ACADEMIC
PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Memory and Language 48 (2003) 320–345

Journal of
Memory and
Language

www.elsevier.com/locate/jml

Individual differences in semantic short-term memory capacity and reading comprehension

Henk J. Haarmann,[a,*] Eddy J. Davelaar,[b] and Marius Usher[b]

[a] Department of Hearing and Speech Sciences, University of Maryland, College Park, MD 20742, USA
[b] School of Psychology, Birkbeck College, University of London, London WC1E 7HX, UK

Received 26 November 2001; revision received 25 February 2002

Abstract

We report three correlation studies, which investigate the hypothesis that individual differences in the capacity of a semantic short-term memory (STM) component in working memory (WM) predict performance on complex language tasks. To measure the capacity of semantic STM, we devised a storage-only measure, the conceptual span, which makes use of a category-cued recall procedure. In the first two studies, where the conceptual span was administered with randomized words (not blocked by categories), we found that conceptual span predicted single-sentence and text comprehension, semantic anomaly detection and verbal problem solving, explaining unique variance beyond non-word and word span. In some cases, the conceptual span explained unique variance beyond the reading span. Conceptual span correlated better with verbal problem solving than reading span, suggesting that a storage-only measure can outperform a storage-plus-processing measure. In Study 3, the conceptual span was administered with semantically clustered lists. The clustered span correlated with the comprehension measures as well as the non-clustered span, indicating that the critical process is memory maintenance and not semantic clustering. Moreover, we found an interaction between subjects' performance on the conceptual span and the effect of the distance between critical words in anomaly detection, supporting the proposal that semantic STM maintains unintegrated word meanings.

Keywords: Short term memory; Working memory; Reading comprehension; Individual differences; Conceptual span; Reading span

It is widely recognized that a limited-capacity working memory (WM) system plays an important role in complex cognition, supporting both the temporary storage and processing of information (for a review see Kintsch, Healy, Hegarty, Pennington, & Salthouse, 1999). A seminal study by Daneman and Carpenter (1980) demonstrated the importance of WM in the domain of language processing. Its major finding was that a storage-plus-processing measure of WM, *the reading span*, predicted accuracy of text comprehension (see also Baddeley, Logie, Nimmo-Smith, & Brereton, 1985; Budd, Whitney, & Turley, 1995; Daneman & Carpenter, 1983; Dixon, Le Fevre, & Twilley, 1989; Engle, Cantor, & Carullo, 1992; LaPointe & Engle, 1990; Masson & Miller, 1983), while a storage-only measure, *the word span*, did not (see also Turner & Engle, 1989). Moreover, when a statistically significant correlation between word span and comprehension is obtained, it tends to be smaller than the correlation between reading span and comprehension (LaPointe & Engle, 1990). The reading span test determines the number of sentence-final words a person can recall immediately after reading aloud a set of sentences and thus emphasizes both storage and processing of words. By contrast, the word span is a storage-only measure, which determines the number of

[*] Corresponding author. Fax: 1-301-314-2023.
E-mail addresses: hhaarmann@hesp.umd.edu (H.J. Haarmann), e.davelaar@psychology.bbk.ac.uk (E.J. Davelaar), m.usher@bbk.ac.uk (M. Usher).

PII: S0749-596X(02)00506-5

words a person can recall in exact serial order immediately after their presentation. Consistent with the WM interpretation of the reading span test, Daneman and Carpenter (1980) found that the ability of low span readers to answer a question about the referent of a pronoun showed a marked deterioration when the number of sentences intervening between the referent and the pronoun was increased, while no such effect was present for high span readers. Furthermore, reading span is a good predictor of word reading times in sentence comprehension (for a review see Just & Carpenter, 1992; Miyake, Just, & Carpenter, 1994b).

While the correlation between comprehension and reading span is well established, its WM interpretation has been subject to debate (Baddeley et al., 1985; Daneman & Merikle, 1996). Jackson and McClelland (1979) found that listening comprehension is one of the most important predictors of reading comprehension. Together with a measure of letter matching it predicted 77% of the variance. Citing this result, Baddeley et al. (1985) asked "How should the correlation between comprehension and working memory span be interpreted?" and suggested that

> The original Daneman and Carpenter result was open to a range of interpretations, from the strong suggestion that their task was a measure of the capacity of a general working memory system, to the relatively weak interpretation that since working memory span itself depended on comprehension, that they were basically replicating the observation of Jackson and McClelland that listening comprehension was the best predictor of reading comprehension (Baddeley et al., 1985, pp. 129–130).

Thus, it seemed possible that the correlation between comprehension and reading span was somewhat trivial, that is, due to a great deal of overlap between the two tasks, in particular, their common sentence processing component. This interpretation was further discussed by Daneman and Merikle (1996) who rephrased it as: "*sentence comprehension (reading span ...) correlates with paragraph comprehension (the criterion comprehension tests)*" (p. 424). Indeed, Daneman and Carpenter (1980) suggested that the inclusion of a sentence processing task in both reading span and comprehension may be crucial to the ability of the reading span to predict comprehension. In particular, they suggested that in both tasks the temporary storage of verbal information could have been worse for poor readers who had to devote some of their limited WM resources to compensate for inefficient reading processes.

Subsequent research has ruled out this weak interpretation of the correlation between comprehension and reading span and has provided support for the stronger WM interpretation (Conway & Engle, 1996; Daneman & Merikle, 1996; Engle et al., 1992; Turner & Engle, 1989). Turner and Engle (1989) found that the background task of the complex span measure does not have to involve sentence processing in order to predict reading comprehension. They found that operation span predicts comprehension as well as reading span. Operation span measures the number of words or digits a person can retain while verifying a sequence of arithmetic problems presented in alternation with the to-be-retained words or digits. More generally, Daneman and Merikle (1996) concluded from a meta-analysis of data from 77 studies that complex span measures, which include a storage and processing component (e.g., reading span, operation span), predict comprehension better than storage-only measures (e.g., word span, digit span), even if the processing component of the complex span task does not involve manipulation of words and sentences. Furthermore, both reading span and operation span still predict comprehension, but to a lesser extent, when individual differences in processing efficiency are statistically controlled (Conway & Engle, 1996; Engle et al., 1992). This finding makes it unlikely that individual differences in complex span are solely due to inter-subject variation in the degree to which a constant capacity is allocated to compensate for differences in processing efficiency. Instead, such differences appear to reflect differences in the capacity of a WM system that supports both storage and processing (Just & Carpenter, 1992).

A shared assumption of all current models of WM is that it is a multi-component system (for a review see Kintsch et al., 1999). However, it is currently not well understood in what component of the WM system individual capacity differences predicting comprehension reside. One possibility is that there are individual differences in the capacity of a WM component that is crucial for dual tasking, that is, the ability to coordinate the performance of two tasks. For example, it has been suggested that individual differences in complex span performance could reflect differences in capacity to alternate attention between different tasks (Kane & Engle, 2000). This view correctly predicts that complex span measures (reading span, operation span) are better predictors of comprehension than simple span measures (word span, digit span), because the former but not the latter type of task involves a dual-task component.

It is nevertheless possible that a storage component plays also a role in predicting comprehension and that the impact of this factor has been underestimated, because most of the studies relied on a phonological measures of span (digit/word span using serial order recall). Recently a number of authors have suggested that the storage of verbal information is supported not only by a phonological short-term memory (STM) (Baddeley, 1986) but also by a semantic STM (Haarmann, Cameron, & Ruchkin, in press; Haarmann & Usher, 2001; Hanten & Martin, 2000; Martin & Freedman, 2001; Martin & Romani, 1994; Martin, Saffran, & Dell, 1996; Martin, Shelton, & Yaffee, 1994; Potter, 1993; Romani & Martin, 1999). In particular, a series of neuropsychological studies by R.C. Martin and her colleagues has provided strong evidence for separate phonological and semantic STM components in WM (Hanten & Martin, 2000; Martin & Freedman, 2001; Martin & Romani, 1994; Martin et al., 1994; Romani & Martin, 1999). Across a series of immediate recall tasks, two patients, E.A. and A.B., showed evidence for a double dissociation between phonological and semantic STM impairment (Martin et al., 1994). While the performance of E.A. indicated greater phonological than semantic STM deficit, A.B. showed the opposite pattern of deficit. For example, E.A.'s performance on an immediate probed recognition task was markedly lower with rhyme probes than with semantic category probes, while A.B. showed the reverse pattern. Unlike E.A., A.B. also showed a normal word length effect (i.e., better recall for short than long words) and a normal modality effect (i.e., better recall with spoken than written presentation), indicating a relatively preserved phonological STM system. Also, unlike E.A., A.B. did not show a normal lexicality effect (i.e., better recall of words than non-words), which suggests a semantic STM deficit (Martin et al., 1994).

Additional evidence for a semantic STM comes from functional neuro-imaging studies, which found that the dorso-lateral prefrontal cortex (DL-PFC) is activated by a semantic but not by a phonological working memory task with identical response demands (Crosson et al., 1999; Gabrieli, Poldrack, & Desmond, 1998). This finding suggests that the DL-PFC may help to sustain the activation of semantically-sensitive item representations (Haarmann & Usher, 2001; Usher & Cohen, 1999). While semantic effects in STM have been demonstrated long ago (Raser, 1972; Shulman, 1972), they have been typically attributed to LTM contributions to recall (Baddeley, 1972; Crowder, 1979). Recently however, Haarmann and Usher (2001) reported semantic effects in immediate recall that cannot be attributed to LTM contributions. Pairs of weak semantic associates were better recalled at recency positions when they were close together in a list than when they were far apart and this semantic-separation effect was much larger in immediate recall than in delayed recall. In addition, Haarmann and Usher (2001) found that the semantic separation effect in immediate recall is still obtained and of similar magnitude when encoding in phonological STM is prevented through articulatory suppression (Baddeley, 1986). They therefore argued that the effect arose in semantic STM, in line with the neuropsychological evidence for such a WM component (Martin & Freedman, 2001).

Accordingly, we believe that there is now compelling evidence that, while phonological STM stores phonologically decaying traces that are refreshed through subvocal rehearsal, semantic STM stores lexical-semantic item representations (i.e., word meanings) that are actively maintained until they can be integrated into a meaning relation with words that occur later in the sentence (Gunter, Jackson, & Mulder, 1995; Haarmann et al., in press; Miyake et al., 1994b). Moreover, while neuropsychological dissociations indicate that an intact phonological STM is not crucial for on-line sentence comprehension (Butterworth, Campbell, & Howard, 1986; Caplan & Waters, 1990, 1999; Martin, 1990; Martin & Romani, 1994), the semantic component seems to play an important role in this process. The larger an individual's semantic STM capacity, the better the chances the meaning of a to-be-integrated word is maintained during the processing of words that intervene between it and words with which it is to be integrated (see Haarmann, Just, & Carpenter, 1997, for a computational model). Thus, we expect semantic STM to be a better and more reliable predictor of comprehension than phonological STM. This expectation is not necessarily at odds with the observation that storage-only measures (i.e., word span and digit span) are poor predictors of comprehension (Daneman & Merikle, 1996), because the previously used storage-only measures rely primarily on phonological STM (Baddeley, 1986) and may not rely much on semantic STM.

Conceptual span test

In the present study, we used a category cued-recall test as a relative index of the capacity of the semantic STM component of WM. Since this system supports the maintenance of concepts associated with words, we will refer to the category-cued recall test as the "conceptual span" test. On each trial in the test, participants silently read a randomly ordered list of nine words, consisting of nouns in three different semantic categories with three nouns per category. Immediately following the presentation of the last word, the name of one of the three categories appears and participants attempt to recall aloud all three words in that category (e.g., *lamp*, *pear*, *tiger*, *apple*, *grape*, *elephant*, *horse*, *fax*, *phone*, *FRUIT?* Correct answer: *apple*, *pear*, *grape*). Their score is the average number of words they could recall out of three words across a series of such trials.

The conceptual span test was designed so as to minimize the contribution of LTM to task performance and maximally engage STM. First, the words are presented at a relatively fast rate (i.e., one word per second) in order to minimize participants' ability to encode the items into a script-type representation in LTM and thereby engage their STM system to a larger extent (Cowan, 2001; Haarmann & Usher, 2001). This presentation rate is thought to be fast enough to enable semantic encoding of individual words in semantic STM (Potter, 1993). Second, participants read all words twice immediately prior to the start of the test, and are presented the word materials during the test from a fixed word pool. Such repeated exposure is likely to induce proactive interference (PI), which affects retrieval from LTM more than from STM (Craik & Birtwistle, 1971; Halford, Maybery, & Bain, 1988) and which may, therefore, help to promote the use of STM rather than LTM in the conceptual span test. Following Cowan (1999, 2001), we regard STM as the capacity-limited, activated part of LTM and assume that PI affects retrieval of inactive representations in LTM.

Investigating the role of the semantic STM component of WM for comprehension by means of the conceptual span test has several advantages. First, the test does not include a dual-task requirement. As a result, positive correlations between conceptual span and comprehension cannot be attributed to individual differences in the ability to perform a dual-task or, alternatively, to an individual's willingness to shift attention away from the secondary processing task to the primary storage task (Waters & Caplan, 1996). In addition, since the conceptual span test does not involve any sentence processing, correlations between conceptual span and comprehension accuracy cannot be attributed to individual differences in sentence processing efficiency, which might arise for example at a non-semantic, syntactic level and which may not be correlated with individual differences in semantic STM capacity. The latter possibility is consistent with the finding of a double dissociation between syntactic judgment ability and semantic STM deficit in brain-damaged patients (Martin & Romani, 1994). Furthermore, the use of a category cued recall procedure is likely to engage semantic STM and minimize the contribution of phonological STM and its sub-vocal rehearsal component. The latter is suggested by the lack of a word length effect in category-cued recall (Haarmann & Usher, 2001) and the relative preservation of category-cued recognition performance in patients with a severe phonological but mild semantic STM deficit (Martin et al., 1994). Moreover, the preexposure to, and repeated use of the words in the conceptual test makes it unlikely that obtained performance differences reflect inter-individual variation in the efficiency of word encoding processes.

