Models of verbal working memory capacity: What does it take to make them work?

Published in final edited form as:
Psychol Rev. 2012 July ; 119(3): 480–499. doi:10.1037/a0027791.

Models of Verbal Working Memory Capacity: What Does It Take to Make Them Work?

Nelson Cowan, **Jeffrey N. Rouder**, **Christopher L. Blume**, and **J. Scott Saults**
University of Missouri

Abstract

Theories of working memory (WM) capacity limits will be more useful when we know what aspects of performance are governed by the limits and what aspects are governed by other memory mechanisms. Whereas considerable progress has been made on models of WM capacity limits for visual arrays of separate objects, less progress has been made in understanding verbal materials, especially when words are mentally combined to form multi-word units or chunks. Toward a more comprehensive theory of capacity limits, we examine models of forced-choice recognition of words within printed lists, using materials designed to produce multi-word chunks in memory (e.g., *leather brief case*). Several simple models were tested against data from a variety of list lengths and potential chunk sizes, with test conditions that only imperfectly elicited the inter-word associations. According to the most successful model, participants retained about 3 chunks on average in a capacity-limited region of WM, with some chunks being only subsets of the presented associative information (e.g., *leather brief case* retained with *leather* as one chunk and *brief case* as another). The addition to the model of an activated long-term memory (LTM) component unlimited in capacity was needed. A fixed capacity limit appears critical to account for immediate verbal recognition and other forms of WM. We advance a model-based approach that allows capacity to be assessed despite other important processing contributions. Starting with a psychological-process model of WM capacity developed to understand visual arrays, we arrive at a more unified and complete model.

A fundamental aspect of cognitive processing is the capacity limit in working memory (WM), that is, heightened access to a limited amount of information recently encountered or recently activated through long-term memory (LTM) retrieval (Baddeley & Hitch, 1974; Cowan, 1988, 1999). Various theorists have suggested that there is a limit in how many psychologically coherent units, or chunks, can be remembered at once (Broadbent, 1975; Cowan, 2001; Miller, 1956). Although that WM capacity limit has been investigated in detail for arrays of simple visual objects in probe recognition tasks (e.g., Luck & Vogel, 1997; Wheeler & Treisman, 2002), including the psychometric estimation of capacity (e.g., Cowan et al., 2005; Morey, Cowan, Morey, & Rouder, 2011; Rouder et al., 2008), there has been no extension of this approach to verbal lists, or generally to materials in which (1) there is likely to be a substantial LTM contribution, or (2) presented items can be combined to form larger chunks of information. In search of a unified understanding of WM capacity across domains, we discuss capacity limits, develop a capacity model for verbal list item recognition, and compare alternative approaches.

Address Correspondence to: Nelson Cowan, Department of Psychological Sciences, University of Missouri, 18 McAlester Hall, Columbia, MO 65211, CowanN@missouri.edu, Tel. 573-882-4232.

The Nature of Capacity Limits

Miller (1956) suggested that what is now termed WM was limited to a specific number of meaningful units or chunks, a “magic number”. Although Miller’s work has proved influential (Shiffrin & Nosofsky, 1994), researchers working in the verbal domain have often suggested that WM capacity is determined by the number of items that may be rehearsed in a limited time, or “magic spell” (Schweickert, Guentert, & Hersberger, 1990). Baddeley, Thomson, and Buchanan (1975) found that immediate serial recall performance for short lists depended on the length of words to be recalled when rehearsal was allowed, indicating that at least part of WM is governed by time or amount of phonological material, as opposed to the number of chunks (see Baddeley, 1986). Other research showed that there was an advantage for spoken as compared to printed verbal items, arising from an auditory sensory memory representation (see Penney, 1989). These lines of research suggested that, on the surface, capacity is not limited to a fixed number of chunks.

Subsequently, however, several theoretical treatments suggested that there still is a chunk capacity limit, but that the limit can only be seen in special circumstances in which the effects of supplementary memory mechanisms, such as rehearsal, are controlled or subtracted out (e.g., Broadbent, 1975; Cowan, 2001). Although the difficulty of eliminating supplementary mechanisms should not be underestimated, considerable progress has been made in the past few years toward understanding a core WM capacity with supplementary factors eliminated.

There has been a separation in the literature between studies of WM capacity using visual versus verbal materials. On one hand, much effort has gone toward theoretical accounts of the recognition of simple objects within visual arrays, with the application of models in which a discrete number of WM slots is filled by items (e.g., D. Anderson, Vogel, & Awh, 2011; Cowan, 2001; Luck & Vogel, 1997; Rouder et al., 2008; Zhang & Luck, 2008). The resulting models generally fit the data well with a constant capacity, and also have been used to fit sequences in addition to arrays (e.g., Cowan, AuBuchon, Gilchrist, Ricker, & Saults, 2011). On the other hand, a smaller vein of recent research has examined WM capacity for verbal items in recall procedures. In this research, special attention has been paid to the possibility that verbal items can be combined mentally to form larger chunks of information (Chen & Cowan, 2005, 2009a; Cowan, Chen, & Rouder, 2004; Gilchrist, Cowan, & Naveh-Benjamin, 2008, 2009; cf. Glanzer & Razel, 1974; Simon, 1974). In these recall procedures, too, evidence of a component of WM limited to a fixed number of chunks has been obtained. The recall procedures, however, do not easily lend themselves to the same types of process models of WM that have proven useful in recognition procedures. It has not yet been ascertained whether the WM capacity models for simple visual objects can be adapted to a verbal domain with complicating factors such as the formation of multi-item chunks, or LTM activation outside of any capacity limit.

Cowan (2001) suggested that the WM component that is limited by how many chunks it may include (typically 3 or 4 concurrently in young adults) is the focus of attention. In this framework performance is based not only on the focus of attention, but also on a larger set of information, the activated portion of LTM, which includes categorized information in the focus of attention as well as less-analyzed or less-organized features activated outside of the current focus of attention. The latter can include activated sensory features (Cowan, 1988), phonological memory and its maintenance by automatic, verbal rehearsal (Baddeley, 1986), and semantic elements (e.g., Potter & Lombardi, 1990). Much recent work has gone toward understanding the role of LTM in WM tasks (e.g., Unsworth & Engle, 2007).

