Applied Psycholinguistics **39** (2018), 275–301
doi:10.1017/S0142716417000182

Selective updating of sentences: Introducing a new measure of verbal working memory

DANIEL FELLMAN, ANNA SOVERI, CHARLOTTE VIKTORSSON, SARAH HAGA, JOHANNES NYLUND, SANNA JOHANSSON, JAKOB EDMAN, FELIX VON RENTELN, and MATTI LAINE
Abo Akademi University

Received: October 11, 2016 Accepted for publication: April 25, 2017

ADDRESS FOR CORRESPONDENCE
Daniel Fellman, Department of Psychology, Abo Akademi University, Fabriksgatan 2, FIN-20500 Turku, Finland. E-mail: dfellman@abo.fi

ABSTRACT
Working memory (WM) is one of the most studied cognitive constructs in psychology, because of its relevance to human performance, including language processing. When measuring verbal WM for sentences, the reading span task is the most widely used WM measure for this purpose. However, comparable sentence-level updating tasks are missing. Hence, we sought to develop a WM updating task, which we termed the selective updating of sentences (SUS) task, which taps the ability to constantly update sentences. In two experiments with Finnish-speaking young adults, we examined the internal consistency and concurrent validity of the SUS task. It exhibited adequate internal consistency and correlated positively with well-established working memory measures. Moreover, the SUS task also showed positive correlations with verbal episodic memory tasks employing sentences and paragraphs. These results indicate that the SUS task is a promising new task for psycholinguistic studies addressing verbal WM updating.

Keywords: reliability; sentence processing; updating; validity; verbal working memory

Working memory (WM) represents a temporary mental platform for maintaining, accessing, manipulating, and coordinating information (Baddeley, 2000; Oberauer, Süß, Wilhelm, & Wittman, 2003). WM is central to volitional mental activities, such as reasoning (Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002), executive control in everyday life (Kane, Brown, et al., 2007), and multitasking (Hambrick, Oswald, Darowski, Rench, & Brou, 2010).

According to the seminal model by Baddeley and Hitch (1974), WM consists of a limited-capacity attentional system (the central executive) that interacts with two autonomous slave systems (phonological loop and visuospatial sketchpad). The phonological loop with its subvocal rehearsal and phonological storage components is most directly related to language processing and language learning, but the other WM systems are relevant to language and language disorders as well,

including the more recently added episodic buffer component (Baddeley, 2003). Verbal WM is a strong predictor of reading comprehension (Daneman & Merikle, 1996), and has shown to be associated with several language-related integration processes, such as resolving lexical ambiguity (Mason & Just, 2007; Miyake, Just, & Carpenter, 1994), disregarding irrelevant details when reading text (Sanchez & Wiley, 2006), and drawing contradicting conclusions from a text (Oberauer, Weidenfeld, & Hörnig, 2006).

Because of the central role of WM in higher order cognitive abilities including language processing, several task paradigms have been developed to measure individual WM capacity. The most frequently used paradigm is the complex span task (CST), and some researchers have even considered the CSTs as the gold standard among WM capacity measures (Conway et al., 2005). What all variants of the CST have in common is that one must keep in mind a series of items while performing an intervening task in between the stimuli to be remembered (Redick et al., 2012). Thus, the task imposes both storage and processing load on WM: the former load varies according to target item number, while the latter one induced by the intervening secondary task remains constant. The first CST was introduced by Daneman and Carpenter (1980), who deployed a *reading span task* in which participants were to simultaneously judge the semantic correctness of sentences and remember the sentence-final words in correct order. Other commonly used CSTs include the *operation span task* (Turner & Engle, 1989), where participants are asked to read and verify a simple math problem and then remember a digit after the operation, and the *symmetry span task* (Kane et al., 2004) that requires symmetry judgments and remembering spatial locations. CSTs have been shown to be associated with various aspects of cognitive performance, such as language comprehension and fluid intelligence (Daneman & Merikle, 1996; Unsworth & Engle, 2007). Moreover, the CSTs typically demonstrate excellent psychometric properties by showing high test–retest reliability, high internal consistency, convergent and discriminant construct validity, and criterion-related validity (Conway et al., 2005; Redick et al., 2012).

Another aspect of WM that has been extensively investigated is the "updating" component. Together with task switching and inhibition of prepotent responses, WM updating has been identified as one of three primary executive processes (Miyake et al., 2000). It can be defined as the ability to store and modify information (Morris & Jones, 1990), achieved by removing outdated items (Ecker, Lewandowsky, & Oberauer, 2014; Ecker, Oberauer, & Lewandowsky, 2014), by adding new items, and/or by modifying information in WM (Kessler & Meiran, 2006, 2008; Kessler & Oberauer, 2015; Oberauer, 2002). Often used WM updating tasks include the N-back task, the Keep Track task, and running memory tasks. What these tasks have in common is that participants are prompted to replace earlier information in WM with newer information as the task progresses (Lechuga, Pelegrina, Pelaez, Martin-Puga, & Justicia, 2016). As with the CSTs, WM updating has shown to be closely related to higher level cognitive abilities such as fluid intelligence and reading comprehension (Carretti, Borella, Cornoldi, & De Beni, 2009; Friedman et al., 2006; Palladino, Cornoldi, De Beni, & Pazzaglia, 2001). Moreover, most of the recent findings indicate that performance in CSTs is a strong

predictor of performance in WM updating tasks (Ecker, Lewandowsky, Oberauer, & Chee, 2010; Schmiedek, Hildebrandt, Lövdén, Wilhelm, & Lindenberger, 2009; Schmiedek, Lövdén, & Lindenberger, 2014; Wilhelm, Hildebrandt, & Oberauer, 2013), albeit less so with the N-back task (see Redick & Lindsey, 2013, for a meta-analytic review). Thus, possibly with the exception of the N-back task that has been criticized for lacking in construct validity (Kane, Conway, Miura, & Colflesh, 2007), current evidence indicates that both the CSTs and WM updating tasks rely heavily on the same complex cognitive processes (Ecker et al., 2010; Schmiedek et al., 2014).

The issue of the ecological validity of the currently available WM tasks has been raised recently. This discussion has mainly taken place in the context of WM training studies, with the notion that training paradigms and pre/post tests often represent rather artificial laboratory tasks (Holmes, 2011; Holmes & Gathercole, 2014; Klingberg, 2010; Moreau & Conway, 2014; Shipstead, Redick, & Engle, 2010). Tasks such as the CST are repetitive and predictable, typically using unrelated digits, letters, or spatial locations as stimuli (but see the reading span task described above). It has been suggested that future tasks should target WM mechanisms from a more ecological perspective (for a review, see Moreau & Conway, 2014). We applied this idea to the field of language processing by devising a new verbal WM updating task that operates with sentences. Reading sentences is something literates do on a daily basis, and the ability to process textual information is related, for example, to academic accomplishment, financial success, and socioeconomic status (Ricketts, Sperring, & Nation, 2014; Ritchie & Bates, 2013). Verbal WM has been considered as crucial for the ability to process sentences, because this ability requires rapid computation of linguistic relations between temporally distal parts of the sentence (Lewis, Vasishth, & Van Dyke, 2006).

Our novel verbal WM updating task is coined the *selective updating of sentences* (SUS) task. The SUS task is a modified version of the paradigm presented by Murty et al. (2011). In their updating task, participants were asked to keep sequences of digits in mind by selectively replacing them with new digits. Their task has been employed in subsequent WM studies (Solopchuk, Alamia, Olivier, & Zenon, 2016; Waris, Soveri, & Laine, 2015; Yu, FitzGerald, & Friston, 2013) with slight modifications, such as using letters instead of digits as stimuli. However, in our task, participants were to update complete semantically feasible sentences by selectively replacing some constituent words. The SUS task calls for constant updating of the semantic contents of a verbal message in WM, thus bearing similarity to real-life communicative situations where corrections and additions commonly take place.

The present study consists of two experiments. Experiment 1 sought to determine the internal reliability and the concurrent validity of the SUS task. The validation was performed by investigating how the SUS task performance was related to the CSTs. In Experiment 2, we replicated the findings of Experiment 1, examined the test–retest reliability of the SUS task, and investigated how it was related to verbal episodic memory tasks measured with sentence and paragraph recall.

EXPERIMENT 1

This experiment examined the internal reliability of the SUS task. Moreover, we evaluated its concurrent validity by correlating the SUS task performances with CST performances.

Method

Participants. The participants consisted of 170 neurologically and psychiatrically healthy university and polytechnics students from the University of Turku, the Turku University of Applied Sciences, and the University of Helsinki. The mean age of the sample was 25.22 years ($SD = 5.18$), and 76.9% were female. Most of them were university students (97.6%) and the remaining ones (2.4%) polytechnics students. The participants were native speakers of Finnish with little or no exposure to any other languages during childhood. Each participant who successfully completed the test battery was rewarded with a movie ticket.