We present here three correlation studies with college-age adults, whose major aim was to measure the contribution of a semantic STM component of WM to language comprehension. The first two studies included the conceptual span test, the reading span test, a series of span tests and a series of comprehension tests. The third study was designed to test the alternative view that the conceptual span test measures clustering ability[1] instead of semantic STM capacity. Moreover, the third study was designed to test the hypothesis that semantic STM maintains unintegrated word meanings to support their on-line integration during sentence processing (Martin & Romani, 1994). Given the capacity-limited nature of semantic STM, this hypothesis predicts an interaction between subjects' conceptual span performance and the effect of the distance between critical words in on-line anomaly detection, such that participants with a low conceptual span show larger distance effects than participants with high conceptual span.

Study 1

Study 1 included the conceptual span test, the reading span test, the word span test, a sentence comprehension test, and a text comprehension test. In accordance with previous results (Daneman & Carpenter, 1980; LaPointe & Engle, 1990; Turner & Engle, 1989), we expected reading span to be a better predictor of sentence and text comprehension than word span. Furthermore, we reasoned that the conceptual span test provides a better measure of the semantic STM component in WM than word span and that this component may be an important determinant of the performance on the reading span test. We therefore expected conceptual span to predict sentence and text comprehension better than word span and possibly as well as reading span. We also investigated whether conceptual span accounts for unique variance in text comprehension above and beyond the variance contributed by reading span. Such a finding could indicate that conceptual span provides a more sensitive measure of semantic STM than reading span. Alternatively, such a finding could indicate that conceptual span measures semantic STM, whereas reading span measures some other ability, such as, domain-specific sentence processing skills or ability to control attention (Kane & Engle, 2000; Kane, Bleckley, Conway, & Engle, 2001).

Method

Participants. Sixty-six undergraduate students from the University of Maryland at College Park, all native speakers of English, participated. They received either a seven-dollar payment or extra-credit for their participation.

Tests

Each participant was tested individually and performed five tests, given in the same order, namely, conceptual span, reading span, word span, text comprehension, and sentence comprehension. A test session lasted about 1 h and 15 min. Three participants did not perform the final sentence comprehension test because of time constraints. Presentation of all tests was visual and computer controlled.

[1] We would like to thank Nelson Cowan for pointing out this alternative view to us.

1. *Conceptual span.* On each trial, participants silently read a sequence of nine nouns (in small letters) followed by a category name (in capital letters), presented at a computer-controlled rate of 1 word/s. The nine words consisted of three groups of three nouns, with each group belonging to a different semantic category, and were presented in a random order. Participants were instructed to try to recall aloud the three nouns in the named category in any order (e.g., *lamp*, *pear*, *tiger*, *apple*, *grape*, *elephant*, *horse*, *fax*, *phone FRUIT?* Answer: *pear grape apple*). The materials for the test came from a pool of 48 nouns with eight nouns in each of six semantic categories. The assignment of categories and nouns within categories to a trial sequence and the selection of the cued category within a trial sequence was random and with replacement. Prior to the test participants were shown each of the eight categories and its nouns and asked to read aloud the nouns while thinking of how it fit within the category. They did this twice in succession. The actual test consisted of two practice trials and 16 test trials. A participant's conceptual span was defined as the number of words recalled across the 16 test trials (the maximum possible score was 48).

2. *Reading span.* This test was an adapted version of the Daneman and Carpenter (1980) reading span test. On each trial, participants read aloud a set of sentences, presented one sentence at a time on a display monitor. As soon as the participants finished reading the last word in a sentence, the experimenter pushed a key that led to the display of the next sentence in the set. At the end of each set a question mark appeared and participants attempted to recall aloud all the sentence-final words in the set in their order of presentation. The set size varied from two to five sentences and there were two trials at each set size. A particular sentence occurred only once in the test, always ended in a concrete noun, and could be from 13 to 16 words long. (An example of a trial at set size 2 is *Josh wanted to finish his homework, but he forgot to go to the store. Chris liked being a sheriff, but he didn't like to wear the hat.* Answer: *store*, *hat*.) The reading span test started with two practice trials at set size 2 and the actual test began at this set size. Each time a participant answered one or two trials at a particular set size correct, the set size was increased with one sentence and participants were warned that such an increase would take place. Testing was discontinued if a participant got zero trials correct at a particular set size. A correct trial was one in which all the sentence-final words in a sequence of sentences were recalled in their order of presentation. A participant's reading span was defined as the total number of correct trials (the maximum possible score was 10).

3. *Word span.* On each trial, participants read aloud a set of words, presented at a computer-controlled rate of one word/s. Immediately after the offset of the last word, a question mark appeared and participants attempted to recall aloud all words in the set in their order of presentation. The length of the sequences varied from three to nine words and there were two trials at each set size. The different nouns were semantically and phonetically as unrelated as possible. The word span test started with two practice trials at set size two and the actual test began at this set size. Each time a participant answered one or two trials at a particular length correct, the length was increased by one word and participants were warned that such an increase would take place. Testing was discontinued if a participant got zero trials correct at a particular set size. A correct trial was one in which all the words in a set were recalled in their order of presentation. A participant's word span was defined as the total number of correct trials. A particular word occurred only once in the test and was always one-syllable long and a concrete noun.

4. *Text comprehension.* The materials were taken from a practice version of the Verbal Scholastic Aptitude Test (VSAT) and consisted of two written stories that were related in theme (i.e., the role of a mentor in the early education experiences of an artist). All participants indicated that they were not familiar with the two stories prior to the test. To avoid re-reading of the stories, the presentation mode was self-paced, line-by-line. With each press of a button, participants would replace the current line of text with the next line of text in the middle of the screen. The two stories consisted of a total of 120 lines of text, or 1180 words. The last line of the last story was followed by 13 written, multiple-choice questions, each of which required a combination of fact retrieval and inference making, either involving the first story, the second story, or a comparison of a similar theme in both stories. There were five answer alternatives per question. The display monitor showed only one question at a time together with its answer alternatives. Participants indicated their answer choice out loud and the experimenter recorded whether or not it was correct. The answer to a question could not be changed once the next question appeared on the display. The score on the story comprehension test was defined as the number of questions answered correctly (maximum possible score was 13).

5. *Sentence comprehension.* On each sentence comprehension trial participants read a stimulus sentence, followed by a verification statement, and indicated whether the statement made a true or false assertion about the meaning of the stimulus sentence. The details of the test were as follows.

Materials. We created 64 stimulus sentences with a main clause and a relative clause (e.g., *The nurse that thanked the doctor helped the patient*). Stimulus sentences varied in their syntactic complexity. They included subject-relative and (more complex) object-relative sentences

Table 1
Descriptive statistics: Study 1

Measure	*N*	Mean[a]	*SD*	Min	Max	Maximum possible score
Conceptual span[b]	66	28.92	5.77	13	40	48
Reading span	66	4.63	1.48	2	8	10
Word span	66	4.48	1.37	1	7	14
Text comprehension	66	5.27	2.08	1	10	13
Sentence comprehension	63	74.00	13.00	52	100	100

[a] The medians were identical to the means rounded to the nearest integer.

[b] The split-half reliability of the conceptual span test (i.e., correlation between scores on even and odd items) was .85 ($p < .001$) after Spearman–Brown correction for test length.

(for a review see Miyake, Carpenter, & Just, 1994a) with right-branching and (more complex) center-embedded relative clauses (for a review see Stromswold, Caplan, Alpert, & Rauch, 1996). The nouns, which referred to human actors, were semantically interchangeable and the degree of their semantic association with one another could be either strong or strong weak. Each stimulus sentence was paired with a verification statement, which probed participants' comprehension of the semantic relationship between one of the nouns and one of the verbs (e.g., *The nurse did the thanking. True/False? The nurse was thanked? True or False*). The order of presentation of the stimulus sentence verification statement pairs was randomized. We also created sentence materials for two practice trials.

Procedure. Each trial consisted of the following events. First, a fixation-cross appeared at the center of the display monitor for 1000 ms. Second, the stimulus sentence was presented one word at a time at the center of the display monitor, each new word replacing the previous one. The word presentation rate was 300 ms per word plus 20 ms for every letter in a word. Thus, a word's presentation duration increased linearly with the number of letters (Miyake et al., 1994a), approximating the effect of word length on eye fixations during reading (Just & Carpenter, 1992). Third, immediately following the last word of the stimulus sentence, the entire verification statement appeared one line lower with a prompt to press one button for "true" and another button for "false". Response-to-key assignment was counterbalanced across participants. The two response keys were '1' and '2'. There was a response deadline of 4 s. Both response accuracy and answer time (i.e., time from onset to verification statement to onset of response) were recorded. Participants did two practice trials followed by 64 experimental trials. After every 16 trials, there was a short one-minute break, during which participants were asked to rest their eyes and focus them at various distances. Participants initiated each next trial with a button press. The score was the percentage correct (out of 64 trials).

Results

Table 1 shows the descriptive statistics for conceptual span, reading span, word span, text comprehension, and sentence comprehension. Table 2 shows the product moment correlations among the span measures and between each of the span measures and each of the comprehension tests. Conceptual span,[2] reading span, and word span each showed moderate and significant correlations with sentence and text comprehension. Conceptual span is still significantly correlated with sentence and text comprehension when individual variation in word span was statistically controlled for ($r = .27$, $p < .05$ and .32, $p < .05$, respectively). When individual variation in conceptual span was statistically controlled for, word span no longer correlated significantly with text comprehension ($r = .19$, $p = .14$), albeit that it still predicted sentence comprehension ($r = .28$, $p < .05$). The magnitude of the correlation with text comprehension was somewhat greater for conceptual and reading span than for word span, while the magnitude of the correlation for

[2] A final analysis examined the correlations between each of the three span measures and sentence comprehension, separately for the semantically related ($M = 75\%$ correct) and unrelated condition ($M = 75\%$ correct) in the comprehension task. The correlation between conceptual span and comprehension accuracy was .33 ($p < .01$) in the related and .33 ($p < .01$) in the unrelated condition. The correlation between word span and comprehension accuracy was .33 ($p < .01$) in the related and .31 ($p < .05$) in the unrelated condition. The correlation between reading span and comprehension accuracy was .36 ($p < .01$) in the related and .20 ($p > .10$) in the unrelated condition. We did not necessarily expect a larger correlation between conceptual span and sentence comprehension in the unrelated than related condition. Weakly associated words may be more difficult to retain in semantic STM than strongly associated words (Haarmann & Usher, 2001). However, in on-line sentence processing, weakly associated words may also engage semantic STM to a lesser extent, due a need to use phonological STM to re-process difficult-to-integrate words.

Table 2
Correlations between measures in Study 1

Measure	Conceptual span	Reading span	Word span	Text comprehension	Sentence comprehension
Conceptual span	1.00	.37**	.32**	.37**	.35**
Reading span		1.00	.38**	.35**	.30**
Word span			1.00	.25*	.34*
Text Comprehension				1.00	.51**
Sentence Comprehension					1.00
Conceptual span (controlling word span)				.32*	.27*
Word span (controlling conceptual span)				.19	.28*

$^{*}p < .05$.
$^{**}p < .01$.

sentence comprehension was only slightly greater for conceptual and word span than for reading span.

To examine whether conceptual span accounts for unique variance of text and sentence comprehension that is not accounted for by reading span, we performed a multiple regression analysis of text and sentence comprehension onto reading span and conceptual span, entering reading span first and conceptual span second. In text comprehension, reading span accounted for 12% of the variance ($R^2 = 12$, $F(1,63) = 8.67$, $p < .01$). Of these 12%, 6% represented a unique[3] contribution by reading span ($p < .5$) and the remaining 6% were shared with conceptual span. Conceptual span accounted for another 7% of the variance ($R^2 = 7$, $F(1,62) = 5.28$, $p < .05$). Together, reading span and conceptual span accounted for 19% of the variance in text comprehension ($R^2 = 19$, $F(2,62) = 7.33$, $p < .01$). In sentence comprehension, reading span accounted for 9% of the variance ($R^2 = 9$, $F(1,61) = 5.92$, $p < .05$). Of these 9%, 3% represented a unique but non-significant contribution by reading span ($p = .17$) and the remaining 6% were shared with conceptual span. Conceptual span accounted for another 6% of the variance ($R^2 = 6$, $F(1,60) = 4.39$, $p < .05$). Together, reading span and conceptual span accounted for 15% of the variance in sentence comprehension ($R^2 = 15$, $F(2,60) = 5.31$, $p < .01$).

The conceptual span is analogous to Sperling's (1960) partial report task and as such allows us to estimate the capacity of semantic STM from the data. The estimated capacity was 3.4 items, which we obtained by multiplying the average number of items recalled per trial (1.81 item) with the number of categories that could be probed in a memory list (i.e., three categories), and by further multiplying in a conservative guessing correction (i.e., 1–3/8). This guessing correction seemed appropriate because three of eight items in a category were probed per trial and because subjects were acquainted with the pool from which words were sampled prior to the test.