Unlike Cowan (2001), some researchers believe that the capacity limit of working memory cannot be attributed to the focus of attention (see the many commentaries in Cowan, 2001). We, however, find the view apt given a 1-to-1 tradeoff between verbal-acoustic and visual-spatial capacity under tightly-controlled dual task circumstances (Saults & Cowan, 2007) and other instances of WM tradeoffs between modalities (e.g., Cowan & Morey, 2007; Morey & Cowan, 2004; Morey et al., 2011; Stevanovski & Jolicoeur, 2007).

To explore capacity limits, one must also be able to identify the chunks. This is a difficult problem inasmuch as there is typically no way to know for sure that the correct chunks have been identified. Cowan (2001) suggested that Miller's (1956) seminal finding that immediate memory includes about 7 items or chunks of information was misleading, inasmuch as there was no control for how the encoded items were combined to form chunks. When Broadbent (1975) considered situations in which such factors appeared to be controlled, he found that only about 3 items could be remembered, and Cowan (2001) came up with a similar estimate of 3 to 5 items in a larger and more systematic survey of the literature (concentrating on memory for items in situations in which multi-item chunking seemed improbable).

In the verbal domain, using recall methods, the capacity-limit concept has been extended to materials in which a set of multiple presented words often can make up a single chunk, and in which the chunk size has been estimated and varied. Instead of presenting all of the various work in this domain (e.g., Chen & Cowan, 2005, 2009a; Cowan et al., 2004; Glanzer & Razel, 1974; Johnson, 1978; Simon, 1974; Tulving & Patkau, 1962) we will focus on one verbal recall study that demonstrates the power of the capacity concept and helped us to decide upon the conditions of the empirical tests of our models. Chen and Cowan (2009a) obtained what we believe to be the cleanest evidence of constant capacity across different chunk sizes. That study included several experimental phases with printed words. Familiarization and cued-recall phases set the stage for list recall. Singletons were presented for familiarization along with consistent word pairs (e.g., *brick-hat, king-desk*). In cued recall, the participant saw the first item in a pair and was to indicate the second item or, when a singleton was presented, indicate that there was no pairing. The complete set of pairs and singletons was presented repeatedly until the participant reached a 100% correct criterion of performance. Then, in list recall, participants received lists of 4, 6, 8, or 12 singletons or 4 or 6 learned pairs, always presented one pair of words at a time for 2 s per pair. The word recall responses were typed and, although recall was to be in the presented serial order, Chen and Cowan were interested in the number of items in WM and therefore ignored serial order errors. Results were scored according to the number of chunks accessed, which refers to the recall of at least one word from a learned chunk (learned pair or familiarized singleton).

Several factors had to be controlled. It was expected that the use of visual materials, with each chunk replacing the previous one on the screen, would limit sensory memory. To assess and control the role of phonological rehearsal, half the participants repeated the word *the* during the presentation of list items, which would presumably limit phonological rehearsal (Baddeley, 1986). The other half had no such constraint, and were presumably free to engage in rehearsal. The results are reproduced in Figure 1. When rehearsal was allowed, mean capacity varied systematically across list type. When rehearsal was prevented, quite remarkably, all the types of lists resulted in a mean of almost exactly 3 chunks accessed, in close alignment with the earlier predictions of Broadbent (1975).

Present Recognition Task and Its Contribution

Our primary goal was to use a verbal list memory task similar enough to visual array memory tasks so that formal process models of capacity limits could be assessed. We also wanted to examine effects of chunking in this domain. To these ends, we develop a forced-choice recognition test for verbal lists. Recognition often shows the existence of information in memory more clearly, with less inter-item interference, than does recall (e.g., Dyne, Humphreys, Bain, & Pike, 1990). Forced choice accuracy conveniently minimizes any issue of response biases and has an expected range of .5 to 1.0, with .5 serving as the baseline accuracy when performance reflects uninformed guessing.

The forced-choice task, while providing data amenable to modeling, required the construction of a derived measure, the *adjusted chunk score*, that takes into account the chunk structure, awarding credit for the recognition of chunks rather than individual items. We show that this derived measure reveals a very distinct pattern of data that was not evident from the proportion of items correct. We examine several factors: effects of a WM chunk capacity that can be considered constant within each individual; an LTM storage probability[1] that is not limited by capacity; and a probability that the encoded chunks can include multiple separate, smaller portions of an intended, presented chunk (chunk decomposition). Testing different combinations of these factors, we show that they all must be included in one model in order to reproduce the pattern of data. Although there theoretically could be some other factor that is needed for other types of stimuli, this model greatly extends the domain of prediction of capacity models, in the most natural and straightforward ways we could conceive.

The paradigm for testing various formal models is as follows. On each memory trial, our participants read a list of singletons (e.g., *rain*) or multi-word chunks (e.g., *leather brief case*) on which they had previously been trained, and then were asked about each item in the list. A list-recognition trial from Experiment 1 is illustrated in Figure 2. The participants' task was to determine, in each response, which of two words was present in the list. The words were drawn from the chunks in the list and tested in a random order that did not preserve the order of words within the learned chunks. During the presentation of lists (but not during list item recognition testing), articulatory suppression was used on all trials to prevent covert verbal rehearsal. The purpose of the experiments was to determine which models would prove sufficient across a wide range of conditions (list lengths and chunk lengths).

To reduce the amount of time necessary for training of the chunks, we used familiar word triplets, such as *leather brief case* (Appendix A). We selected them in such a manner that, for a given participant, some triplets were randomly selected to be used as singletons, with only the last word of the triplet appearing in the stimulus materials (e.g., *case*); others to be familiar pairs, with the last two words appearing (e.g., *brief case*); and still others to appear as complete triplets (e.g., *leather brief case*). Conditions for this experiment were selected so as to include a wide range of chunk sizes and list lengths (Table 1). For a given participant, the assignment of examples to chunk sizes remained the same throughout the experiment, and specific chunks were re-used on multiple trials.

[1]In our task, LTM could contribute in two ways. First, the participant could have an episodic memory indicating that the probe word was in the list. Second, the participant could have a heightened state of LTM activation of the probe word without knowing why. These mechanisms may be the recognition and familiarity processes of Jacoby (1991), but we make no attempt to differentiate these two versions of LTM activation.