Procedure. The data were collected with an in-house programmable Internet-based test platform. It allows researchers to create, distribute, and manage psychological experiments. The experiments are distributed over the Internet by sending a link to the participants who can complete the experiment on their own computer. All tasks in an experiment are at first automatically installed on the participant's computer and then run on that computer, obviating issues with the varying speed of Internet connections. The system guides the user through the experiment and finally sends the performance data to a secure university server for analysis.

The participants were recruited via e-mail lists. The whole test session was performed online via either a desktop or a laptop computer that was connected to the Internet. The participants were instructed to carry out the test session in a quiet place where they would not be disturbed or interrupted. The participants completed the tasks one after another within a single session. The presentation order of the tasks was randomized for each participant. The participants could take a short break between the tasks, and they were advised to cancel the task completion if they experienced fatigue, headache, or other discomforts. Before the online test session could start, the participants had to give informed consent. The test battery included a background questionnaire followed by five computerized tasks and took about 70 min to complete.

SUS. Our novel Finnish-language version of the SUS task was designed based on the structure of the WM updating task by Murty et al. (2011). The task is illustrated in Figure 1. In this task, sentences were presented on the computer screen for 4000 ms, one word per box. The participant was to encode and keep in mind the initial sentence, after which it disappeared and a blank screen was shown for 500 ms, followed by an updating stage with a new row of boxes. At the updating stage, two of the boxes contained a word while the rest remained empty. The participant was prompted to replace the old words with the newly presented words in their WM, while at the same time maintaining the unchanged words in the original sentence. Finally, a row of empty boxes appeared on the screen, and the participants were

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

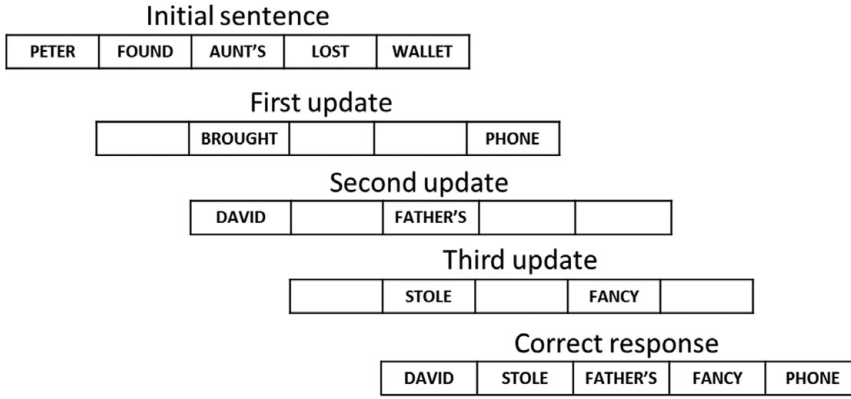

Figure 1. An English translation of a sample trial of the selective updating of sentences task with three updating stages in a sentence with five words. Note that there are no articles in the Finnish language.

to type up the latest version of the sentence, including the most recent updates. The updates of the sentence constituents were pseudorandomized so that their positions could not be predicted. The participants completed altogether 12 trials (a single trial consisted of an initial sentence followed by its updating stages), the order of which was randomized for each participant. The task consisted of three blocks with 4 trials in each block. One of the blocks included sentences with two updating stages, one block with three updating stages, and one block with five updating stages. In each block, the sentence length ranged from four to seven words so that 1 trial of each sentence length was presented in all blocks. The sentence lengths and the number of updating stages incorporated in the SUS task were based on an extensive piloting phase where we noticed that a span ranging from four to seven words and two to five updating stages appropriately revealed individual differences in task performance.

The Finnish sentences in the SUS task were ordinary declarative sentences of both transitive and intransitive type, following the canonical SVX word order. Complex syntactic structures were avoided. Information content in relation to sentence length was quite high due to the frequent use of attributes. The stimulus sentences, for instance, included predicative clauses (Koulun loputtua poika oli nälkäinen "After the school ended the boy was hungry"), transitive clauses (Pirteä mies myi koripallon "The cheery man sold a basketball"), intransitive clauses (Nuoret pojat ilahtuivat lahjasta "The young boys were happy for the gift"), ownership clauses (Minun miehelläni on uusi auto "My husband has a new car"), and existential clauses (Perheen talon katolla oli lintu "At the roof of the family's house there was a bird"). When the sentences were updated and replaced by other words, they still remained semantically and syntactically plausible (see Figure 1 for an example).

CSTs. Following the procedure of Unsworth, Heitz, Schrock, and Engle (2005), we deployed three computerized CSTs, namely, the reading span task, the opera-

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

tion span task, and the symmetry span task. As CSTs are rather time-consuming to administer, we shortened our CSTs so that there was one trial per each sequence length instead of three as was the case with the original versions (Unsworth, 2010; Unsworth et al., 2005). It is worth noting that shortened versions of CSTs have shown to be reliable measures of WM (Foster et al., 2015), and have demonstrated adequate psychometric properties (Oswald, McAbee, Redick, & Hambrick, 2015), even though the intercorrelations between the CSTs tend to decrease slightly (Foster et al., 2015; Gonthier, Thomassin, & Roulin, 2016; Oswald et al., 2015).

READING SPAN. Following Daneman and Carpenter (1980), we devised a Finnish version of the reading span task. In this task, the participants were asked to read a set of sentences occurring on the screen and make an acceptability judgment, thus deciding if the sentence was semantically correct (e.g., Havanna is the name of the capital city of Cuba) or not (e.g., The cat developed a spade after it had repaired the friend). Subsequently, the participant was to memorize the last word of each sentence, and finally recall the last words by their order of presentation.

Each sentence was shown on the screen for up to 8000 ms. The next sentence appeared as soon as the participant provided the response. The participants recalled the words to be remembered by typing them up into an empty box that was shown at the end of each trial. The task included seven trials (with two to eight items to recall) with one trial per each sequence length, and the order of the trials was randomized.

OPERATION SPAN. Following Turner and Engle (1989), an operation span task was implemented. Here the participant was to solve simple math equations (e.g., $5 - 2 + 6 = 9$?) while simultaneously trying to memorize a set of unrelated digits. The digit to be remembered was displayed on the computer screen for 1000 ms, followed by a fixation point (an asterisk) for 500 ms. After that, the math equation appeared for 6000 ms. At the end of each trial, a recall grid was shown, and the participant was to recall the unrelated digits in the order they were presented. The participant completed six trials. The sequence lengths ranged between 4 and 9 (one trial per sequence length), and the order of the trials was randomized.

SYMMETRY SPAN. As the third CST, we implemented a symmetry span task following Kane et al. (2004). Here the participant was to recall sequences of spatial locations while doing symmetry judgments. A gray 3×3 matrix, with one of the squares colored in white, was presented on the screen for 1000 ms. The participant was to memorize the location of the white square. Immediately afterwards, three white 3×3 matrices filled with black squares (the location of which varied in each matrix) were shown on the computer screen for 6000 ms. During this time, the participant was to make a symmetry judgment by visualizing the patterns of the black squares in the upper two matrices as a single pattern, and by determining whether this assembled pattern corresponded with the pattern of the third matrix located in the lower part of the screen. The participants performed altogether five trials (with three to seven items to be recalled) with one trial per sequence length. The order of the trials was randomized.

Scoring. In the SUS task, one point was given for each correctly recalled word that was typed up in the corresponding box. The percentage of correctly recalled words on all trials was used as the dependent variable in the analyses. For the CSTs, we used a partial-credit scoring system (Conway et al., 2005), where the number of correctly recalled elements per trial was counted, regardless of trial length (e.g., two correctly recalled elements from a three-element trial were worth as much as two correctly recalled elements from a five-element trial). The highest possible score was 35 for the reading span task, 39 for the operation span task, and 25 for the symmetry span task. Accuracy scores for the intervening task, calculated as the percentage of correctly solved processing problems, have previously been suggested to have a lower limit of 85% (Conway et al., 2005; Unsworth et al., 2005). However, previous tests in our laboratory indicate that this accuracy criterion may be too strict, leading to a sizable number of excluded participants. Therefore, we used a binomial probability (one tailed, $p < .05$) to define the cutoff value for above-random individual performance. In other words, we set the cutoff to the point where the number of correct responses on the intervening yes/no task in the CSTs reached the p value below .05. The binomial probability analysis revealed the following cutoff scores: reading span task $\geq$ 65.71% correctly solved problems, operation span task $\geq$ 66.67% correctly resolved problems, and symmetry span task $\geq$ 72.00% correctly solved problems.