Discussion

The multiple regression results demonstrated that conceptual span and reading span each accounted for a unique portion of the variance in text comprehension. This could indicate that conceptual span indexes semantic STM capacity, whereas reading span indexes a different ability, such as control of attention (Kane & Engle, 2000; Kane et al., 2001). The partial correlation results were also consistent with the hypothesis that conceptual span indexes semantic STM. Conceptual span still predicted sentence and text comprehension when individual differences in word span were statistically controlled for, possibly indicating that conceptual span places a greater emphasis on retention of lexical items in semantic STM compared to word span. By contrast, word span did no longer predict text comprehension when individual differences in conceptual span were statistically controlled for, but still predicted sentence comprehension. This could indicate that word span relies to a greater extent than conceptual span on phonological STM, which may be especially important for the processing of the kinds of difficult sentences that were included in the sentence comprehension task. The storage of a verbatim representation of the sentence in phonological STM may provide a back-up mechanism to support the re-processing of difficult sentences, when their immediate online meaning integration fails (Baddeley, 1986; Martin & Romani, 1994; Martin et al., 1994). That sentences were indeed difficult to process is suggested by the

[3] The unique variance contribution of a span measure was calculated as the square of its semi-partial correlation with the criterion measure.

overall error rate of 26% in the sentence comprehension task (see Table 1).[4]

Previous studies that used the VSAT as a measure of text comprehension (Daneman & Carpenter, 1980; LaPointe & Engle, 1990), found somewhat larger correlations between word span and text comprehension and between reading span and text comprehension than were obtained in Study 1. While we found a correlation of .25 ($n = 66$) between word span and text comprehension, Daneman and Carpenter (1980) reported non-significant correlations ranging from .35 to .37 ($n = 20$) and LaPointe and Engle (1990) reported significant correlations ranging from .37 to .49 ($n = 80$). Similarly, we found a correlation of .35 ($n = 66$) between reading span and text comprehension, whereas Daneman and Carpenter (1980) observed correlations ranging from .49 to .59 and LaPointe and Engle (1990) observed a correlation of .54. Our study included only two passages from the VSAT. By contrast, Daneman and Carpenter (1980) and LaPointe and Engle (1990) used VSAT scores based on the administration of the entire test, making it perhaps a more sensitive measure for capturing individual variation in text comprehension, and possibly explaining why they obtained larger correlations. To obtain similar-sized correlations, we used more difficult text materials in Study 2, which also attempts to contrast the semantic and phonological processes in STM.

The estimated average capacity of semantic STM (i.e., 3.4 items) resembles the storage capacity (i.e., three items), which we obtained previously in a category cued recall experiment in which memory lists with words in 6 different semantic categories were presented and one word was cued per trial (Haarmann & Usher, 2001, Experiment 2). This capacity is around 3–4 items, which has been argued by Cowan (2001) to be a better estimate of the STM capacity than the traditional 7 ± 2 items when the influence of strategies (such as rehearsal of items in the phonological loop or deep encoding of items in LTM via semantic elaboration) is prevented. The focus on semantic item information present in the conceptual span task makes rehearsal in the phonological loop less likely (e.g., we found that word length effects, which are attributed to rehearsal, are present in serial recall but not in semantic cued recall, Haarmann & Usher, 2001). Moreover, the repeated sampling of words from a small finite pool of words and categories makes it difficult for participants to reliably retrieve items from LTM, due to the likely build-up of pro-active interference across trials (see also Discussion of Study 3).

Study 2

Study 2 correlated each of four span measures (i.e., conceptual span, reading span, word span, and non-word span) with each of four comprehension measures (i.e., pronoun texts, GRE texts (Graduate Record Examinations), verbal problem solving, and semantic anomaly detection). One aim of Study 2 was to test the prediction that semantic STM capacity, as measured by the conceptual span task, is a better predictor of comprehension than phonological STM capacity, as measured by a non-word span task with pseudo-words. We also included a word span task, which we assumed to receive mixed contributions from semantic STM and phonological STM (cf. discussion above) and which we expected to predict comprehension better than non-word span but not as well as conceptual span. As a further way of addressing the same issue, we carried out multiple regression analyses to test the prediction that individual differences in semantic STM capacity, as measured by the conceptual span task, contribute unique variance to the prediction of comprehension, above and beyond comprehension variance explained by individual differences in phonological STM capacity, as measured by the non-word span task. Moreover, we carried out multiple regression analyses to test our prediction that this extra variance contribution is larger for conceptual span than for word span, reasoning that the semantic task emphasis of conceptual span would make it a better measure of semantic STM capacity than word span. Following Martin et al. (1994), we obtained a word-minus-non-word span measure by subtracting the non-word span score from the word span score. Martin et al. reported neuropsychological evidence that "*the difference could be attributed to the availability of lexical and semantic information to support the retention of the word lists*" (p. 89). Accordingly, we predicted that (1) conceptual span should correlate well with word span and with word-minus-non-word span (due to the common semantic STM component) and that (2) non-word span should correlate well with word span (due to the common phonological STM component) but not with conceptual span (due to the lack of a shared STM component). We also regressed conceptual span onto non-word and word span to provide an additional method for determining how much unique variance word span contributes to conceptual span above and beyond its phonological contribution (captured by non-word span).

A second aim of Study 2 was to check whether we could obtain higher correlations between reading span

[4] The overall error rate in our study was slightly higher than that in a previous sentence comprehension study by King and Just (1991) who obtained an overall error rate of about 22% (estimated from their Fig. 2) and used similar subject- and object-relative sentences. In our study, word presentation was experimenter-paced and there was a 4 s response deadline, whereas in the King and Just study word presentation was subject-paced and there was no response deadline. These procedural differences might explain why in our study the overall error rate was slightly higher.

and text comprehension in the range of those found in previous studies (from .49 to .59, Daneman & Carpenter, 1980; LaPointe & Engle, 1990) by using more difficult texts. To achieve this goal, we used two comprehension tasks, each of which was likely to be more difficult than the VSAT comprehension task in Study 1. The first comprehension test critically relied on the ability to relate a pronoun to an earlier antecedent (i.e., proper noun) in a text, where one or more sentences could intervene between the pronoun and its antecedent, and where several competing proper names had been introduced prior to the antecedent. Daneman and Carpenter (1980) devised such a text comprehension task in order to engage WM and found that it correlated highly (i.e., .90) with reading span. The second comprehension test was based on paragraphs and questions taken from the verbal test of the GRE.

A third aim of Study 2 was to replicate Study 1's finding that conceptual span predicts sentence processing performance with a different task. Whereas in Study 1 we used a sentence comprehension task, here we used an on-line semantic anomaly detection task. Each semantic anomaly involved a critical word pair (e.g. "*Did the visitor that admired the plants* <u>*write*</u> *the* <u>*wallet*</u> *in the room?*"), which could occur at variable locations in the sentence and whose words could be separated by intervening words. Moreover, the task required an immediate response (i.e., as soon as the anomaly was detected), discouraging the use of phonological STM as a back-up mechanism that requires a time-consuming re-processing of the sentence and, encouraging instead the use of semantic STM to support the on-line maintenance and comparison of word meanings. The semantic anomaly judgment task has been used extensively in previous neuropsychological research. A pronounced performance decrement on this task has been found in patients with a semantic STM deficit but not in patients with a phonological STM deficit. It occurred when the memory load was high and when the critical words could not be immediately integrated into the sentence context (Hanten & Martin, 2000; Martin & Romani, 1994). For these reasons, we expected that conceptual span would predict anomaly detection performance, but that non-word span, and possibly also word span, would be less good predictors. Furthermore, we expected that reading span like conceptual span would also predict on-line anomaly detection, because it has been found that reading span predicts the effect of semantic information on on-line word reading times during sentence comprehension (Just & Carpenter, 1992; Miyake et al., 1994a).

A final aim of Study 2 was to extend our investigation of the role of semantic STM beyond sentence and text comprehension by testing the prediction that conceptual span correlates highly with verbal problem solving ability. It is known that individual differences in WM predict the ability to solve verbal problems, such as syllogistic analogies (Gilhooly, Logie, Wetherick, & Wynn, 1993; Gilhooly, Logie, & Wynn, 1999) and class-inclusion problems (Howe, Rabinowitz, & Powell, 1998). WM enables verbal problem solving not only by supporting the comprehension of the sentences that constitute a problem but also by providing a computational workspace in which the solution to a problem can be generated and tested (Carpenter, Just, & Shell, 1990; Greeno, 1973). The latter involves comparing the abstract relations among the meaning elements of different propositions within and across sentences (Carpenter et al., 1990; Hummel & Holyoak, 1997) as well as keeping track of and managing the different problem-solving states, including the problem-solving goals and sub-goals (Carpenter et al., 1990; Ernst & Newell, 1969). Since we assume that semantic STM is the component within WM that helps to actively maintain such abstract meaning elements, we expected individual differences in the capacity of semantic STM to predict verbal problem solving ability.

Method

Participants. Sixty undergraduate students from University of London, all native speakers of English, participated in the study. They were paid five pounds for their participation.

Tests

1. *Conceptual span*. The conceptual span test was the same as the one used in Study 1 except that it included 10 instead of 16 test trials.

2. *Reading span*. The reading span test was the same as the one used in Study 1 with two differences. First, there were five instead of two trials at each set size. As soon as a participant responded with three trials correct at a given set size, the set size was increased with one sentence. When a participant got less than three of five trials correct at a given set size, testing was discontinued. Second, the reading span score was equal to the largest set size at which a participant got three of five trials correct. Half a point was added to the score, if a participant got two out of five trials correct at the next largest set size. Both in terms of the number of trials per length and scoring method, the reading span test in Study 2 was identical to one used by Daneman and Carpenter (1980).

3. *Word span*. The word span test was the same as the one used in Study 1. While the word span score was defined as the number of correct trials in Study 1, it was defined as the largest set size for which a correct response was obtained in Study 2.

4. *Non-word span*. The materials consisted of pseudo words that were constructed by rearranging the letters from the words in the word span test (e.g., dag, lund).

As a result, the non-word stimuli had the same phonological, orthographic, and visual features as the word stimuli in the word span test.

5. *Pronoun texts*. This test consisted of four narrative passages similar in structure to the ones used in Daneman and Carpenter's, 1980 text comprehension experiment. Each passage was approximately 140 words long and consisted of 12 sentences. Each passage introduced several different persons by name and its final sentence contained a pronoun whose antecedent was the last named person. The number of sentences intervening between this pronoun and its antecedent was 1, 2, 3, and 4 in the first, second, third, and fourth passage, respectively. Participants silently read the sentences of a passage one at time and pressed a key to display the next sentence on a computer display. Following the last sentence in the passage, they were presented with three open-ended questions, which they had to answer out loud. The first question probed whether a participant knew the correct antecedent noun of the pronoun mentioned in the last sentence. The two remaining questions were about explicitly stated facts in the passage. The score for this test was the number of correctly answered questions. Its maximum possible value was 12.

6. *GRE texts*. This test consisted of five passages that were taken from the verbal reasoning part of the GRE test. The first passage was 620 words long, and the other four passages were each approximately 250 words long. After each passage, there were four multiple-choice questions about facts that were either stated or implied in the text. Each question had four choices and participants had to circle the correct answer. Each passage was typed on a sheet of paper and the questions were always typed on a separate page following it. Based on pilot data, participants were given a maximum of 20 min to silently read all passages and answer the questions following it. Participants were allowed to re-examine the text when answering the questions. The score for this test was the number of correct answers. Unanswered questions were counted as wrong answers. The maximum possible score was 20.

7. *Verbal problem solving*. This test consisted of 20 multiple-choice questions about verbal problems. There were three types of problems (see Appendix A for an example of each). Two problem descriptions were taken from a graduate record examinations (GRE) test. They were each accompanied by three multiple-choice questions with five answer alternatives (problem type 1). Two further problems were taken from a book by Barrett and Williams (1990) and consisted of a statement and four facts. Participants were asked to choose which two of the four facts were necessary to make the statement true (problem type 2). The remaining eight problems were also taken from the book by Barrett and Williams (1990) and consisted of several statements regarding the ordinal relations among a set of entities (e.g., locations of persons seated in a row) (problem type 3). These problems were followed by one or two multiple-choice questions with three, four, or five answer alternatives. Participants were shown one example each of problem types 2 and 3 prior to the test. All problems and questions were typed on paper. Based on pilot data, participants were given 10 min to read the entire set of problems and circle the correct answer alternatives. Only one of the answer alternatives of a question was correct. Unanswered questions were counted as wrong answers. The maximum possible score was 20.

8. *Anomaly judgment. Materials and design*. The materials for this test consisted of 44 question sentences of a variety of types (e.g., *Did the reporter that conducted the interview hear the secret in the room? In what class did she never get any questions from the pupils?*). Half of the sentences were semantically sensible and half were semantically anomalous. For every semantically sensible sentence of a particular length (i.e., number of words) and syntactic structure there was a semantically anomalous sentence of the same length and syntactic structure (e.g., Sensible: *Did the cashier that collected the receipts place the papers on the register?* versus Anomalous: *Did the visitor that admired the plants write* the *wallet in the room?*). A semantic anomaly was created by replacing a word in a sensible sentence (e.g., *Did the visitor that admired the plants leave the wallet in the room* changed into *Did the visitor that admired the plants write the wallet in the room?*). To encourage participants to analyze and store the meaning of all parts of a sentence, anomalies could occur at different locations within the sentences (e.g., inside the main clause *Did the visitor that admired the plants write the wallet in the room?*, inside the relative clause *Did the soldier that wore the grill send the report to the barracks*, or spanning the main and relative clause *Was he welcomed to the reception that was rung in his father's honor?*) and with a variable number of intervening words separating the two critical words (e.g., one intervening word *Did the visitor that admired the plants write the wallet in the room?* [15 of 20 anomalous trials] versus an intervening relative clause *What wallet did the visitor that admired the plants write in the room?* [5 of 20 anomalous trials]).