Theoretical Modeling Approach

We base our models on a simple, discrete-slot model of capacity limits, similar in spirit to the previous work on the analysis of visual arrays (Cowan, 2001; Rouder et al., 2008; Zhang & Luck, 2008). The slot models have proven to be useful in that domain, so the basic question is how these models must be augmented to make them successful in this more complex situation, as well. As a check on our approach we also ask whether these augmented models depend on the capacity-limited mechanism, as we expect, or whether they can do without it. The issues we address are the needed modifications for a fixed slot model to explain the results of a complex test situation. In particular, we focus on conditions in which meaningful units vary in both number and length, and recognition may be accomplished by either recognition of multi-word chunks or their constituent parts. This resembles ordinary life, in which one might remember a presented multi-word ensemble in full (e.g., *tax-exempt code*) or only in part (e.g., *tax-exempt*).

Processes that Might Account for List Item Recognition

To account for results in the procedure we have just described, our models of forced-choice recognition were designed to assess contributions of the following processes:

Chunk capacity—Based on the literature to be discussed and new research, we first state a very simple hypothesis that has not been widely accepted because it has not had adequate support. It is that the information retained in a trial within a WM task can be accurately described as a fixed number of integrated units or chunks (cf. Miller, 1956), and that retrieval of some information about the chunk necessarily implies retrieval of the rest of the chunk on the same memory trial. This hypothesis is meant to apply only when covert verbal rehearsal cannot be used (Chen & Cowan, 2009a). We have found it necessary also to state further hypotheses, however, because the fixed-capacity hypothesis does not entirely hold up as we had initially expected.

LTM contribution—Previous work has established that there are LTM contributions in what had been considered WM tasks (Cowan, 1988; Healy, Fendrich, Cunningham, & Till, 1987; Towse, Cowan, Hitch, & Horton, 2008; Unsworth & Engle, 2007; Waugh & Norman, 1965). For the present data set we have found it necessary to hypothesize that a fixed WM capacity can be supplemented by another mechanism that is not capacity-limited, in particular the activated portion of LTM (cf. Cowan, 1988; Unsworth & Engle, 2007). We propose that this supplementation should occur when there is sufficient contextual information for LTM retrieval. Given past findings of fixed capacity, the contextual information is presumed insufficient for serial recall (e.g., Chen & Cowan, 2009a). It is also presumed insufficient for recognition when a small number of memoranda are used many times, with items drawn from this small pool on each trial (e.g., Luck & Vogel, 1997; Rouder et al., 2008). In the latter circumstance, we believe that most models of memory would show that contextual cues to retrieval of the items are unusable because there is too much proactive interference; the cues are insufficient to discriminate one trial from the next (e.g., Gillund & Shiffrin, 1984). What is new with the present work is our incorporation of chunking and LTM into a process model for the estimation of a fixed WM capacity in recognition tasks, which we test in the verbal domain.

Chunk decomposition—Third, we have found it necessary also to hypothesize that some allowance must be made for the imperfect use of associative information in chunk formation. For example, if the phrase *leather brief case* is encoded or retained by our participant not as a single chunk as we intended, but as two separate chunks, *leather* and *brief case*, then allowance must be made for this *chunk decomposition* if constant capacity is to be observed.[2]

After presenting the method and data, we compare nine models that incorporate different combinations of WM capacity, LTM retrieval, and chunk decomposition.

Method

In Experiment 1, we presented lists in which the number of words and the number of chunks both varied. The specific conditions, their notation, and the number and types of chunks are shown in Table 1. Exemplifying two types of notation, *2x6* refers to a list with six pairs, whereas *123* refers to a list with a singleton, a pair, and a triplet (in any order). As indicated in the table, many conditions have six words, but differ in the number and size of the presented chunks (e.g., two triplets, three pairs, six singletons, or a mixture of chunk sizes). There also were lists with six larger chunks (six pairs or six triplets), so that effects of chunk size could be observed within six-chunk lists. The purpose of Experiment 2 was to determine whether the modeling mechanisms found for materials within a typical WM task range of up to 6 presented chunks would need to be modified to account for materials well outside of that range, but with the same amount of phonological material (i.e., in place of 6 singletons, pairs, or triplets, lists of 6, 12, or 18 singletons). The *1x2* condition of Experiment 2 provided two chunks as in the *33* condition of Experiment 1.

Experiment 1

Participants—Twenty-six undergraduates (17 female, 9 male, mean age 19.08 years) from an introductory psychology course completed the experiment for course credit.

Materials and procedure—The experiment consisted of three parts: chunk familiarization, articulatory suppression training, and list memory. The words were in Courier New 18-point lower-case typeface.

From a stimulus pool of 36 word triplets (shown in Appendix A), the computer program randomly selected for each participant 12 triplets to be used as complete triplets (e.g. *leather brief case*), 12 triplets for which only the last two words were used (e.g., *game day*, called here pairs), and 12 triplets for which only the last word was used (e.g., *knob*, called here singletons). These triplets, pairs, and singletons were maintained throughout the session for a given participant.

For familiarization of the chunks, participants were instructed that a word-chunk would be shown to them for one second, just above the center of the computer screen. It was followed immediately by a single probe word just below the center of the screen, surrounded by question marks. As soon as the probe word appeared, the participant had 1 s to press a key indicating whether this word was part of the chunk just shown. If not, the probe word was from one of the 11 other chunks of the same size. A 1-s feedback screen indicated correct, incorrect, or too slow (> 1s). This procedure was repeated for 144 trials, within which each word was shown at least 4 times (twice within a chunk at study and twice as a test probe word; among the latter, once as a probe present in the studied chunk and once as a probe absent from the chunk). The number of presentations of some words was often larger and depended on the chunk size, because each position of every chunk was tested exactly twice (once with each kind of probe).

A brief articulatory suppression training session included a computer-presented metronome. This was followed by the list memory task (Figure 2), in which a list of chunks was

[2]Another possibility is that an individual would retain a chunk in memory but not have access to all of its constituent words. For our chunk access measure this would not matter. Moreover, our training method promoted the abstraction of individual words from the multi-word chunks.

presented on each trial. Each chunk was shown on screen in its entirety for one second, without pauses between stimuli, regardless of chunk size. In each of eight blocks of trials, one trial with each of the eight different study list types (see Table 1) was presented, with trials randomly ordered. Thus, there were 64 test trials. As illustrated in Figure 2, the chunks did not have to be presented in the order conforming to the condition name; for example, the *1113* condition could be manifest in a list as 3 – 1 – 1 – 1, or as 1 – 3 – 1 – 1, etc.