Results of Experiment 1

Scores on any task that deviated more than 3.5 *SD* from the group mean were defined as univariate outliers. Based on this, one participant was excluded from the analyses on the reading span task. Regarding the binomial cutoff score for the intervening task in the CSTs, 19 participants were excluded in the operation span task, and 13 participants in the symmetry span task. All participants scored above the binomial cutoff score in the reading span task.

The data was also screened for multivariate outliers. In the multivariate outlier analysis, we replaced missing data (i.e., the univariate outlier in the reading span task and the participants in the CSTs that scored below our binomial cutoff in the intervening task) with variable means (Tabachnick & Fidell, 1996). We computed the Cook D values as >1 considered as outliers (Cook & Weisberg, 1982). No participant exceeded this critical value. We also performed a second multivariate outlier analysis by investigating the Mahalanobis distance value for each participant, using the χ^2 table ($p < .001$; Tabachnik & Fidell, 2007). The cutoff score for the Mahalanobis distance was set to 18.47 due to the number of predictor variables ($n = 4$). One participant showed a distance score exceeding this critical value and was excluded from the analyses on all tasks. The final sample included in the analyses thus consisted of 169 participants.

Table 1 presents item-level descriptives and item-total correlations (ITCs) for the SUS task. A positive ITC means that a higher score on a given item predicts a higher average score on the other items in the task, while a low ITC indicates that the item does not covary with the rest of the items in the task. Almost all item scores were normally distributed except for the item with four words and two updates, where the distribution was both negatively skewed and showed a high

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

Table 1. *Item descriptives and corrected ITC for the selective updating of sentences task in Experiment 1*

Item							Range	
Words	Updates	*M*	*SD*	Skew	Kurtosis	ITC	Potential	Actual
4	2	98.22	6.99	−4.165	18.397	.257	0–100	50–100
4	3	89.50	16.04	−1.543	2.310	.339	0–100	25–100
4	5	88.17	19.30	−1.543	1.525	.394	0–100	25–100
5	2	85.92	21.00	−1.430	1.182	.378	0–100	20–100
5	3	70.89	24.68	−0.651	−0.113	.371	0–100	0–100
5	5	81.89	24.71	−1.504	1.780	.323	0–100	0–100
6	2	85.01	21.26	−1.497	1.752	.452	0–100	0–100
6	3	85.50	18.95	−1.366	1.995	.495	0–100	0–100
6	5	74.36	20.00	−0.849	1.271	.412	0–100	0–100
7	2	75.74	19.64	−1.073	1.932	.479	0–100	0–100
7	3	69.57	22.29	−0.751	0.369	.418	0–100	0–100
7	5	44.13	18.89	−0.212	0.454	.220	0–100	0–100

Note: $N = 169$. Means are the percentage of correctly recalled words. ITC, item-total correlation.

kurtosis. The lowest ITCs were observed for the item with seven words and five updates ($r = .220$) and for the item with four words and two updates ($r = .257$). However, as a widely used rule of thumb states that correlations between .2 and .3 are acceptable for the ITC (Everitt, 2002), we chose to include all items in the SUS task.

We analyzed first the effects of our task manipulations (number of updating stages; sentence length) within the SUS task by a two-way repeated measures analysis of variance (ANOVA; see Figure 2). A main effect of updating stages indicated that performance decreased significantly as a function of the number of updates, $F\,(2, 167) = 112.44$, $p < .001$, partial $\eta^2 = 0.401$. There was also a main effect of sentence length, with the performance significantly decreasing with longer sentences, $F\,(3, 166) = 202.49$, $p < .001$, partial $\eta^2 = 0.547$. Moreover, the interaction term was statistically significant, $F\,(6, 163) = 38{,}09$, $p < .001$, partial $\eta^2 = 0.185$, primarily stemming from a more pronounced sentence length effect in the most demanding updating condition.

Descriptives for all tasks are summarized in Table 2. The number of correctly recalled items was calculated for each trial and Cronbach α was calculated across all trials. Internal consistency was adequate for the SUS task, the operation span task and the symmetry span task. The current version of the reading span task showed a slightly lower internal consistency. In general, the scores in each task were normally distributed. In the reading span task, the operation span task, and the symmetry span task, the intervening tasks showed high kurtosis, indicating ceiling effects. However, similar results have previously been observed in CSTs (Gonthier et al., 2016; Redick et al., 2012), as the intervening task is a "secondary" task to be solved with a rather minor effort.

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

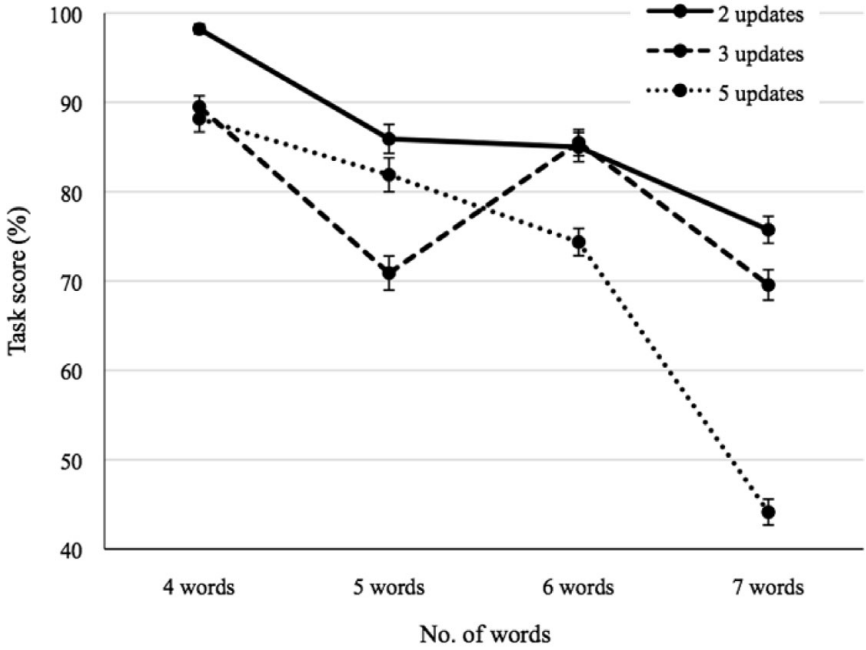

Figure 2. The mean task performance as a function of updating stages and sentence length on the selective updating of sentences task in Experiment 1. Error bars represent standard errors of mean.

Table 2. *Descriptives and Cronbach alpha values for all test variables in Experiment 1*

Measure	*M*	*SD*	Skew	Kurtosis	α
SUS	76.96	10.42	−0.452	−0.074	0.740
Reading span task scores					
Working memory	22.79	4.51	−0.069	0.174	0.629
Intervening task	93.13	5.07	−2.020	4.477	
Operation span task scores					
Working memory	22.69	8.84	−0.557	−0.231	0.782
Intervening task	86.26	8.93	−0.366	−0.771	
Symmetry span task scores					
Working memory	13.31	5.92	−0.079	−0.530	0.686
Intervening task	90.67	6.76	−0.568	−0.208	

Note: The range of possible task scores is 0–100 for the SUS task, 0–35 for the reading span task, 0–39 for the operation span task, and 0-–25 for the symmetry span task. SUS, selective updating of sentences task.

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

Table 3. *Intercorrelations for all variables in Experiment 1*

Measure	SUS	RSpan	OSpan
SUS	—		
RSpan	.349	—	
OSpan	.253	.372	—
SymSpan	.241	.376	.506

Note: All correlations were statistically significant ($p < .01$). $N = 144–169$. SUS, selective updating of sentences task; RSpan, reading span task; OSpan, operation span task.

The concurrent validity of the SUS task was assessed by the intercorrelations between the SUS task and the CSTs (see Table 3). The SUS task showed statistically significant positive correlations with all CSTs, albeit these correlations were somewhat lower than the correlations between the CSTs.

Discussion of Experiment 1

The aim of Experiment 1 was to study the internal consistency and the concurrent validity of our novel SUS task. The SUS task manipulations worked as expected, as increased WM load in the form of both the number of updating stages and the sentence length led to significant performance decrements. The range of difficulty was conveniently large, varying from nearly 100% on the easiest trial to approximately 40% correct recall on the most demanding one. The SUS task exhibited also adequate internal consistency. Concurrent validity was examined with intercorrelations between the SUS task and three well-established WM measures, namely, different CSTs. Although the task intercorrelations in general were not particularly high, the correlations with the SUS task were all positive and statistically significant, suggesting that the SUS paradigm taps WM.