Procedure. Each participant encountered a particular sentence only once and all sentences differed in lexical content. Each participant received the trials in a different random order. Prior to the test, participants received examples of semantically anomalous and semantically sensible sentences. On each trial during the test, participants silently read the sentence, which was presented one word at a time in the center of the screen at a rate of 2.3 words/s. Participants were instructed to press a key as soon as the sentence stopped making sense, and to do nothing if it was sensible. As soon as a key was pressed, the presentation of the sentence was aborted and the next trial began. The computer recorded both response accuracy and latency. The score on the test was the percentage correct.

Results

Table 3 shows the descriptive statistics for conceptual span, reading span, word span, non-word span, pronoun texts, GRE texts, verbal problem solving, and semantic anomaly judgment. Table 4 shows the product moment correlations among the span measures themselves and among the span measures and each of the comprehension tests. The average storage capacity in the conceptual span task was estimated at 3.1 items (see Study 1 for the calculation method and for discussion).

Correlations among span measures. Reading span correlated with both conceptual span ($r = .47$, $p < .01$) and word span ($r = .56$, $p < .01$) but not with non-word span ($r = .07$, $p = .57$). The size of the correlation between non-word and conceptual span ($r = .27$, $p < .5$) was significantly smaller ($t(57) = 2.34$, $p < .5$) than the size of the correlation between word span and conceptual span ($r = .52$, $p < .01$)[5] and it was also significantly smaller ($t(57) = 3.11$, $p < .01$) than the size of the correlation between non-word and word span ($r = .58$, $p < .01$). The difference span measure, word-minus-non-word span, correlated highly with conceptual span ($r = .40$, $p < .01$) and approached the size of the correlation between word and conceptual span ($r = .52$, $p < .01$) (difference between correlations, $t(57) = 1.46$, $p < .10$). To determine more directly how much word span contributes to predicting conceptual span above and beyond its phonological contribution, we regressed conceptual span onto non-word span (entered-first) and word span (entered-second). Non-word span accounted for 7% of the variance [$R^2 = 7$, $F(1,58) = 4.74$, $p < .5$], which were entirely shared with word span. Word span added another 20% of variance [$R^2 = 20$, $F(1,57) = 15.20$, $p < .001$]. Together, non-word span and word span accounted for 27% of the variance [$R^2 = 27$, $F(2,57) = 10.38$, $p < .001$].

Correlations among span and comprehension measures. We report the correlation and multiple regression results for each of the four comprehension measures separately. The correlation results indicate how well each span measure predicts the comprehension measures. The multiple regression analyses determine whether conceptual span adds any variance in predicting comprehension above and beyond variance predicted by either non-word span, or word span, or reading span (see Table 5). In these analyses, non-word span, or word span, or reading span was entered first and conceptual span was entered second. Furthermore, to compare how much variance word span versus conceptual span contribute to predicting variance in comprehension above and beyond variance contributed by non-word span, we not only regressed each of the comprehension measures onto non-word span and conceptual span but also onto non-word span and word span.

Verbal problem solving. Verbal problem solving was predicted by conceptual span ($r = .51$, $p < .01$) and word span ($r = .29$, $p < .5$). The correlation with verbal problem solving was marginal for reading span ($r = .22$, $p < .10$) and non-significant for non-word span ($r = .20$, $p = .12$). In terms of size, the correlation between conceptual span and verbal problem solving was larger than the correlation between word span and verbal problem solving ($t(57) = 1.94$, $p = .05$) and the correlation between reading span and verbal problem solving ($t(57) = 2.43$, $p < .01$). It also was larger than the correlation between non-word span and verbal problem solving ($t(57) = 2.18$, $p < .03$). Moreover, stepwise-multiple regression analyses in which non-word span, word span, or reading span was entered first and conceptual span entered second, showed that conceptual span explained variance in verbal problem solving above and beyond variance explained by any of the other span measures (see Table 5, row 1). Conceptual span explained 21%, 18% and 22% unique variance in verbal problem solving above and beyond reading span, word span, and non-word span, respectively.

Anomaly judgment. Anomaly judgment was predicted by conceptual span ($r = .42$, $p < .01$), word span ($r = .35$, $p < .01$), reading span ($r = .32$, $p < .5$), and non-word span ($r = .28$, $p < .5$). We also correlated each of the span measures with the latencies for correct trials in the anomaly judgment task and found no significant correlations. This finding rules out a speed-accuracy trade-off explanation of the correlation between the span measures and anomaly judgment accuracy. Statistically the correlations of each of the span measures with accuracy of anomaly judgment did not differ in size (all $p > .15$). However, stepwise-multiple regression analyses, in which non-word span, word span, or reading span was entered first and conceptual span entered second, showed that conceptual span explained variance in anomaly judgment over and beyond variance explained by any of the other span measures (see Table 5, row 2). To determine the unique variance contribution of conceptual span above and beyond non-word and word span, we extended the regression analysis reported in Table 5 and entered conceptual span on a third step after non-word span (entered first) and word span (entered second). Conceptual span accounted for 7% of unique variance in anomaly judgment above and beyond

[5] The test for the significance of the difference between dependent correlations (i.e., correlations obtained from the same sample) was computed with the *diffdef.exe* program, accompanying an article by Crawford, Mychalkiw, Johnson, and Moore (1996). This program implements Howell's (1997) procedures for such a test.

Table 3
Descriptive statistics: Study 2

Measure	*N*	*M*	*SD*	Min	Max	Maximum possible score
Conceptual span	60	19.17	3.63	12	27	30
Reading span	60	3.93	.84	2	5	5
Word span	60	4.92	.85	3	7	9
Non-word span	60	3.37	.58	2	5	9
Pronoun texts	60	8.05	2.11	2	12	12
GRE texts	60	12.92	4.33	3	19	20
Anomaly judgment (percent correct)	60	68	14.8	25	95	100
Verbal problem solving	60	14	3.29	5	19	20

Table 4
Correlations of span measures with each other and with the comprehension measures: Study 2

Measure	Non-word span	Conceptual span	Reading span	Pronoun texts	GRE texts	Anomaly judgment	Verbal problem solving
Word span	.58**	.52**	.56**	.36**	.40**	.35**	.29*
Non-word span	—	.27*	.07	.14	.12	.28*	.20
Conceptual span	—	—	.47**	.39**	.34**	.42**	.51**
Reading span	—	—	—	.56**	.38**	.32*	.22

*$p < .05$.
**$p < .01$.

non-word and word span [$R^2 = 7$, $F(2, 56) = 4.76$, $p < .05$]. Together, non-word span, word span, and conceptual span accounted for 24% variance ($R^2 = 24$, $F(3, 56) = 5.47$, $p < .01$).

Pronoun texts. Comprehension of pronoun texts was predicted by reading span ($r = .51$, $p < .01$), conceptual span ($r = .39$, $p < .01$), and word span ($r = .36$, $p < .01$), but not by non-word span ($r = .14$, $p = .30$). In terms of size, the correlations between conceptual span and pronoun texts and between word span and pronoun texts were marginally larger than the correlation between non-word span and pronoun texts ($t(57) = 1.67$, $p = .10$, and $t(57) = 1.94$, $p < .10$, respectively). The correlation between reading span and pronoun texts was not significantly larger than the correlation between conceptual span and pronoun texts ($t(57) = 1.48$, $p = .14$), whereas it was marginally larger than the correlation between word span and pronoun texts ($t(57) = 1.91$, $p < .10$). Moreover, stepwise-multiple regression analyses, in which non-word span, word span, or reading span was entered first and conceptual span entered second, showed that conceptual span explained variance in comprehension of pronoun texts over and beyond variance explained by non-word and word span (see Table 5, row 3). Conceptual span and word span predicted similar amounts of unique variance (i.e., 14% and 12%, respectively) in the comprehension of pronoun texts above and beyond variance predicted by non-word span.

GRE texts. Comprehension of GRE texts was predicted by word span ($r = .40$, $p < .01$), reading span ($r = .38$, $p < .01$), and conceptual span ($r = .34$, $p < .01$) but not by non-word span ($r = .12$, $p = .30$). In terms of size, the correlations between conceptual span and GRE texts and between word span and GRE texts were marginally ($t(57) = 1.45$, $p < .10$) and significantly ($t(57) = 2.53$, $p < .01$) larger than the correlation between non-word span and GRE texts and, respectively. By contrast, the correlations between word span, reading span, and conceptual span and GRE texts did not differ significantly in size (all *p*s $> .30$). Moreover, stepwise-multiple regression analyses, in which non-word span, word span, or reading span was entered first and conceptual span entered second (see Table 5, row 4), showed that conceptual span explained variance in comprehension of GRE texts over and beyond variance explained by non-word span. Conceptual span and word span accounted for 10 and 16% unique variance, respectively, in the comprehension of GRE texts above and beyond the variance predicted by non-word span (see Table 5, row 4).

Discussion

It has been previously suggested that non-word span indexes phonological STM capacity and that word span receives mixed contributions from phonological and semantic STM (Haarmann & Usher, 2001; Martin &

Table 5
Percent variance accounted for in stepwise multiple regressions of each of the four comprehension measures onto different pairs of span tests in Study 2

	First[a] Second[b]	Rspan[c] Cspan[f]	Wspan[d] Cspan	Nwspan[e] Cspan	Nwspan Wspan
Verbal problem solving	First total[g]	5^{*}	8^{**}	4^{ns}	4^{ns}
	First unique[h]	0	0	0	0
	First shared[i]	5	8	4	4
	Second unique[j]	21^{****}	18^{****}	22^{****}	4^{ns}
	Overall[k]	26^{****}	26^{****}	26^{****}	8^{****}
Anomaly judgment	First total	10^{**}	12^{**}	8^{**}	8^{**}
	First unique	2^{ns}	2^{ns}	3^{ns}	1^{ns}
	First shared	8	10	5	7
	Second unique	9^{**}	8^{**}	12^{***}	5^{*}
	Overall	19^{***}	20^{***}	20^{***}	13^{**}
Pronoun texts	First total	32^{****}	13^{***}	2^{ns}	2^{ns}
	First unique	19^{***}	4^{ns}	0	0
	First shared	13	9	2	2
	Second unique	2	6^{**}	14^{***}	12^{***}
	Overall	34^{****}	19^{***}	16^{***}	14^{**}
GRE texts	First total	14^{***}	16^{***}	2^{ns}	2^{ns}
	First unique	6^{**}	7^{**}	0	0
	First shared	8	9	2	2
	Second unique	4^{*}	3^{*}	10^{**}	16^{***}
	Overall	18^{***}	19^{**}	12^{**}	18^{***}

Note. Table values represent percentages. ns, non-significant, $p > .10$.

[a] First, span test entered on first step.

[b] Second, span test entered on second step.

[c] Rspan, reading span.

[d] Wspan, word span.

[e] Nwspan, non-word span.

[f] Cspan, conceptual span.

[g] First total, variance in comprehension measure accounted for by the first predictor variable (given by the square of its R).

[h] First unique, unique variance in comprehension measure accounted for by the first predictor variable (i.e., square of the semi-partial correlation of the first span test with the comprehension measure, controlling for the second span test).

[i] First shared, variance in comprehension measure shared by the first and second span test (calculated by subtracting first unique from first total).

[j] Second unique, unique variance in comprehension measure accounted for by the second span test only (given by the square of its R change).

[k] Overall, overall variance in comprehension measure accounted for by the first and second span test (given by R^2 of the model including the first and second span test). In order, the degrees of freedom of the within- and between sums of squares of the F values of R were 1 and 58 for the span test that was entered first, 1 and 57 for the span test that was entered second, and 2 and 57 for the overall regression model.

$^{*}p < .10$.

$^{**}p < .05$.

$^{***}p < .01$.

$^{****}p < .001$.

Romani, 1994; Martin et al., 1994). Moreover, we reasoned that conceptual span receives a minor contribution from phonological STM and major contribution from semantic STM, due to its emphasis on the retrieval of lexical-semantic item information. Consistent with this interpretation, we found higher correlations between non-word and word span (likely to reflect phonological STM) and between word span and conceptual span (likely to reflect semantic STM), than between non-word and conceptual span. In addition, word span added a substantial unique variance contribution in predicting conceptual span (likely to reflect semantic STM) above and beyond the smaller variance contribution that it shared with non-word span (likely to reflect phonological STM). Finally, the word-minus-non-word span measure, which has been proposed as an index of semantic STM capacity (Martin et al., 1994), also predicted conceptual span. Thus, our findings with normal participants are consistent with results from brain-damaged patients, which suggest that phonological and semantic STM are separate components in verbal STM (Hanten & Martin, 2000; Martin & Freedman, 2001; Martin & Romani, 1994; Martin et al., 1994; Romani & Martin, 1999).

The results of Study 2 furthermore confirmed our prediction that semantic STM capacity, as measured by the conceptual span task, is a better predictor of comprehension than phonological STM capacity, as measured by the non-word span task. Conceptual span but not non-word span predicted comprehension of GRE texts, pronoun texts, and verbal problem solving.[6] Conceptual span predicted comprehension of anomaly judgment better than non-word span did, explaining 12% of unique variance in addition to the 5% of variance it shared with non-word span. This finding suggests that semantic STM plays a larger role in on-line anomaly judgment than does phonological STM. Furthermore, the correlation between non-word span and pronoun texts, GRE texts, and verbal problem solving was low and non-significant, consistent with the neuropsychological evidence that phonological STM is not critical for comprehension. Our finding that a likely index of semantic STM capacity, conceptual span, predicts text comprehension can be reconciled with Romani and Martin's (1999) finding that their semantic STM patient, A.B., performed at normal levels on a story comprehension task. In their stories, words were easy to integrate into each sentence (Romani & Martin, 1999), minimizing the need for semantic STM to store unintegrated word meanings to enable their on-line semantic integration (Romani & Martin, 1999). By contrast, the storage of unintegrated word meanings may have been important to comprehend the more difficult texts in our study, for example, to integrate a pronoun with its antecedent (Daneman & Carpenter, 1980) or to support inference generation (St. George, Mannes, & Hoffman, 1997), either while participants were reading the texts or answering the questions.