During the list study phase of the memory task, participants carried out articulatory suppression by saying the word *the* aloud two times per second, starting at the appearance of a screen 3 s before the first chunk of the list and ending with a screen lasting 1 s after the last chunk (Figure 2). Within the list to be remembered, all the words of one chunk were shown on screen concurrently and in order, for 1 s per chunk without pauses between chunks. In the test phase of each list-recognition trial, all of the words in each chunk were probed in a random order, so that the words within a multi-word chunk were usually not tested adjacent to one another. For each word tested, two words were presented on screen concurrently, one on the left and one on the right, separated by a question mark. The task was to indicate which of these words appeared in the list. The word that was not in the list may have appeared already in the study list of other trials but not the current one; each particular foil word appeared only once per trial. The foil word always came from a chunk of the same length as the correct target word, in order to prevent chunk length from being used as a clue. The task for each test pair was to determine which word was in the current studied list. There was a short break between trial blocks.

Given the use of the same 72 words, forming 36 chunks of 1–3 words within 64 lists, there was a high degree of repetition of the chunks from one list to the next. This amount of repetition was intentional and is common in short-term memory studies, as it creates proactive interference that limits the amount of LTM that can be used in the task.

Experiment 2

This experiment included lists of singletons equal to the length of Experiment 1 lists measured in words (6, 12, or 18 words), and also provided 2-word lists of singletons to determine the performance level when demands were minimal, and for comparison with the 33 condition in Experiment 1.

Participants—Participants were 27 undergraduate students from an introductory psychology course who were not in Experiment 1. Of these, 1 male and 3 females were excluded because of a proportion correct of .51 or below (i.e., near chance) in at least one condition, resulting in a final sample of 8 females and 15 males (mean age 18.83 years). No other participant performed below .55 on any condition. Using this same exclusion rule, no participant had to be omitted from Experiment 1 but the extremely long lists of singletons in Experiment 2 were of course more daunting.

Materials and procedure—The materials and procedure were the same as in Experiment 1 except for the list conditions (noted above). The stimuli were all the final words from the 36 base triplets (e.g., the word *case* from *leather brief case*). In the familiarization phase, each singleton appeared twice to be studied, once as a correct probe and once as a lure. In the list phase, there were still eight blocks of trials as in Experiment 1, but each block included only four trials (one per condition) so the number of trials per condition remained the same as in Experiment 1.

Results

In our results and theoretical analysis, we distinguish between a *presented chunk*, meaning a word or set of words that was intended to be a chunk according to the training regimen and list presentation (i.e., a presented singleton, pair, or triplet), and the unqualified use of the word *chunk* as a mental concept, referring to the unit as it may exist in the participant's memory at the time of testing of part of that presented chunk. Accuracy (proportion of responses correct) provides a simple description of the results, but our modeling-oriented measure is one based on counting chunks accessed, an adjusted chunk score inspired by the access measure of Chen and Cowan (2009a) in recall.

Accuracy

The observed accuracy is provided in Figure 3A (points with error bars). The role of chunking may be observed by considering those conditions with 6 items on the list. Performance is best when these six items are arranged into fewer, larger chunks and decreases as the number of chunks increases (two chunks, *33* condition; three chunks, *222* and *123* conditions; four chunks, *1113* and *1122* conditions; six chunks, *1x6* condition). A one-way, repeated-measures ANOVA for these six-item conditions provides statistical evidence for the effect, $F(5, 125) = 18.08$, $\eta_p^2 = .42$. We evaluate the evidence with the Bayes factor (Jeffreys, 1961).[3] The Bayes factor, denoted B, for this contrast is $B \approx 43$ billion, indicating that the data are about 43 billion times more probable under a model with effects than under the null model without them. The direction of the effect is that there is better performance for lists with fewer, larger chunks.

The accuracy data also reveal an effect of chunk size. A comparison of performance in the *1x6*, *2x6*, and *3x6* conditions reveals that performance declines as the number of items per chunk increases, even though each list contains exactly 6 chunks ($F(2, 50) = 23.15$, $\eta_p^2 = .48$, $B = 1.6 \times 10^{15}$).

A regression of all accuracy scores in Experiment 1 reveals a role for number of presented chunks as well as the number of words that comprise each list. For each participant in that experiment, a mean accuracy was calculated for trials with each particular combination of list length in words and number of presented chunks. Together they accounted for 54% of the variance and both factors were significant. The partial correlation between the number of presented chunks and accuracy was −.46, $t(153) = -6.69$, $B \approx 18.7$ million; between the number of list items and proportion correct, −.39, $t(153) = -5.21$, $B \approx 23,000$. The number of presented chunks uniquely accounted for 13% of the variance in proportion correct, the number of words uniquely accounted for 8%, and the other 33% of the variance accounted for was shared between presented chunks and words.

[3]For this F ratio, $p < .001$. However, we do not use p-values to evaluate the evidence in this report. Along with several authors (e.g., Berger & Berry, 1988; Edwards, Lindman, & Savage, 1963; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Sellke, Bayarri, & Berger, 2001; Wagenmakers, 2007) we believe that inference by p values is flawed because it does not allow the researcher to accept the null as a valid description of the data. Thus, inference by p values overstates the evidence against the null. We provide Bayes factors, which, for the contrasts reported herein, are the probability of the data under the alternative hypothesis that there is an effect, relative to the probability of the data under the null. Specification of the alternative is needed, and we adopt the default priors advocated by Jeffreys (1961) and discussed by Rouder et al. (2009). Expressions we used for evaluating the Bayes factor for t-tests, linear regression, and ANOVA are presented in Rouder, Speckman, Sun, Morey and Iverson (2009), Liang, Paulo, Molina, Clyde, & Berger (2008), and Rouder, Morey, Speckman, and Province (submitted; available at http://pcl.missouri.edu), respectively. Bayes factors may be seen as more conservative than p values in the sense that significant p values often correspond to marginal evidence. Thus, for our 2-tailed t tests comparing the two experiments, for which df=47, a p value of .05 corresponds to a Bayes factor favoring the alternative by 1.23-to-1. Certainly, as the p value decreases the Bayes factor increases; for instance, p values of .01, .001, and .00002 correspond to Bayes factors of 4.55-to-1, 33.33-to-1, and a 1000-to-1, respectively. See Wetzels et al. (2011) for the practical differences between inference by p values and Bayes factor.