EXPERIMENT 2

To replicate and expand upon the results from Experiment 1, we conducted a second experiment in which we reexamined the internal consistency and investigated the test–retest reliability of the SUS task. Moreover, as the SUS task was designed to be a more naturalistic sentence-level WM measure, we examined the correlations between the SUS task and verbal episodic memory tasks at sentence and paragraph level, assuming that these tasks reflect more real-life sentence processing. Finally, hierarchical regression analyses were performed to probe whether the SUS task possessed some predictive value in the performance in the verbal episodic memory tasks over and above the well-established CSTs.

In this experiment, we included two sentence recall tasks in the test battery. Sentence recall performance has been shown to predict language proficiency skills,

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

such as language-processing speed and processing of complex sentence structures (Poll, Miller, & van Hell, 2016; Riches, Loucas, Baird, Charman, & Simonoff, 2010). Moreover, it is a valuable tool for language assessment with children as it draws upon a wide range of language-processing skills such as language comprehension, vocabulary knowledge, and grammar abilities (Archibald & Joanisse, 2009; Klem et al., 2015; Ziethe, Eysholdt, & Doellinger, 2013).

Besides the sentence recall tasks, we also chose to include a paragraph recall task. Furthermore, a computerized word fluency task was administered in which the participants were to generate words beginning with a certain letter in a given time. As regards the WM tasks, we included the same versions of the SUS task, the reading span task, and the operation span task that were employed in Experiment 1. Moreover, we incorporated two additional WM tasks to the test battery. In the minus 2 span task, the participants were to recall a series of digits after subtracting 2 from each digit. In the alphabet working memory task, letters presented to the participants were to be memorized and transformed into other letters according to cues.

Method

Participants. A total of 101 undergraduate students, recruited from the University of Turku, the Turku University of Applied Sciences, and the University of Helsinki took part in this study. The participants were screened for psychiatric illnesses, neurological illnesses, and for possible use of CNS medication. Moreover, they underwent a language background screening to ensure that only participants with a monolingual language background were included. Altogether 21 participants were not eligible, and the final sample thus consisted of 80 participants. They had an average age of 24.4 years ($SD = 3.6$), and 74.7% were female ($n = 59$).

Procedure. These data stem from a pretest session of a WM training study that will be reported elsewhere. Each participant who successfully completed all phases of the training study (pretest, training, posttest) received a compensation of 70 euros. The data were collected with our in-house programmable Internet-based test platform. The pretest was performed as computer class sessions with 1–12 participants. The test battery included a background questionnaire followed by 10 computerized tasks, and it took about 2.5 hr to complete. As in Experiment 1, the presentation order of the tasks was randomized for each participant. The participants could take a short break between the tasks and were advised to terminate participation if they experienced fatigue, headache, or other discomforts.

WM tasks. The participants completed five WM tasks, namely, the SUS task, the reading span task, the operation span task, the minus 2 span task, and the alphabet working memory task. The SUS task, the reading span task, and the operation span task were identical to those described in Experiment 1.

MINUS 2 SPAN TASK. In the minus 2 span task (Waters & Caplan, 2003), sequences with digits occurred successively on the screen, and the participant was

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

to subtract 2 from each digit. For example, a correct response for a sequence of 4–8–3–5–6 would be 2–6–1–3–4. Each digit appeared on the screen for 1000 ms, followed by a fixation point for 500 ms. After each trial, a recall grid with horizontally aligned boxes ranging from number 1 to 9 prompted the participant to respond. The task included 12 trials with 2 trials of each sequence length. The sequence lengths ranged from four to nine digits, and the trials appeared in a randomized order.

ALPHABET WORKING MEMORY TASK. In this task (Was, Rawson, Bailey, & Dunlosky, 2011), the participant was presented with either one letter or two alphabetically nonadjacent letters for 2500 ms, followed by a transformation phase according to direction and number cues (–3, –2, –1, +1, +2, +3), which remained on the screen until the participant decided to proceed with the task. At the transformation phase, the task was to mentally move either up or down the alphabet according to the cues (e.g., JO + 3 = MR). This was followed by an empty column where the participant was to respond by typing the transformed letter/letters. The participants performed altogether 18 trials with 9 trials including a single letter to be remembered and 9 trials with two letters to recall. The forward and backward recoding directions (– or +) and recoding distances (1, 2, or 3) varied systematically in both trial lengths. The order of the trials was randomized for each participant.

Verbal episodic memory tasks. We also implemented four tasks that tapped verbal episodic memory. These tasks were sentence recall in Finnish, sentence recall in English,[1] paragraph memory, and word fluency.

SENTENCE RECALL. Here we administered both a Finnish and an English version of the sentence recall task to ensure that task intercorrelations were not language specific. The sentence recall versions were similar in terms of the number of words in the sentences and the scoring procedure. In this task, words of a sentence were presented successively on a computer screen at a rate of one word per 1000 ms. Immediately after a sentence sequence ended, the participant was asked to reproduce the sentence by typing it up in an empty column.[2] The sentences were 18–22 words long, and their contents tapped diverse topics in science, nature, and history. Both the Finnish and the English versions consisted of five sentences presented in a randomized order.

PARAGRAPH MEMORY. In this task, the participant was instructed to read a paragraph and memorize key points rather than trying to remember every word in the text. Nevertheless, when recalling the story, the participant was instructed to use the original words. The paragraphs were shown on a computer screen with no time limit for reading. When the participant decided to proceed, an empty box appeared on the screen, and the participant was to write down as much of the paragraph as possible. The length of the paragraphs ranged between 57 and 59 words. Two trials were completed in a randomized order. Besides the proportion of correctly recalled words, we also applied a semantical scoring procedure. See the Scoring section below for a more detailed description of the two scoring procedures for this task.

WORD FLUENCY. This was a computerized Finnish version of the word fluency task by Benton and Hamsher (1978). In this task, a letter was shown on a computer screen, and the participant was asked to type as many words as possible beginning with that letter in 60 s. The letter was visible on the upper section of the screen while an empty column was displayed on the lower section. The participant was to type the first relevant word that came to mind and then press the "Enter" button. After that the column went blank again, and the participant could immediately type the next word. This procedure was repeated until 60 s had passed. The participant was to perform three trials with different letters, and the presentation order of the trials was randomized.

As the present sample was derived from a pretest session in a training study that included both pre- and postmeasurements, there were two different versions of each verbal episodic memory task. Thus one half of the participants received a different version than the other half did. Yet, the task versions were designed to be equally demanding by matching them in terms of the complexity of words, the rate of high/low-frequency words, and the complexity of syntactic structure where applicable. Independent-samples *t* tests were conducted to examine whether the mean scores on the two task versions differed significantly. There were no statistically significant differences in the mean scores on sentence recall in Finnish, $t\ (78) = 1.959$, $p = .054$, sentence recall in English, $t\ (78) = 1.737$, $p = .086$, paragraph memory, $t\ (78) = 0.182$, $p = .856$, or paragraph semantic memory, $t\ (78) = 0.932$ $p = .354$. The only significant differences between the groups were observed on word fluency, $t\ (78) = 2.030$, $p = .047$. However, the difference in the word fluency task was not expected to exert any major influence on the present analyses that focus only on task intercorrelations. Thus, the task versions were lumped together in the following analyses.

Scoring. The scoring systems for the SUS task, the reading span task, and the operation span task were the same as in Experiment 1. In the sentence recall tasks, the proportion of correctly recalled words, regardless of the order in which they were recalled, was used as the outcome variable. Moreover, a word was considered correct if the participant was able to recall the word stem correctly. Suffix alterations were ignored, as well as pure orthographical errors.

In the minus 2 span task, one point was given for each correctly recalled digit in a correct position. The total score was the proportion of correct items in correct positions across all trials (Waters & Caplan, 2003). The highest possible score was 78. In the alphabet working memory task, the proportion of correctly recalled trials per minute was used as the dependent variable (Was et al., 2011). Thus, there was no maximum score in this task. In the paragraph recall, the first dependent variable was the proportion of correctly recalled words, corresponding to the measure used in the sentence recall tasks. In this variable, Test Version 1 had a maximum score of 116, and Test Version 2 a maximum score of 118. The second variable tapped paragraph semantic recall, measuring the proportion of correctly recalled semantic contents that had been determined beforehand when designing the paragraphs. In the paragraph semantic recall, the highest possible score was 27. Regarding the word fluency task, a total score was calculated as the sum of unique correctly reported words in the three trials.

Statistical analysis

Internal consistency and concurrent validity. As in Experiment 1, the internal consistency of the SUS task was investigated with item-level descriptives and ITCs. A two-way ANOVA was also conducted to investigate how the SUS task manipulations (i.e., number of updates, sentence length) altered task performance. Concurrent validity was examined with correlations between the SUS task, the CSTs, and the two other measures tapping verbal WM.