Conceptual span may provide a more sensitive index of semantic STM capacity than word span due to the explicit semantic task emphasis that is present in conceptual span (i.e., recall the words in a cued semantic category). Consistent with this reasoning, we found that conceptual span was a better predictor of verbal problem solving than word span. In addition, conceptual span explained unique variance above and beyond variance accounted for by word span in anomaly judgment, comprehension of pronoun texts, and verbal problem solving. Two results indicate that this unique variance contribution of conceptual span is likely to reflect semantic STM and not phonological STM. First, non-word span, which we believe to index phonological STM, did not explain any variance in the comprehension of pronoun texts and verbal problem solving. Second, conceptual span explained unique variance in anomaly judgment in a 3-factor regression model in which non-word span, word span, and conceptual span were added on the first, second, and third step, respectively. The only comprehension measure that was predicted equally well by conceptual span and word span was GRE texts. The correlations between GRE texts and conceptual span and between GRE texts and word span did not differ significantly in size and conceptual span did not add unique variance above and beyond variance contributed by word span.

As in previous studies, we obtained a high correlation between a storage-and -processing measure, reading span, and a measure of text comprehension (pronoun texts in our study and VSAT in Daneman & Carpenter, 1980, and LaPointe & Engle, 1990). Consistent with the idea that a storage-plus-processing measure predicts comprehension better than a storage-only measure (Daneman & Carpenter, 1980; Daneman & Merikle, 1996), we found that reading span correlated descriptively more highly with pronoun texts than did word span and conceptual span. However, when statistically tested, this difference in correlation sizes was only marginally significant in the case of word span and failed to reach significance in the case of conceptual span. Moreover, the idea that a storage-plus-processing measure predicts comprehension better than a storage-only measure did not hold in general. First, comprehension of GRE texts correlated more highly with word span than with conceptual span than with reading span, but these differences in the size of the correlation were not significant. Second, anomaly judgment correlated more highly with conceptual span than with word span than with reading span, but also these differences in the size of the correlation were not significant. Third, conceptual span was a better predictor of verbal problem solving than reading span or word span, in spite of the fact that conceptual span is a storage-only measure and reading span a storage-plus-processing measure. A final result we wish to emphasize was that conceptual span accounted for unique variance in anomaly judgment and verbal problem solving above and beyond variance accounted for by reading span. Our results could indicate that conceptual span and reading span measure different abilities. Conceptual span may provide a more sensitive index of the capacity of semantic STM than reading

[6] Performance on the non-word span task showed less variability than performance on any of the other span measures (see Table 3). However, it is unlikely that this lack of variability or unreliability explain why non-word span did not correlate with GRE text, pronoun texts, and verbal problem solving. The reason is that non-word span, in spite of its smaller variability, showed a high and statistically significant correlation with word span (.58, $p < .01$). This correlation is likely to reflect the contribution of phonological STM (see text).

span. Reading span, on the other hand, may index the ability to control attention, including the ability to resist pro-active interference in retrieval from LTM (Kane & Engle, 2000). Whether, and to what extent, these latter abilities also contribute to conceptual span remains to be determined. The ability to control attention to resist proactive interference in LTM may be especially important to comprehend pronoun texts where linking the pronoun to its antecedent proper noun several sentences earlier may involve retrieval from LTM (Kintsch, 1998), which is likely to be interfered with by similar proper nouns occurring in the same text. Another possibility is that in the reading span task (and in text comprehension) there is a greater need for domain-specific linguistic skills (e.g., syntactic parsing) than in the conceptual span task (and in verbal problem solving). This might explain why reading span predicted text comprehension better than verbal problem solving, whereas the opposite was the case for conceptual span.

Study 3

Our interpretation of conceptual span is that it indexes the capacity of semantic STM. However, in the conceptual span task, the words are presented in an order that mixes different categories. Therefore, an alternative interpretation of conceptual span is possible, namely, that the test indexes how well and how quickly participants mentally organize a mixed list into its semantic categories. Whereas clustering ability may be in part determined by the capacity of semantic STM, it may also be determined by the effectiveness of other processes, such as, the efficiency by which a subject matches a noun's meaning to its semantic category. Nevertheless, we hypothesized that conceptual span provides primarily a measure of semantic STM and not of clustering ability (i.e., aspects of that ability that are unrelated to semantic STM). Accordingly, we predicted that conceptual span correlates substantially with comprehension, not only when conceptual span uses mixed lists but also when it uses clustered lists. In clustered lists, words are grouped by their semantic category so that clustering ability should play no or only a minimal role. Study 3 aimed to test this prediction, contrasting a clustered and non-clustered version of the conceptual span task. A text comprehension task and an online semantic anomaly judgment task served as the comprehension measures.

A second aim of Study 3 was to follow-up on the semantic anomaly judgment results of Study 2. The semantic anomaly judgment task in Study 2 did not manipulate distance, that is, the number of words intervening between the two words that caused the anomaly. In fact, those two words were adjacent in most cases. We hypothesized normal participants to have more difficulties with the on-line detection of a semantic anomaly when the distance is long than when it is short, especially when their semantic STM capacity is low. The semantic anomaly task in Study 3 was designed to test this prediction. The critical anomalies involved an adjective and noun, which could be either separated by one adjective or by three adjectives. The semantic anomaly task was inspired by the results of an experiment carried out by Martin and Romani (1994) with semantic and phonological STM patients and matched normal controls. As reviewed above, they found that the presence of a semantic STM deficit was predictive of a pronounced distance effect in semantic anomaly judgment (more errors with increased distance between an adjective and noun or a noun and a verb) when the to-be-retained words could not be immediately integrated into the sentence context. Hanten and Martin (2000) reported the same findings for head-injured children with semantic and phonological STM deficits, suggesting that semantic STM stores unintegrated word meanings to support their on-line semantic integration during sentence processing (Hanten & Martin, 2000; Martin & Romani, 1994).

A third and final aim of Study 3 was to investigate whether we could replicate previous findings on the correlation of primacy and recency components of cued recall span measures with complex cognitive tasks. Cantor, Engle, and Hamilton (1991) measured the probed recall of nine items (either all words or all digits) and performance on the VSAT and found that performance for the last three of the nine items (recency component) correlated with VSAT, whereas performance for the first three items (primacy component) did not. The correlation for item triplet at recency and VSAT was .32 for words and .30 for digits, respectively, whereas the correlation for the item triplet at primacy and VSAT was −.07 for words and .11 for digits, respectively. In a series of experiments, Cohen and Sandberg (1977) measured the correlation between performance on a 9-digit probed serial recall task and an intelligent test in children. They found that the recency triplet correlated well (lowest $r = .45$) with the intelligence test, whereas the primacy triplet did not (highest $r = .24$ not significant). Likewise, we expected to find a correlation between performance on the third, but not first cluster in the clustered conceptual span task our two criterion measures, text comprehension and anomaly judgment. Such a finding is theoretically important because it would help to corroborate our claim that the source of the predictive ability of the clustered conceptual span task lies in individual differences in the capacity of semantic STM (reflected best by cued recall performance at recency) and not in individual differences in the capacity to retrieve items from episodic LTM (reflected best by cued recall performance at primacy). Such a differentiation is important from the perspective of a dual-store account of retrieval in immediate recall, which postulates that participants may try to retrieve items from STM and episodic LTM (Atkinson & Shiffrin, 1968; Baddeley, 1970; Craik & Levy, 1970; Glanzer, 1972; Haarmann & Usher, 2001; Levy & Baddeley, 1971; Nairne, Neath, & Serra, 1997).

Method

Participants. Sixty-four participants, forty-nine participants from the University of Maryland and fifteen from the University of London, all native speakers of English, participated in the study. The American participants received extra-credit or 7 dollars for their participation. The British participants received five pounds for their participation.

Tests

Each participant was tested individually and performed four tests, given in the same order, namely, non-clustered conceptual span, clustered conceptual span, text comprehension, and semantic anomaly judgment. The non-clustered conceptual span test was administered before the clustered conceptual span to prevent practice with clustering words into their semantic categories from affecting task performance on the non-clustered span test. A test session lasted about 45 min. Presentation of all tests was visual and computer controlled.

1. *Non-clustered conceptual span.* This test was the same as the conceptual span test used in Study 1.

2. *Clustered conceptual span.* The word pool and procedure for this test were the same as the conceptual span test used in Study 1 with two exceptions. First, words were *clustered* by semantic category. Second, instead of three words per category there were four words per category, to prevent ceiling level performance, which may result when small clusters of mutually supportive, adjacent words are stored in semantic STM (cf. Haarmann & Usher, 2001, Experiment 1). The following is an example of a trial in the clustered conceptual span test: *monkey*, *sheep*, *cow*, *horse*, *fan*, *video*, *phone*, *fax*, *banana*, *grape*, *lime*, *orange*, *ELECTRIC?* Correct answer: *fax*, *phone*, *fan*, *video* (cued recall in any order). The first, second, and third category in the memory list were probed on 5, 6, and 5 trials, respectively, for a total of 16 trials. The position of the probed category in the memory list (first, second, or third category) was randomized across trials. As was the case for the non-clustered version of the conceptual span test, there were two practice trials.

3. *Pronoun texts.* This test was the same as the pronoun texts test used in Study 2.

4. *Anomaly judgment. Materials and design.* The materials of this test consisted of 68 sentences of a variety of types. Half of the sentences were semantically sensible and half were semantically anomalous. Of the 68 sentences, 28 sentences were experimental sentences and 40 were filler sentences. The experimental sentences comprised a 2 × 2 within-subject design, crossing the factors distance (short, long) and anomaly (sensible, anomalous) and including seven trials per condition. All sentences included adjective–noun combinations. In the experimental sentences, an adjective–noun combination could be either sensible (e.g., "*The man liked the curly, brown hair of the woman in the car*") or anomalous (e.g., "*The boys admired the curly, new car of the secretary in the office*"). The distance between the adjective and the noun could be either short (one intervening adjective, e.g., the curly new car) or long three[7] intervening adjectives (e.g., *He was concerned about the heavy long steep narrow footpath strewn with rocks*). Assignment of adjectives and nouns to distance condition was random. There were also 40 filler sentences, half of which were sensible and half of which were anomalous. In some of the anomalous filler sentences, the anomaly involved the verb and the noun in the middle of the sentence (e.g., *He lifted the bright sun outside the factory.*), in some it involved the verb and sentence-final noun (e.g., *The owner was drying the wet clothes after diving into the house.*), and in some it involved an adjective–noun, as in the experimental sentences, but not the first adjective (e.g., *They didn't like the fat mean old milk that the man had bought*).

Procedure. The order of trials was randomized and each participant received them in the same order. On each trial during the test, participants silently read the sentence, which was presented one word at a time in the center of the screen at a rate of 450 ms per word (plus 30 ms for every letter in a word). Participants were instructed to press a no-key as soon as the sentence stopped making sense, and to press a yes-key if at the end of the sentence it turned out the sentence was sensible. The left and right fingers of the right hand were used to press the 1- and 3-key on the keypad, respectively. Finger-to-button assignment was counterbalanced across participants, with half of the participants using the 1-key for a sensible response and the 3-key for an absurd response and half of the participants using the reverse key assignment. There was a 1.5 s response deadline in order to prevent ceiling level performance. The computer recorded both response accuracy and latency. Prior to the test, participants received 10 practice trials.

Results

Table 6 shows the descriptive statistics for the span measures, that is, non-clustered conceptual span,

[7] We included one more adjective in our long distance condition than Martin and Romani (1994) to prevent ceiling level performance, because our participants were college-age adults, whereas their participants were matched in age to the older patients and thus likely to make more errors.

Table 6
Descriptive statistics: Study 3

Measure	*N*	*M*	*SD*	Min	Max	Maximum possible score
Non-clustered Conceptual span	64	30	5.31	19	40	48[a]
Clustered conceptual span	64	47	7.93	27	60	64[b]
Cluster 1 of conceptual span	64	13.9	2.90	7	20	20
Cluster 2 of conceptual span	64	15.8	4.00	6	21	24[c]
Cluster 3 of conceptual span	64	17.4	2.78	7	20	20
Text comprehension	64	7.61	2.28	3	12	12
Adjective–noun anomalies						
Across both distances	64	81	16	27	100	100
Short distance	64	88	12	44	100	100
Long distance	64	73	25	22	100	100

[a] Total number correct across16 trials with three words in each of three categories.

[b] Total number correct across 16 trials with four words in each of four categories.

[c] Cluster 2 was probed in six trials, whereas clusters 1 and 3 were each probed in five trials, explaining the different maximum possible scores.

Table 7
Correlations between span and comprehension measures in Study 3

Measure	Pronoun texts	Anomaly across distance	Anomaly short distance	Anomaly long distance
Non-clustered[a] conceptual span	.42**	.51**	.30*	.45**
Clustered conceptual span	.39**	.31*	.08	.34**
Cluster 1[b,c]	.22	.19	.03	.24
Cluster 2	.37**	.28*	.12	.29*
Cluster 3	.38*	.33*	.20	.28*

[a] The correlation between non-clustered and clustered conceptual span was .67**.