Presentation and output position effects—Given the complexity of our task of modeling *how much* is held in WM, we simplify the task and its solution by not attempting to model *which* items are remembered. In particular, we do not model presentation and output position or order effects, even though they figure prominently in various models of performance on immediate memory tasks (for a review see Lewandowsky & Farrell, 2011). The assumption here is that if the capacity-limited component is filled with several specific chunks, this occurs at the expense of other chunks. Nevertheless, we note that position or order effects occurred. There was the typical bow-shaped accuracy function across presentation positions; accuracies on the first and last items presented were typically .05 to .10 above the minimal score among the medial serial positions. Additionally, one might expect that across recognition questions in the response period of a particular trial, each question and response would interfere with memory for the information not yet tested. Output position effects occurred, but were small. For example, in the *3x6* condition of Experiment 1, performance averaged across Response Positions 1–6, 7–12, and 13–18 yielded accuracies of .72, .69, and .67, respectively. Chunk information removed even that small difference: Performance in this same *3x6* condition averaged .69 on the first, second, and third item tested from each triplet. In the 18-singleton condition of Experiment 2, the accuracies for Response Positions 1–6, 7–12, and 13–18 were .66, .64, and .61, respectively, with the same small output position effect magnitude as in Experiment 1. In sum, by collapsing across presentation and output positions in this procedure, we are ignoring some regularity regarding which items are recognized, to focus on capacity limits and what role they play.

Chunk Scoring and the Adjusted Chunk Score

Chunk Scoring—Given the clear importance of chunks according to the accuracy measure, our modeling of the data was based on a data transformation that emphasized the importance of chunks. We did not wish to make the assumption that knowledge of one item from a presented pair or triplet necessarily implied knowledge of all items from that pair or triplet in case participants had only partial information about some presented pairs or triplets.

The measure we used was one in which credit was awarded for correct responding on a presented singleton, and for correct responding on at least one item from a presented pair or triplet. This was termed the *chunk score* because it reflects performance on a related set of words (e.g., *leather brief case*) that were presented together to encourage the formation of a mental chunk including those words. After correcting the chunk scores for guessing (resulting in an adjusted chunk score), we use it to compare models and find that a fixed-capacity model based on the intended, presented chunks is not adequate. A modified model was adequate (Model VIII); it allows a separate LTM contribution, as well as a decomposition, with some probability, of a presented chunk into multiple chunks in the mental representation. We now explain this use of chunk scores.

The lax criterion of just having to know at least one item per presented chunk in order to receive credit for that chunk ensures that participants cannot receive credit for multiple units derived from the same presented chunk (until that assumption is explicitly added later to some of the models in the form of chunk decomposition). For example, if a participant remembered both *leather* and *case* we would not want the participant's score to include these as multiple items in memory. We thus adhere to the strong assumption that they are indeed tied together as a chunk, such that remembering one always implies that the other also was remembered and that these multiple items shared a single slot in WM. When this strong assumption seemed to fail, we then explicitly considered a chunk decomposition factor to

reflect the extent to which the assumption departed from practical reality (partly learned sets, momentary failure to retrieve an association, and so on).

Proportion correct and the chunk score provide complementary views of the data, leading to complementary theoretical insights. The proportion correct brings the effect of chunking into relief. Chunk scores allow an investigation of how far one can get by assuming that WM includes a fixed number of chunks regardless of the nature of the materials. It turns out to be very consistent with the data to say that there is such a fixed-capacity component and that it is important in WM, but inconsistent to believe that it alone can account for the end-product of WM; even with rehearsal prevented, additional complementary factors turn out to play a role.

Adjusted Chunk Score—One limitation of lax scoring in the chunk score measure is that guessing leads to better performance for bigger chunks. For example, if a chunk contains a single item, then it is recognized correctly from guessing with probability of .5; if the chunk contains three items, then any item can be correct by chance and so this rate rises to ($1 - .5^3 = 7/8$). For the *j*th chunk in a list, the guessing base rate is

$$g_j = 1 - .5^{m_j}, \quad (1)$$

where m_j is the number of items in the *j*th chunk (1 =singleton, 2 =pair, 3 =triplet). Let i denote the type of list used on a trial. In Experiment 1 and Experiment 2, there were 8 and 4 types of lists, respectively (see Table 1). The baseline guessing performance for the *i*th list, denoted G_i, is obtained by summing g_j over the chunks in the list: $\sum_{j=1}^{n_i} g_j$. For example, the guessing baseline for the *123* list is the sum .5 + .75 + .875, which evaluates to 2.125. Note that there are no free parameters in computing baseline guessing rates because these rates reflect a constraint in the employed two-alternative forced-choice recognition-memory paradigm.

To address the issue of guessing we define the *adjusted chunk score*, z_i, as the portion of chunk score that is not due to guessing. Let y_i denote the chunk score, and note that y_i reflects both guessing and mnemonic influences:

$$y_i = z_i + \left(1 - \frac{z_i}{n_i}\right) \sum_j g_j. \quad (2)$$

Rearranging terms, the adjusted chunk score is

$$z_i = \frac{y_i - \sum_{j=1}^{n_i} g_j}{1 - \sum_{j=1}^{n_i} g_j / n_i}. \quad (3)$$

This adjusted chunk score thus is derived by subtracting the guessing baseline and rescaling the range so that the expected result based on guessing alone is $z_i = 0$ and the maximum possible value is n_i, the number of presented chunks in the *i*th list. Negative values of z_i are permissible and correspond to below-chance performance, which happened occasionally for a few participants.

The adjusted chunk score is a parameter-free, linear transform of the chunk score measure that allows for a more natural comparison across different list conditions. Trends in the adjusted chunk scores across conditions depend on the processes by which participants answer the recognition questions, and, consequently, these trends may be used to assess processing. If each presented chunk were recognized from a chunk-capacity-limited WM either completely or not at all, and with no additional help from any supplementary source of memory, then the adjusted chunk score for a condition within an individual would depend only on the number of chunks in the list and the individual's chunk capacity; it would not depend on the length of chunks within the list.