Test–retest reliability. A subset ($n = 37$) of the participants included in this study served as an active control group in a WM training study. These participants performed the SUS task on two separate occasions in a pre- and posttest setup, allowing for the examination of test–retest reliability. The active control group practiced with a computerized quiz task that called for long-term memory but did not load on WM. The delay between the two testing sessions was 5 weeks for all participants. We used the Pearson product moment correlation coefficient (Pearson r) to estimate the test–retest reliability of the SUS task.

Regression analysis. Hierarchical two-step multiple regression analyses were used to explore the relationship between the predictors (i.e., the CSTs and the SUS task), and the dependent variables of interest (i.e., sentence and paragraph recall tasks). The CSTs were entered as predictors at Stage 1 of the regression model, and the SUS task was introduced at Stage 2.

Results of Experiment 2

Total scores for any task that deviated more than 3.5 *SD* from the mean were defined as univariate outliers. There were, however, no outliers of this kind in the present experiment. As regards the intervening task in the CSTs, all participants scored above our binomial cutoff score in the reading span task. In the operation span task, nine participants scored below our cutoff score and were excluded from the analyses in this task. The presence of multivariate outliers was also investigated. Again, we replaced missing data (i.e., the participants in the operation span task that scored below our binomial cutoff score) with variable means (Tabachnick & Fidell, 1996). Multivariate outliers were searched for by using the Cook D (Cook & Weisberg, 1982) and the Mahalanobis distance value, χ^2 (10, 80) = 29.59, $p = .001$ (Tabachnick & Fidell, 2007). These did not lead to any exclusions: no participant showed a value of >1 on the Cook D measure, and no participant had a Mahalanobis distance score exceeding the critical value of 29.59. The final sample included in the analyses thus consisted of 80 participants.

Table 4 shows item-level descriptives and ITCs for the SUS task. In general, the SUS item scores were normally distributed, but in line with Experiment 1, the item with four words and two updates was negatively skewed and showed a questionable ITC score. However, as the goal in Experiment 2 was to replicate our previous results, we chose to retain all items. To examine whether SUS task performance remained stable over time, we performed a test–retest reliability analysis. The SUS task had a 5-week test–retest reliability of $r = .707$, 95% confidence interval [0.528,

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

Table 4. *Item descriptives and corrected item-total correlations for the selective updating of sentences task in Experiment 2*

Item							Range	
Words	Updates	*M*	*SD*	Skew	Kurtosis	ITC	Potential	Actual
4	2	97.19	7.95	−2.500	4.357	.173	0–100	75–100
4	3	84.38	23.66	−1.560	1.879	.406	0–100	0–100
4	5	84.06	19.18	−1.075	0.714	.297	0–100	25–100
5	2	82.00	20.77	−0.900	−0.112	.421	0–100	20–100
5	3	67.50	24.31	−0.546	−0.208	.454	0–100	0–100
5	5	78.75	22.35	−1.325	2.248	.412	0–100	0–100
6	2	83.75	18.75	−1.035	0.287	.543	0–100	33.3–100
6	3	87.08	16.56	−1.186	0.680	.610	0–100	33.3–100
6	5	73.75	19.63	−0.603	−0.138	.441	0–100	16.7–100
7	2	74.11	20.47	−1.107	1.765	.550	0–100	0–100
7	3	66.43	23.96	−0.548	−0.334	.380	0–100	0–100
7	5	47.14	21.13	−0.223	−0.523	.300	0–100	0–85.7

Note: $N = 80$. Means are the percentage of correctly recalled words. ITC, item-total correlation.

0.863], indicating an adequate level of test–retest reliability (Strauss, Sherman, & Spreen, 2006).

As in Experiment 1, we ran a two-way ANOVA to investigate how the participants reacted to the task manipulations (number of updates, sentence length) on the SUS task (see Figure 3). Both main effects were statistically significant, indicating decreased performance with more updates, $F\ (2, 78) = 45.13$, $p < .001$, partial $\eta^2 = 0.364$, and longer sentences, $F\ (3, 77) = 83.87$, $p < .001$, partial $\eta^2 = 0.515$. The interaction term was also significant, $F\ (6, 74) = 15.10$, $p < .001$, partial $\eta^2 = 0.160$. This reflected performance variability with different updating/length combinations, with the largest change concerning the longest sentences where the recall performance following the maximum number of updating stages dropped considerably.

Table 5 displays the descriptive statistics and Cronbach αs for each variable in Experiment 2. The skewness and kurtosis values indicate that the scores were normally distributed for each task. For the WM tasks, the coefficient αs were adequate for the SUS task, the operation span task, and the minus 2 span task. For the reading span task and the alphabet working memory task, the coefficient αs were low to marginal. Regarding the verbal episodic memory tasks, the coefficient αs were good for the paragraph recall task and the sentence recall in English, and adequate for the sentence recall in Finnish and the paragraph semantic recall task.

The intercorrelations for the tasks included in Experiment 2 are shown in Table 6. The SUS task showed statistically significant correlations with all WM tasks. In line with Experiment 1, the strongest correlation was observed between the SUS task and the reading span task. When investigating the correlations with the verbal episodic memory tasks, statistically significant positive correlations were

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

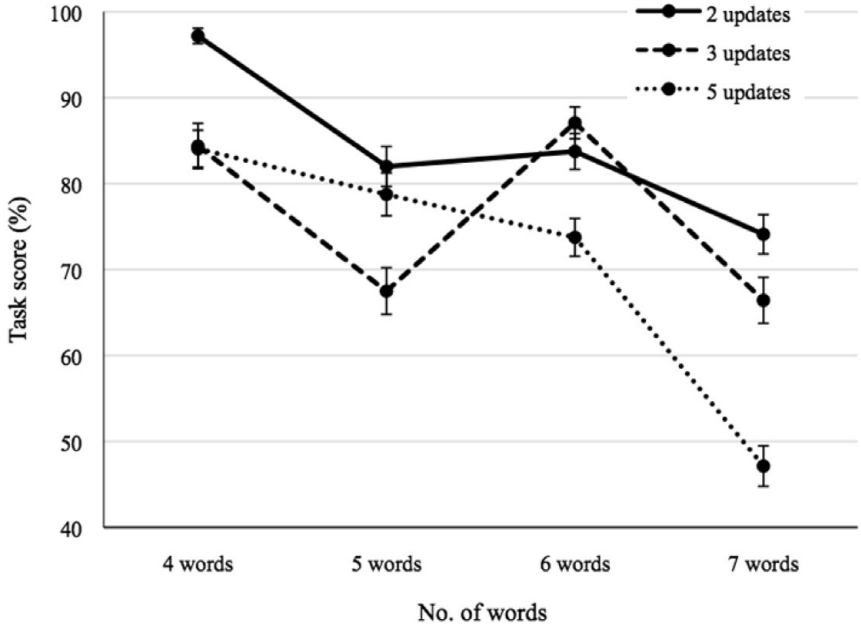

Figure 3. The mean task performance as a function of updating stages and sentence length on the selective updating of sentences task in Experiment 2. Error bars represent standard errors of mean.

observed between all verbal episodic memory tasks and the SUS task. Regarding the correlations between the other WM tasks and the sentence recall tasks, the results showed that the strongest correlations were with the reading span task, the operation span task, and the minus 2 span task. In the paragraph recall tasks, the strongest correlations were observed with the reading span task. However, the minus 2 span task also correlated significantly with the paragraph recall tasks, albeit those correlations were slightly lower than with the reading span task. In summary, the SUS task, the CSTs and the minus 2 span task were the WM tasks that were most strongly related with the verbal episodic memory tasks. Furthermore, the SUS task was the only WM task that showed a positive correlation of a moderate strength with the sentence recall in Finnish.

The results of the hierarchical regression analyses predicting sentence and paragraph recall performance are presented in Table 7. When predicting sentence recall in Finnish at Stage 1, the reading span task and the operation span task contributed significantly to the regression model, $F\,(2, 68) = 10.204$, $p < .001$, and accounted for 23.1% of the variance in this task. Introduction of the SUS task explained an additional 14.8% of the variance in the sentence recall in Finnish, and this change in R^2 was significant, $F\,(3, 67) = 15.905$, $p < .001$. In the sentence recall in English, Stage 1 revealed that the CSTs explained 23.3% of the variance in the task performance, $F\,(2, 68) = 10.350$, $p < .001$. Adding the SUS task to the regression

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

Table 5. *Descriptives and Cronbach alpha values for all variables in Experiment 2*

Measure	No. of Trials	*M*	*SD*	Skew	Kurtosis	α
SUS	12	75.34	11.36	−0.411	−0.506	0.775
Reading span task	7	21.61	4.35	0.092	1.174	0.623
Operation span task	6	20.69	8.51	−0.432	−0.304	0.755
Minus 2 span task	12	48.19	11.63	−0.317	0.637	0.784
Alphabet working memory task	18	2.79	0.86	0.662	0.927	0.542
Sentence recall in Finnish	5	77.85	10.59	1.015	1.037	0.710
Sentence recall in English	5	68.19	15.43	−0.651	−0.187	0.828
Paragraph recall	2	78.73	20.41	−0.403	−0.120	0.880
Paragraph semantic recall	2	21.29	4.98	−0.800	0.258	0.766
Word fluency	3	61.45	12.48	−0.201	0.861	0.897

Note: The range of possible task scores is 0–100 for the SUS task, 0–35 for the reading span task, 0–39 for the operation span task, 0–78 for the minus 2 span task, 0–∞ for the alphabet working memory task, 0–100 for the two sentence recall tasks, 0–116/118* for paragraph recall, 0–27 for paragraph semantic recall, and 0–∞ for the word fluency task. The infinity symbol (∞) indicates the upper limit of the score scale is infinite. In paragraph recall, test version 1 had a maximum score of 116, and test version 2 a maximum score of 118. In the alphabet working memory task, two of the item variables had zero variance and they were thus removed from the Cronbach alpha calculations.