[b] In order, clusters 1, 2, and 3 refer to the category that is presented in the first, second, and third position of the memory lists in the clustered conceptual span task.

[c] In order, the correlations between clusters 1 and 2, 1 and 3, and 2 and 3 were .60 ($p < .01$), .22 ($p < .10$), and .59 ($p < .001$).

* $p < .05$.

** $p < .01$.

clustered conceptual span (total across the three clusters, and for the first, second, and third cluster separately) and for the complex language processing tasks, that is, text comprehension, and judgment of adjective–noun anomalies (overall, and for short and long distance condition). Table 7 shows the product moment correlations between the span measures and between the span measures and each of the comprehension tests. The average storage capacity was 3.5 items for non-clustered conceptual span (see text of Study 1 for the calculation method and for discussion). It was about 1 item higher ($t(64) = 23.1$, $p < .001$), that is, 4.4 items for clustered conceptual span (where a guessing factor of 4/8 instead of 3/8 was used), which is likely to reflect the mutual support among a cluster of same-category items in semantic STM.

Text comprehension. The correlation between non-clustered conceptual span and text comprehension was .42 ($p < .01$), the correlation between clustered conceptual span and text comprehension was .39 (p<.01), and the correlation between non-clustered and clustered conceptual span was .67 ($p < .01$). The correlation between non-clustered conceptual span and text comprehension (.42) and between clustered conceptual span and text comprehension (.39) did not differ in size ($t(61) = .32, p = .75$, see earlier Footnote 4). The correlation between non-clustered conceptual span and comprehension of pronoun texts in Study 2 ($r = .39$, $p < .01$) and in Study 3 (.42, $p < .01$) did not differ in size either ($z = .26, p = .80$).[8] In a stepwise multiple regression of text comprehension onto clustered conceptual span (entered first) and non-clustered conceptual span (entered second), clustered conceptual span accounted for 15% of the variance ($R^2 = 15$,

[8] The test for the significance of the difference between independent correlations (i.e., comparing correlations obtained from two samples) was computed with the *indepcor.exe* program, accompanying an article by Crawford et al. (1996). This program implements Howell's (1997) procedures for such a test following which both correlations are first converted to Fisher's *z*' and the difference between them divided by the standard error of the difference to yield a normal curve deviate (*z*).

$F(1,62) = 10.77, p < .01$). Of these 15%, a non-significant 2% represented a unique contribution by clustered conceptual span ($p = .22$) and the remaining 13% were shared with non-clustered conceptual span. Non-clustered conceptual span accounted for another 5% of the variance, albeit at a marginally significant level ($R^2 = 5$, $F(1,61) = 3.47$, $p < .10$). Together, clustered and non-clustered conceptual span accounted for 20% of the variance in text comprehension ($R^2 = 20$, $F(2,61) = 7.33$, $p < .01$).

Anomaly judgment accuracy. The accuracy of judging the sensibility/anomaly of adjective–noun combinations was analyzed.[9] Both clustered and non-clustered conceptual span correlated with anomaly judgment, but this correlation was significantly better ($t(61) = 2.22$, $p < .5$) for non-clustered conceptual span ($r = .51$, $p < .01$) than for clustered conceptual span ($r = .31$, $p < .5$). In a stepwise multiple regression of anomaly judgment onto clustered conceptual span (entered first) and non-clustered conceptual span (entered second), clustered conceptual span accounted for 10% of the variance ($R^2 = 10$, $F(1,62) = 6.21$, $p < .5$), which were entirely shared with non-clustered conceptual span. Non-clustered conceptual span accounted for another 16% of the variance ($R^2 = 16$, $F(1,61) = 12.03$, $p < .01$). Together, clustered and non-clustered conceptual span accounted for 26% of the variance in anomaly judgment ($R^2 = 26$, $F(2,61) = 9.7$, $p < .01$).

Participants were less sensitive in judging adjective–noun combinations in the long (73% correct) than short distance adjective–noun condition (88% correct) ($F(1,63) = 24.0$, $p < .001$, $MSe = .029$). Participants also took more time to accurately judge adjective–noun combinations in the long (852 ms) than short distance adjective–noun condition (800 ms) ($F(1,63) = 8.96$, $p < .01$, $MSe = 15681$). Clustered conceptual span correlated better ($p < .5$) with sensitivity in detecting adjective–noun anomalies in the long ($r = .34$, $p < .5$, one-tailed) than short distance condition ($r = .08$, $p = .52$). The interaction of clustered conceptual span with distance was significant, as indicated by a moderated multiple regression of judgment accuracy onto clustered conceptual span (varying continuously), distance (short, long), and their cross-product (Fchange(1,124) = 4.53, $p < .05$, compared to a model without the interaction term).

Following previous research with complex span measures (Miyake et al., 1994a), we divided our participants into high, low, and mid span, to further investigate the interaction between span and memory load (i.e., distance between critical words in the sentences of the anomaly judgment task). We aimed to place the upper, middle, and lower one-third of the participants into the high, low, and mid span categories, respectively. Due to ties in the conceptual span scores, the actual numbers were 23 high span participants (or upper 36% of participants), 22 mid span participants (or middle 34% of participants), and 19 low span participants (or lower 30% participants) in clustered conceptual span and 24 high span participants (or upper 37% of participants), 21 mid span participants (or middle 33% of participants), and 19 low span participants (or lower 30% of participants) in non-clustered conceptual span. The percentages correct were entered into a mixed factor ANOVA, crossing the factors distance (short, long) and conceptual span (high, mid, low). Distance was treated as a repeated measurements factor and conceptual span as a pseudo-experimental grouping factor. We calculated two such ANOVAs, one for clustered conceptual span and one for non-clustered conceptual span. In the ANOVA for clustered conceptual span, we found a main effect of distance ($F(1,63) = 23.87$, $p < .001$, $MSe = .027$, discussed above), an interaction between distance and clustered conceptual span ($F(2,61) = 5.07$, $p < .01$, $MSe = .027$), and a trend towards an effect of clustered conceptual span ($F(2,61) = 2.85$, $p < .7$, $MSe = .046$). The interactive effect of distance and clustered conceptual span on proportion correct in anomaly judgment is depicted in Fig. 1. In the short distance condition, judgment accuracy was below ceiling level and not differentiated by clustered conceptual span (88, 90, and 88% correct for low, mid, and high span participants, respectively, $F(2,62) = .12$, $p = .89$, $MSe = .015$). By contrast, in the long distance condition, judgment accuracy was differentiated by conceptual span ($F(2,61) = 4.38$, $p < .5$, $MSe = .058$) due the fact that low span subjects made 24% more errors than high span subjects ($F(1,40) = 7.56$, $p < .01$, $MSe = .067$). Compared to the short distance condition, judgment accuracy in the long distance condition did not decrease

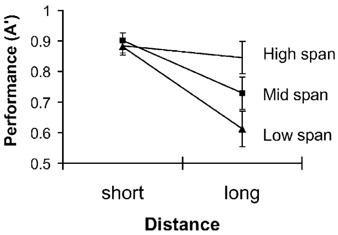

Fig. 1. Performance (A') in the anomaly judgment task as a function of clustered conceptual span and distance between the adjective and the noun in Study 3. Short, 1 intervening adjective; long, 3 intervening adjectives. Vertical bars denote standard errors. For the calculation of A' see footnote 9.

[9] Several participants showed a tendency to be over-rejecting in the anomaly judgment task. To correct for response bias, A' was used as an accuracy measure in this task. A' provides an unbiased estimate of the proportion correct in a two-alternative forced choice procedure (Pollack & Norman, 1964) and was calculated as $A' = .5 + (y - x)(1 + y - x)/4y(1 - x)$, where x = false alarm rate and y = hit rate (Grier, 1971, see e.g., Linebarger, Schwartz, & Saffran, 1983).

reliably for high span participants (4% more errors, $F(1,22) = 1.58$, $p = .22$, $MSe = .010$), whereas it did decrease for both mid span participants (17% more errors, $F(1,21) = 19.14$, $p < .001$, $MSe = .016$) and low span participants (27% more errors, $F(1,18) = 11.18$, $p < .01$, $MSe = .055$, but not significantly more for low than for mid span participants, $p = .26$). In the ANOVA for non-clustered conceptual span, we found a main effect of distance ($F(1,63) = 23.87$, $p < .001$, $MSe = .027$, as discussed above) and a main effect of non-clustered conceptual span ($F(2,61) = 7.6$, $p < .01$, $MSe = .038$). The interaction between distance and non-clustered conceptual span fell short of significance ($F(2,61) = 2.25$, $p = .11$, $MSe = .027$).

Anomaly judgment response times. We also analyzed response times for correct judgments adjective–noun combinations in the anomaly judgment task. Participants' response times were longer in the long (852 ms) than short distance adjective–noun condition (800 ms) ($F(1,63) = 8.96$, $p < .01$, $MSe = 15681$). Response times did not show an interaction between distance and clustered or non-clustered conceptual span, excluding a speed-accuracy trade-off as an explanation for the interaction of those same two factors in the accuracy analysis. Neither non-clustered nor clustered conceptual span correlated with response time in the short distance condition, in the long distance condition, or across both distance conditions.

Serial position effects in clustered conceptual span. In addition to the overall score on the clustered conceptual span test, we also calculated the score for the first, second, and third cluster (i.e., semantic category) in the memory lists separately and correlated each of these scores with text comprehension, anomaly judgment across both distance conditions, anomaly judgment in the short distance condition, and anomaly judgment in the long distance condition (see Table 7). There was a main effect of the serial position of the cluster in the clustered conceptual span test ($F(2,126) = 34.89$, $p < .001$, $MSe = 5.63$), that took the form of a recency effect, such that performance was better for the third than second cluster ($F(1,61) = 21.89$, $p < .001$, $MSe = 5.31$) and better for the second than first cluster 1 ($F(1,61) = 15.39$, $p < .001$, $MSe = 5.28$) (see Table 6). The general pattern of correlation results which emerged was that conceptual span performance on the second and third cluster each correlated moderately and significantly with text comprehension ($r = .37$, $p < .01$ and $r = .38$, $p < .01$, respectively), anomaly judgment across both distance conditions ($r = .28$, $p < .05$ and $r = .33$, $p < .05$, respectively), and anomaly judgment in the long (but not short) distance condition ($r = .29$, $p < .05$ and $r = .28$, $p < .05$, respectively). By contrast, conceptual span performance on the first cluster did not correlate significantly with any of the criterion measures. This lack of a correlation does not seem to be due to some measurement problem with the first cluster measure. The variability of the first cluster measure was about the same as for the third cluster (see Table 6), which did correlate with the criterion measures, and correlation of the first cluster measure with the second cluster measure was relatively high ($r = .60$, $p < .01$).

Discussion

The results of Study 3 suggest that the ability of the conceptual span task to predict comprehension rests to a large extent on the fact that it provides an index of the capacity of semantic STM over and beyond its potential for indexing the ability to cluster a mixed list of words into its semantic categories. This is evident from two sets of findings. First, clustered conceptual span, in which clustering ability should play no role or a minimal role, still correlated substantially with text comprehension (with a correlation value as high as for non-clustered span) and moderately with anomaly judgment (somewhat lower than non-clustered span did). Second, clustered conceptual span shared a substantial part of variance with non-clustered conceptual span in accounting for text comprehension (13% of 17%) and anomaly judgment (10% of 26%). These findings not withstanding it is also evident from our results that non-clustered conceptual span may measure to some extent additional individual differences in clustering ability, given that non-clustered conceptual span accounted for unique variance in text comprehension (4% of 17%, albeit it only at a marginally significant level) and anomaly judgment (16% of 26%). Clustering ability itself may require semantic STM to support the on-line reorganization of category exemplars. However, clustering ability may also rely on factors that are independent of semantic STM capacity. Therefore, we recommend use of the clustered conceptual span task to index semantic STM capacity in future studies.

Our results provide further support for the proposal in the neuropsychological literature that semantic STM is separable from phonological STM and that semantic STM stores unintegrated word meanings to support on-line semantic integration during sentence processing (Hanten & Martin, 2000; Martin & Romani, 1994; Romani & Martin, 1999). In Study 2, we found that on-line semantic anomaly judgment was predicted by conceptual span, a measure of semantic STM capacity but not by non-word span, a measure of phonological STM capacity. Furthermore, in Study 3, we found that conceptual span interacts with memory load during on-line semantic anomaly judgment. In particular, low and mid span participants showed an effect of memory load (i.e., decreased performance when the adjective and noun creating the anomaly were separated by more adjectives), whereas high span participants did not (for similar interactions between reading span and memory load see Daneman &

Carpenter, 1980, and Miyake, Carpenter, & Just, 1994a,b). Martin and colleagues had previously reported such an interactive effect for anomalies involving adjective–noun and verb–noun combinations when comparing neuropsychological patients with a relative selective deficit in semantic STM with normal control patients (Hanten & Martin, 2000; Martin & Romani, 1994). Moreover, in a sentence production task, Martin and Freedman (2001) found that patients with a semantic STM deficit, but neither patients with a phonological STM deficit nor normal controls, showed a pronounced delay in speech onset when the semantic memory load in the first noun phrase was increased from one to two nouns. The finding that conceptual span predicts text comprehension is consistent with the hypothesis that semantic STM stores unintegrated word meanings to support their on-line semantic integration (see Discussion of Study 2). Taken together, these findings suggest that semantic STM stores unintegrated word meanings to enable their on-line semantic integration and that the capacity of this system varies along a continuum in both normal individuals and STM patients.

