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

Kempler, D., Almor, A., Tyler, L. K., Andersen, E. S., & MacDonald, M. C. (1998). Sentence comprehension deficits in Alzheimer's disease: a comparison of off-line and on-line processing. *Brain and Language, 64,* 297-316. doi:10.1006/brln.1998.1980

Kleider, H.M., Parrott, D.J., & King, T.Z. (2009). Shooting behaviour: How working memory and negative emotionality influence police shoot decisions. *Applied Cognitive Psychology, 23,* 1-11. doi:10.1002/acp.1580

Kline, R.B. (1998). *Principles and practice of structural equation modelling.* New York: Guilford Press.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity? *Intelligence, 14,* 389-433. doi:10.1016/S0160-2896(05)80012-1

McCabe, D. P. (2010). The influence of complex working memory span task administration methods on prediction of higher level cognition and metacognitive control of response times. *Memory & Cognition, 38,* 868-882. doi:10.3758/MC.38.7.868

Montgomery, J. W. (2003). Working memory and comprehension in children with specific language impairment: What we know so far. *Journal of Communication Disorders*, *36*, 221-231. doi:10.1016/S0021-9924(03)00021-2

Roberts, R. J., Hager, L. D., & Heron, C. (1994). Prefrontal cognitive processes: Working memory and inhibition in the antisaccade task. *Journal of Experimental Psychology: General*, *123*, 374-393. doi:10.1037/0096-3445.123.4.374

Rosen, V. M., Bergeson, J. L., Putnam, K., Harwell, A., & Sunderland, T. (2002). Working memory and apolipoprotein E: what's the connection? *Neuropsychologia, 40,* 2226-2233. doi:10.1016/S0028-3932(02)00132-X

Stone, M., Gabrieli, J. D. E., Stebbins, G. T., & Sullivan, E. V. (1998). Working and strategic memory deficits in schizophrenia. *Neuropsychology, 12,* 278-288. doi:10.1037/0894-4105.12.2.278

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

Unsworth, N, & Engle, R.W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114,* 104-132. doi:10.1037/0033-295X.114.1.104

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,* 635-654. doi:10.1080/09658210902998047

Unsworth, N. (in press). The many facets of individual differences in working memory capacity. *The Psychology of Learning and Motivation.* doi:10.1016/bs.plm.2016.03.001

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100,* 426-432. doi:10.1037/a0022790

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, 49A,* 51-79. doi:10.1080/713755607

Table 1

Descriptive statistics for WM performance in the three complex span tasks.

Task	*M*	*SD*	range	skew	kurtosis	α
Reading span	.81	.11	.54-1.00	-.19	-.65	.97
Operation span	.93	.08	.56-1.00	-1.88	5.68	.70
Mixed span	.88	.08	.67-1.00	-.68	-.17	.95

Table 2

Descriptive statistics of reaction times and error rates for the processing task of each complex span task.

Task	*M*	*SD*	range	skew	kurtosis
	Reaction times in ms				
Reading span	5162	649	3966-6764	.64	.06
Operation span	3218	691	1928-4743	.11	-.80
Mixed span	5107	708	3871-7582	.74	.80
	Error rates in %				
Reading span	1.97	2.29	0.00-10.00	1.43	1.82
Operation span	4.79	3.25	0.00-13.33	0.47	-0.48
Mixed span	2.29	2.65	0.00-11.67	1.70	2.81

Table 3

Pearson correlations of the composite scores accounting for performance in the respective processing task. Inverse Bayes factors (BF_{10}) are displayed in brackets.

	CS_{RT_OST}	CS_{RT_MST}	CS_{acc_RST}	CS_{acc_OST}	CS_{acc_MST}
CS_{RT_RST}	0.363 (18.2)	0.713 (4.5×10^9)	0.510 (4197.3)	0.120 (0.2)	0.220 (0.8)
CS_{RT_OST}	—	0.362 (17.8)	-0.030 (0.2)	0.349 (12.4)	0.102 (0.2)
CS_{RT_MST}		—	0.354 (14.0)	0.169 (0.4)	0.432 (166.0)
CS_{acc_RST}			—	0.123 (0.2)	0.551 (33787.3)
CS_{acc_OST}				—	0.3 (2.4)

Note. CS = composite score. RST = reading span task. OST = operation span task. MST = mixed span task. The subscript *RT* refers to the composite score including reaction time. The subscript *acc* refers to the composite score including accuracy.