Table 6. *Intercorrelations for all tasks in Experiment 2*

Measure	1	2	3	4	5	6	7	8	9	10
1. SUS	—									
2. RSpan	.412**	—								
3. OSpan	.300*	.205	—							
4. M2Span	.329**	.453**	.360*	—						
5. AWM	.279*	.148	.289*	.260*	—					
6. SR_Fin	.535**	.352**	.393**	.354**	.170	—				
7. SR_Eng	.413**	.352**	.393**	.280*	.241*	.643**	—			
8. Para_Rec	.254*	.392**	.102	.225*	.063	.417**	.306**	—		
9. Para_Sem	.284*	.359**	.110	.239*	.069	.391**	.210	.931**	—	
10. W_fluency	.231*	.289**	.212	.348**	.109	.272*	.089	.157	.173	—

Note: $N = 71$–80. SUS, selective updating of sentences task; RSpan, reading span task; OSpan, operation span task; M2Span, minus 2 span task; AWM, alphabet working memory task; SR_Fin, sentence recall in Finnish; SR_Eng, sentence recall in English; Para_Rec, paragraph recall; Para_Sem, paragraph semantic recall; W_fluency, word fluency task. *$p < .05$. **$p < .01$.

model explained an additional 6.7% of the variance and this change in R^2 was significant, $F\ (3, 67) = 6.365$, $p = .014$.

As regards the regression analysis with the paragraph recall task as the outcome variable, the CSTs contributed significantly, $F\ (2, 68) = 5.417$, $p = .007$, to the

https://doi.org/10.1017/S0142716417000182 Published online by Cambridge University Press

Table 7. *Summary of hierarchical regression analysis for variables predicting sentence recall in Finnish, sentence recall in English, paragraph recall, and paragraph semantic recall*

	Sentence Recall Tasks				Paragraph Recall Tasks			
	SR_Fin		SR_Eng		Para_Rec		Para_Sem	
Predictor	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2	β
Step 1	.231***		.233***		.137**		.140**	
Reading span task		0.283*		0.287*		0.364**		0.366**
Operation span task		0.335**		0.334**		0.027		0.035
Step 2	.148***		.067*		.012		.022	
Reading span task		0.103		0.167		0.313*		0.297
Opration span task		0.238*		0.269*		0.000		−0.002
SUS		0.443***		0.297*		0.125		0.171

Note: In the operation span task, 9 participants scored below the binomial cutoff score in the intervening task. Thus, the intercorrelations in the operation span task includes 71 participants ($N = 71$); excluded cases listwise. SR_Fin, sentence recall in Finnish; SR_Eng, sentence recall in English; Para_Rec, paragraph recall; Para_Sem, paragraph semantic recall.
*$p < .05$. **$p < .01$. ***$p < .001$.

regression model and explained 13.7% of the variance in task performance at Stage 1. Adding the SUS task at Stage 2 did not significantly increase R^2, F (3, 67) = 0.930, $p = .338$. In the paragraph semantic recall, Stage 1 explained 14.0% of the variance, leading to statistically significant model, F (2, 68) = 5.554, $p = .006$. Adding the SUS task at Stage 2 did not significantly change R^2, F (3, 67) = 1.750, $p = .190$.

Discussion of Experiment 2

The aim of Experiment 2 was to replicate and extend the findings in Experiment 1 by examining the test–retest reliability of the SUS task and investigating how the task was related to a broader array of WM and verbal episodic memory measures. In this experiment, the SUS task manipulations (updating stages, sentence length) showed the expected effects, the range of task difficulty was large, and the task exhibited an adequate internal consistency. Furthermore, the SUS task showed an adequate level of test–retest reliability over a 5-week period. An examination of the task intercorrelations showed that the SUS task correlated positively with all the WM tasks employed and showed the highest correlations with the sentence recall tasks (both in Finnish and in English) among all the WM tasks. Together with the CSTs and the minus 2 span task, the SUS task exhibited the strongest associations with the verbal episodic memory tasks employing sentences and paragraphs. Finally, hierarchical regression analysis showed that the SUS task shared some unique variance with the sentence recall tasks over and above the CSTs. Overall, the results from Experiment 2 concur with those of Experiment 1, supporting the conclusion that the SUS task is a reliable and valid verbal WM updating task.

GENERAL DISCUSSION

The present study set out to devise and test a novel verbal WM updating task. This was motivated by the notion that most commonly used WM tasks employ rather artificial materials (Moreau & Conway, 2014), and more "real-life" tasks are rare. Hence, we sought to validate a novel WM task, the SUS task, which operated with semantically and syntactically well-formed sentences rather than commonly used unrelated pieces of information such as random digit or letter series. The results from Experiment 1 indicated that the SUS task exhibited adequate internal consistency. Moreover, it demonstrated concurrent validity through statistically significant correlations with well-established WM tasks (three complex span tasks). The results of Experiment 2 confirmed the findings from Experiment 1, and showed further that the SUS task was related to sentence- and paragraph-level verbal episodic memory. Based on these results, we conclude that the SUS task is a reliable and valid verbal WM updating task that also taps on memory for more real-life language materials.

While the SUS task exhibited the expected positive correlations with other WM and verbal episodic memory measures, it should be noted that these correlations and hence the shared variance was not particularly high. This issue is not related specifically to the SUS task, as the intercorrelations between the other WM measures were in general at the same level, and followed earlier findings reported in

the literature (Foster et al., 2015; Gonthier et al., 2016; Oswald et al., 2015). The overall modest intercorrelations between different WM tasks (and executive tasks in general) may be due to a number of factors, such as paradigm- and stimulus-specific effects, availability of effective performance strategies, and measurement error. These tasks are inherently rather complex, and they are thus bound to reflect the interplay of several factors. As the SUS task is specifically calling for updating, the cognitive demands it makes differ from, for example, span tasks that should tap more on the WM storage capacity.

Besides shared WM resources, the observed associations between the SUS task and the sentence recall tasks might also be in part related to a common performance strategy, namely, chunking. It is well known that chunking is utilized in sentence recall: in general, individuals can memorize up to 2–3 times more words in a natural sentence compared to an arbitrary string of words or digits (Brener, 1940), thus far outweighing the capacity limits associated with WM. In addition, in WM updating, Solopchuck et al. (2016) observed recently that stronger reliance on chunking was associated with higher scores in a WM updating task with digit series that had to be constantly updated. As the SUS task is constructed in a similar fashion, use of chunking may certainly improve the task score here as well. Chunking is a fundamental mechanism in language and has been shown to predict variances in online processing of complex sentences (McCauley & Christiansen, 2015).

In line with previous research, we expected that an increase in updating stages would decrease sentence recall in our SUS task (Solopchuk et al., 2016; Waris et al., 2015), and the same would be true for increasing sentence length as well. These expectations were borne out in the results, but we also noted that there was variability in the recall performances. For example, some sentences containing less updating stages tended to be more challenging than sentences with more updating stages (e.g., the trial with five words and three updates was less accurately recalled than the trial with five words and five updates). One reason for this could be that the location of the updating stages in the sentences varied across trials, and the update locations in some sentences could have been more challenging than in others. For example, one would expect that verb updates are more demanding than noun updates, as verbs are in general more difficult to remember than nouns due to their abstractness and lesser degree of specificity (Earles & Kersten, 2000; Huttenlocher & Lui, 1979). With regard to sentence length, comprehension of longer sentences requires more effort due to an increasing distance between the argument and its head (Gibson, 2000; Levy & Keller, 2013; Lewis & Vasishth, 2005; McElree, Foraker, & Dyer, 2003), and it has been argued that verbal WM capacity is particularly important in dependency resolution (Nicenboim, Vasishth, Gattei, Sigman, & Kliegl, 2015). However, in our task the overall performance decline with increasing sentence length was not totally uniform. In particular, the recall of the five- and six-word sentences was quite similar. This may be due to several reasons. As noted above, the location of updates in sentences may play a role. In addition, linguistic factors such as syntactic complexity (Humphries, Binder, Medler, & Liebenthal, 2006), the number of high/low-frequency words (Rayner & Duffy, 1986), and morphological features (Allen, Badecker, & Osterhout, 2003) relevant to sentence processing may in part contribute to this variability. Nonetheless, we found a significant main effect of sentence length on recall performance,

and this factor is important for maintaining a wide range of task difficulty and thus providing sensitivity to detect individual differences in WM capacity.