One possible interpretation for the correlation we found between performance on the conceptual span and the comprehension (and reasoning) tests, may involve the ability to resist proactive interference (PI). Because the conceptual span test presents words from the same semantic categories across trials (using a small pool with replacement), PI is likely to be induced. In fact inducing PI was a rationale for the study, since we wanted to minimize LTM contributions and thus obtain a cleaner measure of the information maintained in the activation based semantic STM system. This is motivated by many studies indicating that STM is better protected from PI than LTM (Cowan, 2001; Craik & Birtwistle, 1971; Goshen-Gottstein, Ashkenazi, & Usher, submitted; Halford et al., 1988; Nairne et al., 1997; Wickens, Moody, & Dow, 1981)[10,11] Nevertheless, two factors may contribute to individual differences in conceptual span. The first factor is the capacity of the semantic STM system (Haarmann & Usher, 2001; see also Cowan, 2001, and Baddeley, 2001). The second factor is the ability to resist PI in LTM. Consistent with this, Kane and Engle (2000) have recently reported that participants with lower operational spans are more strongly affected by PI manipulations. The power of the conceptual span to predict comprehension could be due to either of these two factors.

However, one the findings in Experiment 3 suggests that an important source for the individual differences in reasoning and comprehension is the STM memory capacity (Cowan, 2001). The correlation between conceptual span and the two comprehension measures was highest and significant for the last two clusters and relatively low and non-significant at the first cluster. These results are consistent with previous findings where performance on immediate probed recall of list-final (but not list-initial) items predicted performance on complex cognitive tasks (Cantor et al., 1991; Cohen & Sandberg, 1977). This indicates that the process that is critical for predicting the correlation is the STM capacity rather than retrieval from LTM and its influence by PI, because items at the beginning of a memory list are more likely to be retrieved from LTM than items at the end of a memory list due to their greater likelihood of displacement from STM (Haarmann & Usher, 2001, experiment 1; Davelaar & Usher, 2001).

A still further finding also suggests that the critical factor at work is the STM memory capacity and not the ability to resist PI, namely, the correlation between conceptual span and comprehension of distant anomalies. This finding can be naturally explained by the fact that in a capacity-limited STM system later items are more likely to displace earlier items, especially when the memory load is high and the memory capacity is low (Haarmann & Usher, 2001). However, it is difficult to see how resistance to proactive interference could explain the same finding. If pro-active interference were at work, the anomalous adjective would interfere with the retention of the following adjectives, but not vice versa so that there would be no effect of number of intervening words (distance).

General discussion

Most simple span measures used in correlation studies of reading comprehension are based on serial order recall (Daneman & Carpenter, 1980; LaPointe & Engle, 1990; Turner & Engle, 1989), a procedure that is likely to strongly engage phonological processes (Baddeley, 1986). Recently, a number of theorists have suggested on the basis of different but convergingconsiderations that there is a non-phonological

[10] In immediate recall, short but not long lists of items are protected from the build-up of PI across lists of same-category items (Cowan, 2001; Halford et al., 1988; Wickens et al., 1981). When long lists are used, recency but not pre-recency items are protected from PI in immediate recall (Craik & Birtwistle, 1971; Goshen-Gottstein et al., submitted). Nairne et al. (1997) found that word length effects in immediate serial recall are only obtained after the first few trials when words are repeatedly sampled from a limited pool, suggesting that word length effects arise only when participants shift their retrieval from LTM to STM because retrieval from LTM becomes more difficult due to PI.

[11] By contrast, in delayed free recall after continuous distraction (CD) both recency and pre-recency items are affected by PI (Goshen-Gottstein et al., submitted), suggesting that recency effects in CD have their locus in LTM (Goshen-Gottstein et al., submitted), rather than in the same memory system as recency effects in immediate free recall (Bjork & Whitten, 1974).

component in verbal STM. For example, Baddeley (2000, 2001) suggested that a new storage component, with capacity of about three items (Baddeley, 2001) and distinct from the phonological loop, needs to be assumed in his model of WM to explain patterns of data in verbal STM. Similarly, Cowan (2001) reviews data that demonstrate a capacity-limited STM system that can hold about three to four items when rehearsal and chunking are prevented, irrespective of whether the to-be-retained items involve verbal or visual representations. A lexical/semantic buffer in STM was proposed by Martin et al. (1994) and Haarmann and Usher (2001) have presented a model in which such a lexical/semantic buffer is implemented by active representations in the pre-frontal cortex, leading to capacity limitations of similar magnitude but which are subject to variations that depend on biological parameters (such as the mutual inhibition and recurrent excitation of the neural circuits of the pre-frontal cortex; Haarmann & Usher, 2001; Usher, Cohen, Haarmann, & Horn, 2001).

Reasoning that the low correlations between existing simple span measures and comprehension (Daneman & Carpenter, 1980; LaPointe & Engle, 1990; Turner & Engle, 1989) may arise from their failure to sufficiently engage semantic STM, we investigated whether comprehension and verbal problem solving would be better predicted by conceptual span, a novel, relative index of the capacity of semantic STM. This result was indeed obtained in Study 2 where conceptual span predicted verbal problem solving performance significantly better than either word or non-word span. In addition, we found that conceptual span accounted for unique variance in verbal problem solving, anomaly judgment, and comprehension of pronoun texts above and beyond variance explained by word span and non-word span. Moreover, in Study 1 we found that word span no longer predicted text comprehension when conceptual span was controlled for, suggesting that the correlation was mediated by capacity differences in semantic STM. These results suggest that semantic STM plays an important role in on-line meaning integration, while phonological STM does not. However, the possibility that phonological STM may support a different aspect of sentence comprehension cannot be excluded. It has been suggested that phonological STM helps to maintain a verbatim representation of the words in a sentence, so it can be re-processed from memory in case immediate, online meaning integration fails (Baddeley, 1986; Martin & Romani, 1994; Martin et al., 1994). Such re-processing may explain why, in Study 1, word span still predicted sentence comprehension when conceptual span was controlled for, and why, in Study 2, word span predicted comprehension of GRE texts as well as conceptual span.

One important issue concerns the nature of the word meanings stored in semantic STM. Our findings fully support the proposal that semantic STM stores unintegrated word meanings to help with their on-line meaning integration during sentence processing (Martin & Romani, 1994; Hanten & Martin, 2000; cf. our discussion of the interaction between distance and conceptual span in Study 3). However, it still seems an open question as to whether semantic STM also supports the storage of integrated word meanings, consistent with the role Baddeley (2001) attributes to the episodic buffer and consistent with computational models that assume a capacity limit for the storage of verbrole-bindings (e.g., John-is-agent-of-love, Hummel & Holyoak, 1997) and propositions in texts (Goldman & Varma, 1995). The proposal that semantic STM stores integrated word meanings raises the question as to why semantic STM patient A.B. performed normal in story recall (Romani & Martin, 1999). Perhaps one possibility is that patient A.B. did have a problem storing integrated word meanings but this problem was not reflected in his story recall performance because the story paragraphs may have expressed familiar themes for which schemas exist in long-term memory. Linking propositions in short-term memory to schemas, for example, via retrieval structures in long-term working memory (Ericsson & Kintsch, 1995), may have minimized the need to actively maintain propositions in semantic STM. These and other considerations suggest a need for further research that tests these two alternative hypotheses on the properties of the semantic STM system, that is, whether it stores merely unintegrated word meanings or also integrated ones.

The conceptual span test may be used to further investigate the role of semantic STM in supporting meaning integration during on-line sentence comprehension. In online sentence comprehension, where the role of phonological STM is believed to be minimal (Butterworth et al., 1986; Caplan & Waters, 1999; Martin & Romani, 1994), there is evidence for a capacity-limited semantic STM. In particular, several studies found that semantic processes occurring during on-line sentence comprehension are modulated conjointly by language WM load and capacity (Gunter et al., 1995; Haarmann et al., in press; Just & Carpenter, 1992; Miyake et al., 1994a; Munte, Schiltz, & Kutas, 1998). To the extent that the semantic effects in these studies arise in semantic STM, we predict that their time course is modulated also by individual differences in conceptual span.

The conceptual span test may also be used to investigate the role of semantic STM in reasoning and fluid intelligence. Semantic STM provides a system for maintaining concepts in or near the focus of awareness (cf. Cowan, 2001), enabling rapid access to stored concepts and rapid computation of the relationships among them. These properties may be especially important in problem solving, analogical reasoning tasks

(Hummel & Holyoak, 1997) and in fluid intelligence tasks (Duncan, Emslie, Williams, Johnson, & Freer, 1995, 2000). Consistent with this, we found in Study 2 that conceptual span correlated best with verbal problem solving (i.e., $r = .51$; see also Cohen & Sandberg, 1977). Correlations between a fluid intelligence task (Cattell's Culture Fair) and storage-plus-processing measures of WM (operation span, reading span and counting span) in the range of .24–.29 have been reported by Engle, Tuholski, Laughlin, and Conway (1999). In that study, the correlations with storage-only measures were lower. In a preliminary study, we found an even higher correlation between conceptual span and the Cattell test ($r = .47$, $p < .01$, $N = 69$). Consistent with Engle et al., however, the correlation between word span and the Cattell was lower and non-significant ($r = .18$, $p > .10$). We suggest that, although it is a storage-only measure, the conceptual span correlates highly with fluid intelligence, supporting the view that semantic STM is involved in the rapid computation of information, whereas phonological STM is used more as a backup system. Future research could investigate the relations between semantic STM and attentional processes in their ability to predict fluid intelligence.

Another interesting issue that should be addressed in future research is the relation between the semantic STM system (Haarmann & Usher, 2001; Hanten & Martin, 2000; Martin & Freedman, 2001; Martin et al., 1994) and a system in the prefrontal cortex that may mediate the use of lexical and task context to control information processing. This system may coincide with the episodic buffer proposed by Baddeley (2001) or the context module proposed by Cohen and Servan-Schreiber (1992). It has been shown that the use of lexical and task context to control information processing is impaired in Schizophrenia patients with formal thought disorder (Bustini et al., 1999; Cohen, Barch, Carter, & Servan-Schreiber, 1999; Cohen, Braver, & O'Reilly, 1996; Cohen & Servan-Schreiber, 1992). If contextual maintenance relies on a lexical/semantic STM buffer in the pre-frontal cortex, which is distinct from the phonological loop, we expect that schizophrenia patients with language characteristics of formal thought disorder (e.g., incoherent speech) will show a pathologically reduced semantic STM capacity, as measured by the conceptual span. By contrast, we expect such patients to show an intact phonological STM, consistent with the finding that they have no problems with immediate serial recall of digits and words (Cohen et al., 1999).

The conclusion of the correlation studies above needs to be qualified with regard to the reliability values of the various measures. The reliability of a measure gives the theoretical maximum of its correlation with other measures (Daneman & Merikle, 1996, in Spearman, 1904). Span measures therefore need to be sufficiently reliable to exclude the possibility that differences in the magnitudes of the correlations of two span measures with a comprehension measure are an artifact of a low reliability in one of the measures. Nunnally (1978) defined .7 as the criterion for minimum reliability adequacy. The internal consistency of conceptual span (split-half = .85) was well above that. Since in the word- and reading-span we used the traditional procedure with a break-off criterion (where trials are not independent), internal consistency measures, such as, Cronbach's alpha and split-half reliability, cannot be computed for our data. However, internal consistencies have been reported in the literature for tasks using a non-traditional span procedure without break-off. For example, Daneman and Merikle (1996) give in their review, the values of .79 and .80, as the average over four experiments, for word- and reading-span-like tests, respectively (see also Engle et al., 1999; Friedman & Miyake, 2000; Oberauer & Süß, 2000). While it is possible that the procedural difference induced by the break-off criterion could affect the reliability values, we believe that this is unlikely to have affected our results for the following reasons. First, high test-retest reliability has been reported for other simple and complex span measures that used the traditional procedure, namely, .83 for digit span (Wechsler, 1981) and .88 for operation span (Klein & Fiss, 1999). Second, in our measures (e.g., Study 2), the reliability was at least .58 for word and non-word span (as determined by the correlation between word and non-word span) and .56 for reading span (correlation between word and reading span), which is higher than the maximum correlation we obtained between conceptual span and a comprehension measure (i.e., correlation between conceptual span and verbal problem solving = .51). This pattern of correlation makes it highly unlikely that a low reliability in the traditional span measures caused them to show an artifactually low correlation with comprehension. Nevertheless, we think that future studies using non-traditional span measures with independent trials (so that internal consistency can be computed) are needed in order to fully establish the relative contributions of processes involved in reading comprehension.

To conclude, we believe that the conceptual span measures individual variation in the capacity of a general item STM. This new procedure avoids potential confounds with processing efficiency and dual-tasking ability present in reading span and reduces the contribution of the decay-based phonological loop present in word span. Conceptual span may, therefore, provide an efficient new tool for investigating the role of semantic STM in meaning integration and cognitive control and for helping to determine the differential contribution of storage and processing components of WM to higher cognitive processes.

Acknowledgments

Preliminary findings were reported at the 40th annual meeting of the Psychonomic Society, Los Angeles, CA, November, 1999. We thank Olga Borisenko, Sandra Carter, Heather Coward, Starr Hume, Becky Karman, Orsolya Kraut, Sally Mahmoud, Tammy Woodburry, and Amir Zahir for their help with running participants in the experiments and Akira Miyake for very helpful comments on the manuscript. We also thank Meredyth Daneman, Randall Engle, Akira Miyake, Randi Martin, and Klaus Oberauer for their advice about span test reliability.