The adjusted chunk score for individuals across the list conditions of both experiments is shown in Figure 3B. There are three trends, and these will prove important in constraining competing theories: The first, labeled A, is a flat trend for lists of length 6 with 3 and 4 chunks. Such a flat trend is consistent with a constant-capacity model (e.g., Rouder et al., 2008). The remaining two trends, labeled B and C, show the effects of lengthening the list through lengthening the number of items in a chunk or the number of chunks, respectively. (The data points for 2-chunk lists were omitted from our trends because they are at ceiling, as one would expect based on past literature, e.g., Cowan, 2001.) Trend B shows a decrease across lists *1X6*, *2X6*, and *3X6*, indicating that the adjusted chunk score decreases with increasing presented-chunk size. The evidence for this trend was assessed by regressing adjusted chunk score onto the number of items per chunk within these lists (either 1, 2, or 3). The resulting mean slope is −.51 (SEM= .14, $t(25) = 3.64$). The evidence for the trend comes from the Bayes factor, which is $B = 26.7$, and indicates that the data are substantially more probable under a model with nonzero slope compared to one with a slope of zero. The interpretation of Trend B is that the probability a chunk is held in memory depends to some degree on the chunk size. Trend C is an increase with list length in Experiment 2. The evidence for this trend was assessed by adjusted chunk score onto list length (6, 12, or 18 singletons). The resulting mean slope is .12 (SEM= .03), and there is substantial evidence for an effect ($t(22) = 3.87$, $B = 39.8$). Trend C indicates that longer lists of singletons result in more retention after correction for guessing. The fact that these trends are in opposing directions will not only highlight the limits on a model with fixed capacity for chunks, but will serve to rule out a number of alternatives as well. The trends are difficult to interpret atheoretically but will prove crucial to the selection between models that differ in their inclusion versus exclusion of capacity limits, incompletion of chunking, and long-term memory.

Model Analysis of List Item Recognition

We developed a collection of 9 models to measure the contributions of component processes to recognition memory. The models differ in how memory is conceptualized, either as a unified system, or consisting of WM and LTM components, and in whether items or chunks are processed. The need to examine multiple models became clear when the simple assumptions of the first model, based only on perfect access to a fixed number of presented chunks, did not fit the data well. The models consider two additional types of processes that could contribute to performance: the availability of information in the activated portion of LTM outside of the capacity-limited region of WM (which we refer to as LTM for simplicity), and the decomposition of chunks, resulting in smaller units because of incomplete learning or temporary forgetting of associations. These contributions are considered in various combinations that seemed sensible. Each of the 9 models was fit to individuals' mean adjusted chunk scores per condition by finding the parameter(s) that minimized squared error. All minimization was done through the R statistical package (R Development Core Team, 2009), using the optim call, which implemented Nelder and Mead's (1965) simplex algorithm. Convergence of the algorithm was achieved with default

settings for all model fits to all individuals. To competitively test the models, we used the Akaike Information Criterion (AIC; Akaike, 1974).[4] AIC may be computed in this case by assuming that the observed mean adjusted chunk score is the predicted value plus normally-distributed noise. In this case, the expression for AIC is $m_0 \log(RSS/m_0) + 2m_1$, where m_0 is the sample size (in this case the number of conditions per participant), RSS is the sum-squared error, and m_1 is the number of parameters.

Overview of Models

We assessed 9 models to understand what mechanisms were needed to account for the data. Our model closest to the capacity estimation literature (Model I) involved WM for chunks but it did not capture the Trends B, and C in the data (Figure 3B). The next two models investigated whether it might be possible to obtain at least as good a fit without a capacity limit, with only a single and thus unified memory mechanism, applied either to chunks (Model II) or to individual items (Model III). Given limited success for any of these models, we next introduced a LTM component, with several different scenarios: that both a capacity-limited WM and an unlimited LTM operated on individual items (Model IV), that WM operated on chunks (Model V), or that both WM and LTM operated on chunks (Model VI). These models were considerably more successful, but with some important deviations from the data. In Model VII, therefore, we next introduced a model with WM only, but with the possibility that presented chunks could decompose into multiple encoded chunks (e.g., “leather brief case” encoded as two chunks, “leather” and “brief case”). This model was apt for Experiment 1 but the chunk status of course had no bearing on the fit for long lists of singletons in Experiment 2. Therefore, in Model VIII, we retained the notion of chunk decomposition but also added in the LTM component. This provided an excellent fit to the data. Finally, in Model IX, we found that the unified memory mechanism enhanced with chunk decomposition was still severely inadequate. By far the best fit occurred for Model VIII.

Models and Predictions

Model I. Working-Memory for Chunks—This model is the instantiation of Cowan’s constant-capacity WM model (Cowan, 2001) for a stimulus set with multi-item chunks. The model is based on the notion that an individual has k WM slots and uses each of them to hold one of the n units in the presented stimulus set. When any one unit is probed, the unit is recognized with a probability denoted d, with $d = k/n$ if there are more presented units than slots, and $d = 1$ if there are more slots than presented units. It is convenient to combine these statements with the min function: the probability that a unit is recognized is $d = \min(1, k/n)$ (Rouder et al., 2008). If a unit is not recognized (with probability $1 - d$), guessing takes place at a certain rate. (In the present two-alternative forced choice situation, the guessing rate is .5). Applied to the present test situation, words are bound into identifiable chunks and each chunk occupies a single slot. The number of slots does not vary across conditions. The equations for chunk score and adjusted chunk score, respectively, are

$$y_i = \begin{cases} k + \left(1 - \frac{k}{n_i}\right) \sum_{j=1}^{n_i} g_j, & k < n_i, \\ n_i & k \geq n_i, \end{cases}$$

[4]AIC is more suitable than the Bayesian Information Criterion (BIC, Schwartz, 1978) in this application. Both methods assume independent-and-identically distributed error terms, which is violated in within-subject designs. Because performance is correlated across individuals, there is less independent information than the overall number of observations suggests. This violation of assumptions is more problematic in BIC than AIC because, in the former, the penalty term includes the overall sample size. This sample size does not reflect within-subject correlations, thus overpenalizing models with more parameters.

and

$$z_i = \min(k, n_i),$$

where *k*, the sole free parameter, is the individual's capacity, which is constant across all experimental conditions. The last equation shows the theoretical utility of adjusted chunk score: it is equal to the number of chunks loaded into WM in the fixed-capacity model (and equal to the capacity provided that the number of chunks equals or exceeds capacity).