A debated issue regarding verbal WM tasks is to what degree performance on such tasks is affected by individual differences in language exposure rather than WM functioning (MacDonald & Christiansen, 2002; Payne et al., 2014; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). While it has been suggested that performance on verbal WM tasks such as the reading span task relies heavily on linguistic proficiency that is driven by differences in prior language exposure (MacDonald & Christensen, 2002), there is also contradictory evidence indicating a lack of a relationship between language exposure and verbal WM performance (Payne et al., 2014). It is thus unclear whether performance on the SUS task would be at least to some extent driven by individual differences related to language exposure. As the present study showed that performance on the SUS task is related to both CSTs and verbally based episodic memory tasks, the SUS task performance may reflect both WM and language abilities.

Limitations and future directions

A possible limitation in Experiment 1 was that the data was collected via the Internet. Despite the advantages of Internet testing, several concerns such as untruthful or careless responding exist, as the researcher does not have control over the testing environment (Feitosa, Joseph, & Newman, 2015; Smith, Roster, Golden, & Albaum, 2016). There is, however, evidence indicating that online cognitive task performance is in general reliable and valid, comparing well to traditional laboratory studies (Crump, McDonnell, & Gureckis, 2013; Enochson & Culbertson, 2015; Germine et al., 2012; Linnman, Carlbring, Åhman, Andersson, & Andersson, 2006). For example, Crump et al. (2013) administered a battery of commonly used cognitive tasks, utilizing the crowdsourcing service Amazon Mechanical Turk. The results showed that the task performance was very similar to those in laboratory settings. These findings are also supported by the present study, as the performances on the WM tasks were very similar in terms of means, standard deviations, and internal consistency to the task performances in Experiment 2. In summary, there is no evidence that well-instructed Internet testing would suffer from any major reliability or validity problems, whereas its benefits are unquestionable: it is cost-efficient and enables access to participants from very diverse backgrounds.

One limitation in this study concerns the verbal episodic memory tasks employed in Experiment 2. Even though the reliability analysis revealed adequate levels of internal consistency in the verbal episodic memory tasks, no further validation was performed. Therefore, it would be important for future studies to compare the SUS task against more ecologically valid measures such as parent questionnaires or some standardized verbal ability measures such as the verbal SAT. Another limitation of this study relates to the ecological validity of the SUS task. Although the aim of the present study was to validate a somewhat more naturalistic verbal WM task, it is evident that the SUS task is still rather artificial when compared with everyday sentence processing. Rather, it may be located in between classical

laboratory tasks on WM and everyday language performances, and would benefit from further validation attempts, for example, against discourse and narrative production measures.

When comparing the SUS task features against traditional CSTs, perhaps the most prominent difference is the manipulation of the updating component in the SUS task. In the CSTs, the storage load (number of maintained items) increases, but the processing load (the intervening task) is held constant throughout the task. In our SUS task, in contrast, both the storage component (i.e., number of words to be remembered) and the processing component (i.e., the number of updates) are varied. In future versions of the SUS task, it would be worthwhile to manipulate these two components separately, as has been done with CSTs (Archibald & Harder-Griebling, 2015).

Summary and conclusions

This study introduced a novel verbal WM updating measure coined as the SUS task. In this task, participants operate with semantically plausible sentences, thus adding to the naturalness of the task. Two experiments indicated that the SUS task has adequate psychometric properties and that it is related both to other WM tasks and to verbal episodic memory at the sentence and paragraph levels. We hope that it will be a useful addition to the psycholinguistic toolbox of measures that tap into short-term memory processes during language processing.

ACKNOWLEDGMENTS

We thank Daniel Wärnå, Otto Waris, and the rest of the BrainTrain group for their help with the study. Matti Laine was supported by grants from the Academy of Finland (Project 260276) and the Abo Akademi University Endowment (the BrainTrain project).

NOTES

1. In Finland, all students study at least two foreign languages, mainly English and obligatory Swedish (the second official language), up to high school. Thus, most of young Finnish adults have at least satisfactory skills in English when entering university.
2. Both experiments also included a self-paced variant of the sentence recall task. This task version, however, resulted in great variance between participants in the reading time of the to be remembered sentences. This was considered problematic in terms of task validity, and therefore, the self-paced version of the task was excluded from both experiments.

REFERENCES

Allen, M., Badecker, W., & Osterhout, L. (2003). Morphological analysis in sentence processing: An ERP study. *Language and Cognitive Processes, 18*, 405–430. doi:10.1080/01690960244000054

Archibald, L. M., & Harder Griebeling, K. (2015). Rethinking the connection between working memory and language impairment. *International Journal of Language & Communication Disorders, 51*, 252–264. doi:10.1111/1460-6984.12202

Archibald, L. M., & Joanisse, M. F. (2009). On the sensitivity and specificity of nonword repetition and sentence recall to language and memory impairments in children. *Journal of Speech, Language, and Hearing Research, 52*, 899–914. doi:10.1044/1092-4388(2009/08-0099)

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417–423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*, 189–208. doi:10.1016/S0021-9924(03)00019-4

Baddeley, A., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8*, 47–89. doi:10.1016/S0079-7421(08)60452-1

Benton, A., & Hamsher, K. (Eds.). (1978). *Multilingual aphasia examination* (rev.) Iowa City, IA: University of Iowa Hospitals, Department of Neurology.

Brener, R. (1940). An experimental investigation of memory span. *Journal of Experimental Psychology, 26*, 467–482. doi:10.1037/h0061096

Carretti, B., Borella, E., Cornoldi, C., & De Beni, R. (2009). Role of working memory in explaining the performance of individuals with specific reading comprehension difficulties: A meta-analysis. *Learning and Individual Differences, 19*, 246–251. doi:10.1016/j.lindif.2008.10.002

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

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman & Hall.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PlOS ONE, 8*, e57410. doi:10.1371/journal.pone.0057410

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

Earles, J. L., & Kersten, A. W. (2000). Adult age differences in memory for verbs and nouns. *Aging, Neuropsychology, and Cognition, 7*, 130–139. doi:10.1076/1382-5585(200006)7:2;1-U;FT130

Ecker, U. K., Lewandowsky, S., & Oberauer, K. (2014). Removal of information from working memory: A specific updating process. *Journal of Memory and Language, 74*, 77–90. doi:10.1016/j.jml.2013.09.003

Ecker, U. K., Lewandowsky, S., Oberauer, K., & Chee, A. E. (2010). The components of working memory updating: An experimental decomposition and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 170–189. doi:10.1037/a0017891

Ecker, U. K., Oberauer, K., & Lewandowsky, S. (2014). Working memory updating involves item-specific removal. *Journal of Memory and Language, 74*, 1–15. doi:10.1016/j.jml.2014.03.006

Enochson, K., & Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PLOS ONE, 10*, e0116946. doi:10.1371/journal.pone.0116946

Everitt, B. (2002). *The Cambridge dictionary of statistics*. Cambridge: Cambridge University Press.

Feitosa, J., Joseph, D. L., & Newman, D. A. (2015). Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English. *Personality and Individual Differences, 75*, 47–52. doi:10.1016/j.paid.2014.11.017

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*, 226–236. doi:10.3758/s13421-014-0461-7

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., Defries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science, 17*, 172–179. doi:10.1111/j.1467-9280.2006.01681.x

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*, 847–857. doi:10.3758/s13423-012-0296-9

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, MA: MIT Press.