Appendix A. Examples of items in the verbal problem solving test in Study 2

Problem type 1

In order to compete the work of a mail order concern it is necessary to have a minimum of three workers each day. Alice can work on Mondays, Wednesdays, and Fridays. Betty cannot report for work on Wednesdays. Carol can report for work on Tuesdays and Wednesdays only. Dorothy cannot work on Fridays. Edith is available anytime except on the first Monday and Thursday of the month.

Which three could you count on to report for work on Friday?

(a) Alice, Betty, and Dorothy
(b) Alice, Carol, and Dorothy
(c) Betty, Carol, and Edith
(d) Carol, Betty, and Allice
(e) Allice, Betty, and Edith

Problem type 2

Adam runs faster than Stuart. (2 answers required)

(a) Swinthin is the champion runner.
(b) Adam can run further than Swinthin
(c) Adam can run as fast as Swinthin.
(d) Swinthin can run faster than Stuart.

Problem type 3

A man drove from Appleby to Trytown. Shortly after passing through Ester he stopped for coffee at Broughton, which was the halfway point on this journey.

Which one is the longest distance?

(a) Appleby to Ester
(b) Ester to Trytown
(c) Broughton to Trytown
(d) Ester to Broughton

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2). New York: Academic Press.

Baddeley, A. D. (1970). Estimating the short term component in free recall. *British Journal of Psychology, 61*, 13–15.

Baddeley, A. D. (1972). Retrieval rules and semantic coding in short-term memory. *Psychological Bulletin, 78*, 379–385.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Baddeley, A. D. (2000). The episodic buffer: a new component of working memory. *Trends in Cognitive Science, 4*, 417–423.

Baddeley, A. D. (2001). The magic number and the episodic buffer (Commentary on target article by Cowan). *Behavioral and Brain Sciences, 24*(1).

Baddeley, A. D., Logie, R., Nimmo-Smith, I., & Brereton, N. (1985). Components of fluent reading. *Journal of Memory and Language, 24*, 119–131.

Barrett, J., & Williams, G. (1990). *Test your own aptitude*. London: Kogan Page.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology, 6*, 173–189.

Budd, D., Whitney, P., & Turley, K. J. (1995). Individual differences in working memory strategies for reading expository text. *Memory & Cognition, 6*, 735–748.

Bustini, M., Stratta, P., Daneluzzo, E., Pollice, R., Prosperini, P., & Rossi, A. (1999). Tower of Hanoi and WCST performance in schizophrenia: problem-solving capacity and clinical correlates. *Journal of Psychiatric Research, 33*, 285–290.

Butterworth, B., Campbell, R., & Howard, D. (1986). The uses of short-term memory: A case study. *The Quarterly Journal of Experimental Psychology, 38A*, 705–737.

Cantor, J., Engle, R. W., & Hamilton, G. (1991). Short-term memory, working memory, and verbal abilities: How do they relate? *Intelligence, 15*, 229–246.

Caplan, D., & Waters, G. S. (1990). Short-term memory and language comprehension: A critical review of the neuropsychological literature. In G. Vallar & T. Shallice (Eds.), *Neuropsychological impairments of short-term memory* (pp. 337–389). Campidge: Campidge University Press.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences, 22*(1), 77–126.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404–431.

Cohen, J. D., Barch, D. M., Carter, C. S., & Servan-Schreiber, D. (1999). Schizophrenic deficits in the processing of context: Converging evidence from three theoretically motivated cognitive tasks. *Journal of Abnormal Psychology, 108*, 120–133.

Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control and schizophrenia: Recent developments and current challenges. *Philosophical Transactions of the Royal Society in London (B): Biological Sciences, 351*, 1515–1527.

Cohen, R. L., & Sandberg, T. (1977). Relation between intelligence and short-term memory. *Cognitive Psychology, 9*, 534–554.

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological Review, 99*, 45–77.

Conway, A. R. A., & Engle, R. W. (1996). Individual differences in working memory capacity: More evidence for a general capacity theory. *Memory, 4*, 577–590.

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

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–185.

Craik, F. I. M., & Birtwistle, J. (1971). Proactive inhibition in free recall. *Journal of Experimental Psychology, 91*, 120–123.

Craik, F. I. M., & Levy, B. A. (1970). Semantic and acoustic information in primary memory. *Journal of Experimental Psychology, 86*, 77–82.

Crawford, J. R., Mychalkiw, B., Johnson, D. A., & Moore, J. W. (1996). WAIS-R shortforms: Criterion validity in healthy and clinical samples. *British Journal of Clinical Psychology, 35*, 638–640.

Crosson, B., Rao, S. M., Woodley, S. J., Rosen, A. C., Bobholz, J. A., Mayer, A., Cunningham, J. M., Hammeke, T. A., Fuller, S. A., Binder, J. R., Cox, R. W., & Stein, E. A. (1999). Mapping of semantic, phonological, and orthographic verbal working memory in normal adults with functional magnetic resonance imaging. *Neuropsychology, 13*, 171–187.

Crowder, R. G. (1979). Similarity and serial order in memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Vol. 13*. New York: Academic Press.

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

Daneman, M., & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*, 561–584.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*(4), 422–433.

Davelaar, E. J., & Usher, M. (2001). Exploring a computational account of item shortterm memory. In *The Third International Memory Conference (ICOM-3), Valencia, Spain*, July 16–21.

Dixon, P., Le Fevre, J., & Twilley, L. C. (1989). World knowledge and working memory as predictors of reading skill. *Journal of Educational Psychology, 80*, 465–472.

Duncan, J., Emslie, H., Williams, P., Johnson, R., & Freer, C. (1995). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology, 30*, 257–303.

Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A., Newell, F. N., & Emslie, H. A. (2000). Neural basis for general intelligence. *Science, 289*(July 21), 457–460.

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

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

Ernst, G. W., & Newell, A. (1969). *GPS: A case study in generality and problem solving*. New York: Academic Press.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*, 211–245.

Friedman, N. P., & Miyake, A. (2000). Differential roles for visuospatial and verbal working memory in situation model construction. *Journal of Experimental Psychology: General, 129*, 61–83.

Gabrieli, J. D. E., Poldrack, R. A., & Desmond, J. E. (1998). The role of left prefrontal cortex in language and memory. *Proceedings of the National Academy of Sciences, 95*, 906–913.

Gilhooly, K. J., Logie, R. H., Wetherick, N. E., & Wynn, V. (1993). Working memory and strategies in syllogistic reasoning tasks. *Memory & Cognition, 21*, 115–124.

Gilhooly, K. J., Logie, R. H., & Wynn, V. (1999). Syllogistic reasoning tasks, working memory, and skill. *European Journal of Cognitive Psychology, 11*, 473–498.

Glanzer, M. (1972). Storage mechanisms in recall. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 5). New York: Academic Press.

Goldman, S., & Varma, S. (1995). CAPing the construction-integration model of discourse comprehension. In C. A. Mannes (Ed.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 337–358). Hillsdale, NJ: Lawrence Erlbaum.

Goshen-Gottstein, Y., Ashkenazi, A., & Usher, M. (submitted). Proactive interference attenuates recency in the continuous distractor task but not in free recall.

Greeno, J. G. (1973). The structure of memory and the process of solving problems. In R. L. Solso (Ed.), *Contempary issues in cognitive psychology: The Loyola symposium*. Washington, DC: Winston.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin, 75*, 424–429.

Gunter, T. C., Jackson, J. L., & Mulder, G. (1995). Language, memory, and aging: An electrophysiological exploration of the N400 during reading of memory-demanding sentences. *Psychophysiology, 32*, 215–229.

Haarmann, H. J., Cameron, K. A., & Ruchkin, D. S. (in press). Short-term semantic retention during on-line sentence comprehension: Brain potential evidence from filler-gap constructions. *Cognitive Brain Research*.

Haarmann, H. J., Just, M. A., & Carpenter, P. A. (1997). Aphasic sentence comprehension as a resource deficit: A computational approach. *Brain and Language, 59*, 76–120.

Haarmann, H. J., & Kraut, O. (2001). The effect of thematic role bindings on storage in verbal short-term memory. In *The Third International Memory Conference (ICOM-3), Valencia, Spain*, July 16–21.

Haarmann, H. J., & Usher, M. (2001). Maintainance of semantic information in capacity-limited item short-term memory. *Psychonomic Bulletin & Review, 8*(3), 568–578.

Halford, G. S., Maybery, M. T., & Bain, J. D. (1988). Set-size effects in primary memory: An age-related capacity limitation. *Memory & Cognition, 16*, 480–487.

Hanten, G., & Martin, R. C. (2000). Contributions of phonological and semantic short-term memory to sentence processing: Evidence from two cases of closed head injury in children. *Journal of Memory and Language, 43*, 335–361.

Howe, M. L., Rabinowitz, F. M., & Powell, T. L. (1998). Individual differences in working memory and reasoning-remembering relationships in solving class-inclusion problems. *Memory & Cognition, 26*, 1089–1101.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury Press.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427–466.

Jackson, M. D., & McClelland, J. L. (1979). Processing determinants of reading speed. *Journal of Experimental Psychology: General, 108*, 151–181.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122–149.

Kane, M. J., & Engle, R. W. (2000). Working memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 336–358.

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.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language, 30*, 580–602.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kintsch, W., Healy, A. F., Hegarty, M., Pennington, B. F., & Salthouse, T. A. (1999). Eight questions and some general issues. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge: Cambridge University Press.

Klein, K., & Fiss, W. H. (1999). The reliability and stability of the Turner and Engle working memory task. *Behavior Research Methods, Instruments, & Computers, 31*, 429–432.

LaPointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 1118–1133.

Levy, B. A., & Baddeley, A. (1971). Recall of semantic clusters in primary memory. *Quarterly Journal of Experimental Psychology, 23*, 8–13.

Linebarger, M. C., Schwartz, M., & Saffran, E. M. (1983). Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition, 13*, 361–392.

Martin, R. C. (1990). The consequences of reduced memory span for the comprehension of semantic versus syntactic information. *Brain and Language, 38*, 1–20.

Martin, R., & Freedman, M. (2001). Verbal working memory: The ins and outs of phonological and lexical-semantic retention. In H. L. Roediger, J. S. Nairne, I. Neath, & A. M. Suprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder*.

Martin, R. C., & Romani, C. (1994). Verbal working memory and sentence comprehension: A multiple-components view. *Neuropsychology, 9*(4), 506–523.

Martin, N., Saffran, E. M., & Dell, G. S. (1996). Recovery in deep dysphasia: Evidence for a relation between auditory-verbal STM capacity and lexical errors in repetition. *Brain and Language, 52*, 83–113.

Martin, R. C., Shelton, J. R., & Yaffee, L. S. (1994). Language processing and working memory: Evidence for separate phonological and semantic capacities. *Journal of Memory and Language, 33*, 83–111.

Masson, M. E. J., & Miller, J. A. (1983). Working memory and individual differences in comprehension and memory of text. *Journal of Educational Psychology, 75*, 314–318.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81–97.

Miyake, A., Carpenter, P. A., & Just, M. A. (1994a). A capacity approach to syntactic comprehension disorders: Making normal adults perform like aphasic patients. *Cognitive Neuropsychology, 11*(6), 671–717.

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

Munte, T. F., Schiltz, K., & Kutas, M. (1998). When temporal terms belie conceptual order. *Nature, 395*, 71–73.

Nairne, J. S., Neath, I., & Serra, M. (1997). Proactive interference plays a role in the word length effect. *Psychonomic Bulletin & Review, 4*, 541–545.

Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Oberauer, K., & Süß, H. M. (2000). Inhibition in working memory. A comment on Jenkins, Myerson, Hale, and Fry (1999). *Psychonomic Bulletin & Review, 7*, 727–733.

Pollack, L., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomical Science, 1*, 125–126.

Potter, M. C. (1993). Very short-term conceptual memory. *Memory & Cognition, 21*(2), 156–161.

Raven, J. C., Court, J. H., & Raven, J. (1977). *Standard progressive matrices*. London: H.K. Lewis & Company.

Raser, G. (1972). Recoding of semantic and acoustic information in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 11*, 692–697.

Romani, C., & Martin, R. (1999). A deficit in the short-term retention of lexical-semantic information: Forgetting words but remembering a story. *Journal of Experimental Psychology: General, 128*, 56–77.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Sperling, G. (1960). The information available in brief visual presentation. *Psychological Monographs, 74*, entire issue.

St. George, M., Mannes, S., & Hoffman, J. E. (1997). Individual differences in inference generation: An ERP analysis. *Journal of Cognitive Neuroscience, 9*, 776–787.

Shulman, H. G. (1970). Encoding and retention of semantic and phonemic information in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 9*, 499–508.

Shulman, H. G. (1972). Semantic confusion errors in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 11*, 221–227.

Stromswold, K., Caplan, D., Alpert, N., & Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language, 52*, 452–473.

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

Usher, M., & Cohen, J. D. (1999). Short-term memory and selection processes in a frontal-lobe model. In D. Heinke, G. W. Humphries, & A. Olsen (Eds.), *Connectionist models in Cognitive Neuroscience* (pp. 78–91). London: Springer.

Usher, M., Cohen, J. D., Haarmann, H. J., & Horn, D. (2001). Neural mechanism for the magical number 4: Competitive interactions and non-linear oscillations (Commentary on target article by Cowan). *Behavioral and Brain Sciences, 24*(1), 151–152.

Waters, G., & 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.

Wechsler, D. A. (1981). *Wechsler adult intelligence scales (revised)*. San Antonio, TX: Psychological Corporation, Harcourt Brace Jovanovich.

Wickens, D. D., Moody, M. J., & Dow, R. (1981). The nature and timing of the retrieval process and of interference effects. *Journal of Experimental Psychology: General, 110*, 1–20.