Figure 4, Panel I shows the mean prediction[5] for adjusted chunk score. As can be seen, although this model can capture the flat Trend A, it cannot capture the decrease with chunk length (Trend B) or the increase with list length (Trend C).

Model II. Unified Memory for Chunks—Although the distinction between a working-memory system and a LTM system is common in memory research (e.g., Cowan, 2005; Unsworth & Engle, 2007), there is a notable school of thought in which memory is conceptualized as a single unified system with common rules or mechanisms across all time frames (e.g., Nairne, 2002). We implement a simple unified memory model for chunks. We assume that there is a fixed probability that a chunk is stored; thus, there is no chunk capacity limit in this model. The resulting equations for chunk score and adjusted chunk score are, respectively,

$$y_i = n_i u + (1-u)\sum_{j=1}^{n_i} g_j,$$
$$z_i = n_i u_i,$$

where *u* is the sole free parameter and denotes the probability that a chunk is stored. Figure 4, Panel II shows the mean prediction for adjusted chunk score. Although this model captures Trend C (albeit poorly), it does not capture Trends A or B. Overall, the model is not successful.

Model III. Unified Memory for Items—A second unified memory model specifies that memory stores the presented items (in this case, words) rather than multi-item chunks. We let *u*, the sole free parameter, denote the probability that an item is stored. The probability that a chunk is recognized is

$$u_j^* = 1-(1-u)^{m_j},$$

i.e., one minus the probability that no item from the presented chunk is held in memory, and the chunk score is

[5]Predictions for all conditions are computed from individual best-fitting parameter estimates that are then averaged in each condition. This approach is preferred for nonlinear models compared to either fitting the model to averaged data, or computing predictions from averaged parameter values.

$$y_i = \sum_j (u_j^* + (1 - u_j^*) g_j)$$

The adjusted chunk score may be computed by transforming the chunk score, but the result does not reduce to an equation that provides additional insight. Figure 4, Panel III shows the mean prediction for adjusted chunk score, and this model is unable to capture Trend B.

We draw no inference as to whether some other unitary model could explain our data but, within the simple modeling framework we have set up for the present purposes, in which we view memory as the potential combination of a limited-number-correct mechanism and a limited-proportion-correct mechanism, the models already considered suggest that the limited-number mechanism (i.e., the limited-capacity mechanism) greatly improves the fit compared to unitary models that do not include such a limit. This conclusion will be strengthened by the remaining models that we consider.

Model IV. WM+LTM for Items—We considered a model in which a capacity-limited WM and LTM contribute to recognition memory. Of course, all of the items are in LTM in some form but what is meant here is that the item being tested is recognized sometimes when it is not in a capacity-limited WM because its LTM representation is in an active state. There is no capacity limit on activated LTM so it is assumed to include a fixed proportion of the items, not a fixed number like WM.

In this model instantiation, both WM and LTM code items rather than chunks. This model had to be considered given that items were tested individually, apart from the chunk in which they had been learned originally or presented during the study phase of each trial. Let $N_i = \Sigma_i m_j$ be the total number of items on a list. The probability that an item is in WM is $w = \min(1, k/N_i)$, and the probability that an item is in either memory system is $p_m = w + (1 - w)\ell$, where ℓ is the probability that an item is in LTM. Consequently, the probability that at least one item in the jth chunk is held in memory is $p_{cj} = 1 - (1 - p_m)^{mj}$, and the chunk score is

$$y_i = \sum_{j=1}^{n_i} (p_{c_j} + (1 - p_{c_j}) g_j).$$

The model has two free parameters: capacity k and LTM probability ℓ. The adjusted chunk score is computed by transforming these chunk scores. Figure 4, Panel IV shows the mean prediction for adjusted chunk score, and this model is unable to capture Trend B.

Model V. WM for Chunks + LTM for Items—One of the advantages of the LTM models is that they account for the increase in adjusted chunk score with longer lists (Trend C). In this model, we combine a capacity-limited WM model with activated LTM for items. The chunk score is given by

$$y_i = \begin{cases} k + \left(1 - \frac{k}{n_i}\right) \sum_{j=1}^{n_i} [L_j + (1 - L_j) g_j], & k < n_i, \\ n_i, & k \geq n_i, \end{cases}$$

where L_j is the probability that at least one item from Chunk j is held in activated LTM; $L_j = 1 - (1 - \ell)^{m_j}$. Free parameters are capacity k and probability that an item is held in activated LTM, ℓ. The adjusted chunk score is computed by transforming these chunk scores.

Figure 4, Panel V shows the mean prediction for adjusted chunk score. This model is an improvement over the WM model for chunks as it captures the increase with list length in Experiment 2 (Figure 3, Trend C). Nonetheless, it makes a prediction opposite to Trend B, the decrease with chunk length. Because items are remembered without capacity limits, the model predicts incorrectly that chunks with more items should yield increased adjusted chunk scores compared to lists with fewer items.

Model VI. WM+LTM for Chunks—This model is the same as the previous with the exception that chunks rather than items are held in activated LTM, as is the case for WM as well. Let ℓ denote the probability that a chunk is held in activated LTM. The chunk score and adjusted chunk score are respectively given by

$$y_i = \begin{cases} k + (n_i - k)\ell + \left(1 - \frac{k}{n_i}\right)(1-\ell)\sum_j g_j, & k \le n_i \\ n_i & k > n_i, \end{cases}$$

and

$$z_i = \begin{cases} k + (n_i - k)\ell, & k \le n_i, \\ n_i, & k > n_i. \end{cases}$$

The free parameters are k and ℓ. Figure 4, Panel VI shows the mean prediction for adjusted chunk score. The model accounts for the increase with list length in Experiment 2 (Trend C), but does not account for the decrease with chunk length (Trend B).