Gonthier, C., Thomassin, N., & Roulin, J. (2016). The composite complex span: French validation of a short working memory task. *Behavior Research Methods, 48*, 233–242. doi:10.3758/s13428-015-0566-3

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

Holmes, J. (2011). Baby brain: Training executive control in infancy. *Current Biology, 21*, R684–R685. doi:10.1016/j.cub.2011.08.026

Holmes, J., & Gathercole, S. E. (2014). Taking working memory training from the laboratory into schools. *Educational Psychology, 34*, 440–450. doi:10.1080/01443410.2013.797338

Humphries, C., Binder, J., Medler, D., & Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of Cognitive Neuroscience, 18*, 665–679. doi:10.1162/jocn.2006.18.4.665

Huttenlocher, J., & Lui, F. (1979). The semantic organization of some simple nouns and verbs. *Journal of Verbal Learning and Verbal Behavior, 18*, 141–162. doi:10.1016/S0022-5371(79)90091-4

Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeys, I., & Kwapil, T. R. (2007). For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. *Psychological Science, 18*, 614–621. doi:10.1111/j.1467-9280.2007.01948.x

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 615–622. doi:10.1037/0278-7393.33.3.615

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

Kessler, Y., & Meiran, N. (2006). All updateable objects in working memory are updated whenever any of them are modified: Evidence from the memory updating paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 570–585. doi:10.1037/0278-7393.32.3.570

Kessler, Y., & Meiran, N. (2008). Two dissociable updating processes in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1339–1348. doi:10.1037/a0013078

Kessler, Y., & Oberauer, K. (2015). Forward scanning in verbal working memory updating. *Psychonomic Bulletin & Review, 22*, 1770–1776. doi:10.3758/s13423-015-0853-0

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. H., Gustafsson, J., & Hulme, C. (2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science, 18*, 146–154. doi:10.1111/desc.12202

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences, 14*, 317–324. doi:10.1016/j.tics.2010.05.002

Lechuga, M. T., Pelegrina, S., Pelaez, J. L., Martin-Puga, M. E., & Justicia, M. J. (2016). Working memory updating as a predictor of academic attainment. *Educational Psychology, 36*, 675–690. doi:10.1080/01443410.2014.950193

Levy, R. P., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language, 68*, 199–222. doi:10.1016/j.jml.2012.02.005

Lewis, R., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*, 375–419. doi:10.1207/s15516709cog0000_25

Lewis, R., Vasishth, S., & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences, 10*, 447–454. doi:10.1016/j.tics.2006.08.007

Linnman, C., Carlbring, P., Åhman, Å, Andersson, H., & Andersson, G. (2006). The Stroop effect on the internet. *Computers in Human Behavior, 22*, 448–455. doi:10.1016/j.chb.2004.09.010

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109*, 35–54. doi:10.1037//0033-295X.109.1.35

Mason, R., & Just, M. (2007). Lexical ambiguity in sentence comprehension. *Brain Research, 1146*, 115–127. doi:10.1016/j.brainres.2007.02.076

McCauley, S. M., & Christiansen, M. H. (2015). Individual differences in chunking ability predict online sentence processing. In D. C. Noelle, T. Matlock, R. Dale, C. Jennings, A. Warlaumont, P. P. Maglio, & J. Yoshimi (Eds.), *37th Annual Meeting of the Cognitive Science Society (CogSci 2015): Mind, technology, and society* (pp. 1553–1558). Austin, TX: Cognitive Science Society.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language, 48*, 67–91. doi:10.1016/S0749- 596X(02)00515-6

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49–100. doi:10.1006/cogp.1999.0734

Miyake, A., Just, M., & Carpenter, P. (1994). Working memory constraints on the resolution of lexical ambiguity: Maintaining multiple interpretations in neutral contexts. *Journal of Memory and Language, 33*, 175–202. doi:10.1006/jmla.1994.1009

Moreau, D., & Conway, A. R. A. (2014). The case for an ecological approach to cognitive training. *Trends in Cognitive Sciences, 18*, 334–336. doi:10.1016/j.tics.2014.03.009

Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of the central executive. *British Journal of Psychology, 81*, 111–121. doi:10.1111/j.2044-8295.1990.tb02349.x

Murty, V. P., Sambataro, F., Radulescu, E., Altamura, M., Iudicello, J., Zoltick, B., . . . Mattay, V. S. (2011). Selective updating of working memory content modulates meso-cortico-striatal activity. *NeuroImage, 57*, 1264–1272. doi:10.1016/j.neuroimage.2011.05.006

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology, 6*, 312. doi:10.3389/fpsyg.2015.00312

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 411–421. doi:10.1037/0278-7393.28.3.411

Oberauer, K., Süß, H., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence, 31*, 167–193. doi:10.1016/S0160-2896(02)00115-0

Oberauer, K., Weidenfeld, A., & Hörnig, R. (2006). Working memory capacity and the construction of spatial mental models in comprehension and deductive reasoning. *Quarterly Journal of Experimental Psychology, 59*, 426–447. doi:10.1080/17470210500151717

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*, 1343–1355. doi:10.3758/s13428-014-0543-2

Palladino, P., Cornoldi, C., De Beni, R., & Pazzaglia, F. (2001). Working memory and updating processes in reading comprehension. *Memory & Cognition, 29*, 344–354. doi:10.3758/BF03194929

Payne, B. R., Grison, S., Gao, X., Christianson, K., Morrow, D. G., & Stine-Morrow, E. A. (2014). Aging and individual differences in binding during sentence understanding: Evidence

from temporary and global syntactic attachment ambiguities. *Cognition, 130*, 157–173. doi:10.1016/j.cognition.2013.10.005

Poll, G. H., Miller, C. A., & van Hell, J. G. (2016). Sentence repetition accuracy in adults with developmental language impairment: Interactions of participant capacities and sentence structures. *Journal of Speech, Language, and Hearing Research, 59*, 302–316. doi:10.1044/2015_JSLHR-L-15-002

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*, 191–201. doi:10.3758/BF03197692

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*, 164–171. doi:10.1027/1015-5759/a000123

Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review, 20*, 1102–1113. doi:10.3758/s13423-013-0453-9

Riches, N. G., Loucas, T., Baird, G., Charman, T., & Simonoff, E. (2010). Sentence repetition in adolescents with specific language impairments and autism: An investigation of complex syntax. *International Journal of Language & Communication Disorders, 45*, 47–60. doi:10.3109/13682820802647676

Ricketts, J., Sperring, R., & Nation, K. (2014). Educational attainment in poor comprehenders. *Frontiers in Psychology, 5*, 445. doi:10.3389/fpsyg.2014.00445

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science, 24*, 1301–1308. doi:10.1177/0956797612466268

Sanchez, C. A., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition, 34*, 344–355. doi:10.3758/BF03193412

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1089–1096. doi:10.1037/a0015730

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: Putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in Psychology, 5*, 1475. doi:10.3389/fpsyg.2014.01475

Shipstead, Z., Redick, T., & Engle, R. (2010). Does working memory training generalize? *Psychologica Belgica, 50*, 245–276. doi:10.5334/pb-50-3-4-245

Smith, S. M., Roster, C. A., Golden, L. L., & Albaum, G. S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research, 69*, 3139–3148. doi:10.1016/j.jbusres.2015.12.002

Solopchuk, O., Alamia, A., Olivier, E., & Zenon, A. (2016). Chunking improves symbolic sequence processing and relies on working memory gating mechanisms. *Learning & Memory, 23*, 108–112. doi:10.1101/lm.041277.115

Strauss, E., Sherman, E. M., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary*. Oxford: Oxford University Press.

Süß, H., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—And a little bit more. *Intelligence, 30*, 261–288. doi:10.1016/S0160-2896(01)00100-3

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. New York: HarperCollins.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson/Allyn & Bacon.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*, 127–154. doi:10.1016/0749-596X(89)90040-5

Unsworth, N. (2010). On the division of working memory and long-term memory and their relation to intelligence: A latent variable approach. *Acta Psychologica, 134*, 16–28. doi:10.1016/j.actpsy.2009.11.010

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133*, 1038–1066. doi:10.1037/0033-2909.133.6.1038

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

Waris, O., Soveri, A., & Laine, M. (2015). Transfer after working memory updating training. *PLOS ONE, 10*, e0138734. doi:10.1371/journal.pone.0138734

Was, C. A., Rawson, K. A., Bailey, H., & Dunlosky, J. (2011). Content-embedded tasks beat complex span for predicting comprehension. *Behavior Research Methods, 43*, 910–915. doi:10.3758/s13428-011-0112-x

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments, and Computers, 35*, 550–564. doi:10.3758/BF03195534

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology, 58*, 250–271. doi:10.1016/j.cogpsych.2008.08.002

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it. *Frontiers in Psychology, 4*, 1–22. doi:10.3389/fpsyg.2013.00433

Yu, Y., FitzGerald, T. H., & Friston, K. J. (2013). Working memory and anticipatory set modulate midbrain and putamen activity. *Journal of Neuroscience, 33*, 14040–14047. doi:10.1523/jneurosci.1176-13.2013

Ziethe, A., Eysholdt, U., & Doellinger, M. (2013). Sentence repetition and digit span: Potential markers of bilingual children with suspected SLI? *Logopedics Phoniatrics Vocology, 38*, 1–10. doi:10.3109/14015439.2012.664652