Table 4

Descriptive statistics of partial credit score, reaction times and error rates for the online version of the reading span task.

	M	*SD*	skew	kurtosis	range	α
Partial credit score	.83	.12	-0.97	0.54	.45 – 1.00	.97
Reaction times in ms	4304	886	0.21	-0.38	2584 – 6674	–
Error rates in %	5.11	3.70	0.78	-0.04	0.00 – 15.00	–

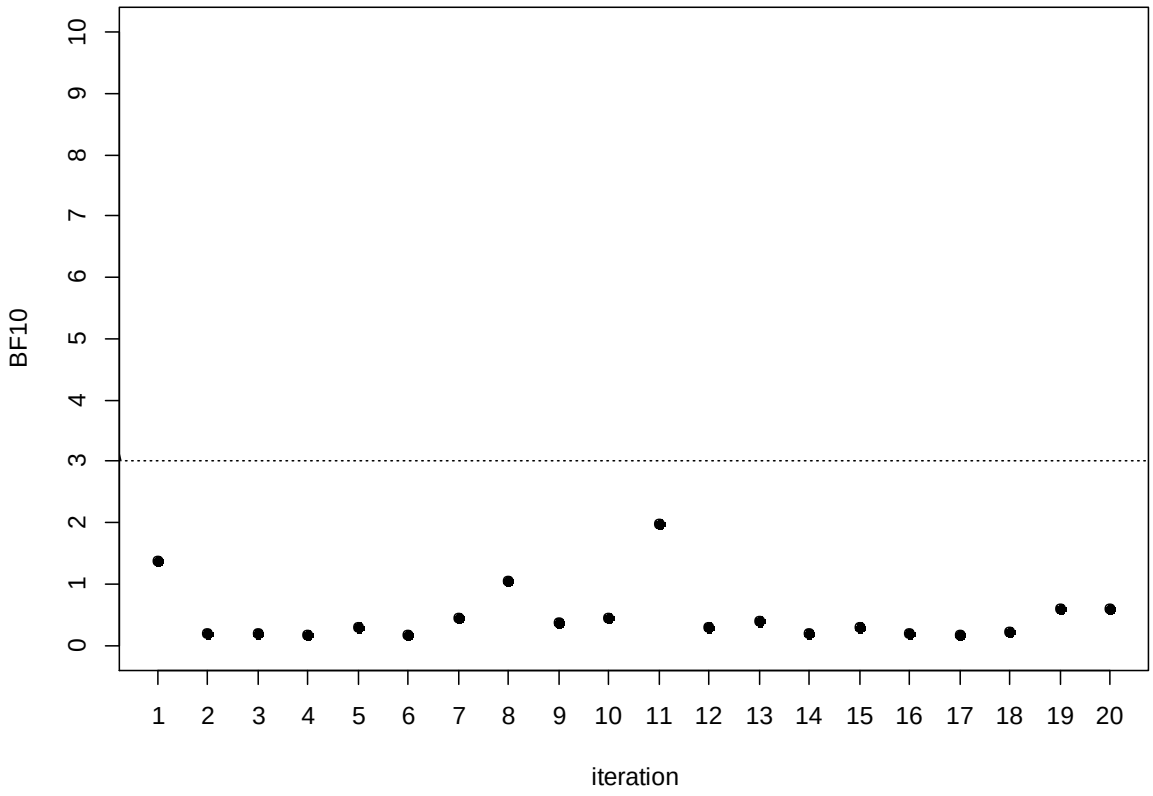

Figure 1. BF_{10} values for comparing WM scores obtained with the laboratory vs. the web-based test. Sample size for the web-based test was adjusted to match that of the laboratory test (N = 72) by randomly drawing an equal amount of participants out of the entire group (N = 120). Twenty independent Bayesian t tests with the grouping variable test situation (laboratory vs. web-based) were run. None of these suggested a difference between the WM scores of the two groups.