Model VII. WM with chunk decomposition—None of the previous 6 models account for Trend B, the decrease in chunk score with increasing chunk size. It is important to place this decrease in context. Certainly, chunking is a salient feature of the data, but the decrease implies that it is not complete or perfect. Larger chunks may not be formed during the encoding process (e.g., perhaps *leather* is not associated with *brief case*) or, if they are formed, the association may not be retrieved at the time of test. To account for the decline in the adjusted chunk score in Trend B, we constructed more flexible models in which associations between items can be unavailable at the time of test (regardless of the reason for the unavailability). When associations are unavailable, effectively the number of memoranda increases.

The present hypothesis of chunk decomposition differs from the fragmentation hypothesis of Jones (1976). According to his hypothesis, the memory representation of a stimulus is a single fragment that includes some or all of the features of the presented object and can be cued by any remembered feature. Applied to our study, the fragmentation hypothesis is a reasonable interpretation of our model without chunk decomposition. In contrast, in our notion of chunk decomposition, the presented chunk gives rise to two or three mental chunks that can be remembered only by occupying two or three slots in working memory.

We assumed that a two-item chunk decomposes to two singletons (meaning that the association between the words either was not formed, or was formed but not retrieved) with

probability *p*. For simplicity, we also assumed that in a three-item chunk in this study includes only 2 associations (between the first and second items, and between the second and third items) and that each of them decomposes with independent and identically distributed probability *p*. Hence, the probability that the chunk decomposes to three singletons is p^2; the probability that it decomposes to a singleton and a pair (which can happen in 2 ways) is $2p(1-p)$; and the probability that it remains intact is $(1-p)^2$. To understand the effect of chunk decomposition, consider the *1113* list type. If the last chunk remains intact, then there are 4 memoranda; if the chunk decomposes to a singleton and a pair, there are 5 memoranda, and if the chunk decomposes completely to three singletons, there are 6 memoranda. The probability of these three events are $(1-p)^2$, $2p(1-p)$, and p^2, respectively.

Let n' denote the number of unobserved or latent chunks available taking into account decomposition, which, naturally, must be no less than the number of presented chunks, i.e., $n' \geq n$. Let $q(n', p)$ be the function that denotes the probability of n' memoranda for failure-of-association probability *p*. For the *1113* list,

$$q(n', p) = \begin{cases} (1-p)^2, & n'=4, \\ 2p(1-p), & n'=5, \\ p^2, & n'=6. \end{cases}$$

With this function, a WM model with chunk decomposition yields a chunk score for the *1113* condition given by

$$y_{1113} = \sum_{n'=4}^{6} q(n', p) \times \min\left(n', \left[k + \left(1 - \frac{k}{n'}\right)\sum_{j=1}^{n'} g_j\right]\right)$$

The calculation of chunk score for a list is the weighted sum (or expected value) over all possible failure-of-association events. For example, in the *3X6* list, each chunk may be in three possible decomposition states (intact, partially, fully decomposed) yielding a total of 729 states. These 729 states map into 13 values of n', from 6 to 18, and into 13 corresponding values of *q*. Chunks are assumed to decompose independently, so, for example, the probability that the first chunk decomposes to 3 singletons, the second decomposes to a pair and a singleton, and the remaining four chunks do not decompose at all is found by multiplication, $p^2 \times p(1-p) \times (1-p)^4$. To express the model predictions for the *i*th list, let $\mathcal{N}'_i$ be the collection of memoranda sizes which vary across trials due to chunk decomposition, and let $q_i(n', p)$ be the probability that chunks decompose into lists of length n' for the *i*th list type and for association-decomposition probability *p*. Then,

$$y_i = \sum_{n' \in \mathcal{N}'_i} q_i(n', p) \times \min\left(n', \left[k + \left(1 - \frac{k}{n'}\right)\sum_{j=1}^{n'} g_j\right]\right). \quad (4)$$

The free parameters are capacity *k* and probability parameter *p*. Adjusted chunk score is given by transforming chunk score according to (3). Figure 4, Panel VII shows the mean prediction for adjusted chunk score. This model accounts for the difficult Trend B, the decrease in adjusted chunk score with chunk length. It does not, however, account for Trend C, the increase with list length in Experiment 2.

Model VIII. WM with chunk decomposition and LTM for chunks—The preceding development and resulting model fits highlight two seemingly necessary elements, at least within a capacity-based WM model. First, there is evidence for a small contribution of a capacity-unlimited memory, as seen by the increase in adjusted chunk score with list length in Experiment 2. This trend is accounted for by adding an activated-LTM component to the model account. Second, there is evidence for an opposing trend in which the adjusted chunk score decreases if lists are increased in lengths by increasing the length of the constituent chunks. This trend is accounted for by adding chunk decomposition. In this model, both LTM and chunk decomposition are implemented simultaneously. The resulting equation for chunk score is

$$y_i = \sum_{n' \in \mathcal{N}'_i} q_i(n', p) \times \min\left(n', \left[k + (n' - k)\ell + \left(1 - \frac{k}{n'}\right)(1-\ell)\sum_j g_j\right]\right). \quad (5)$$

The free parameters are capacity k, LTM probability of chunk recognition ℓ, and decomposition probability p. However, the ability to accurately locate p and ℓ is leveraged by Experiments 1 and 2, respectively and exclusively. This reliance is especially clear for Experiment 2, where all presented chunks are singletons, and, consequently, chunks do not decompose. Although it is impossible to estimate p for participants in Experiment 2, this inability is of no concern because for these lists, the value of p does not govern performance. The more salient concern is with Experiment 1, where all three parameters govern performance. The main problem here is an instability in which parameters p and ℓ trade off within a participant, sometimes wildly. To stabilize estimation, we set ℓ to the average value obtained in Experiment 2, $\ell = .125$. Had the experiments been conducted on the same participants, a procedure that would have introduced several other concerns, this fixing of ℓ would be unnecessary. This estimation procedure, despite being non-optimal given individual variation in LTM, proved sufficient for good fits. Figure 4, Panel VIII shows the mean prediction for adjusted chunk score, and the account is highly accurate.

For further assurance that the best-fitting model, Model VIII, provided a good description of the data, we fit it to the accuracy data. Let $d = \min(1, k/n')$, where n' is the number of chunks in mind following any decomposition of presented chunks. Then the accuracy for a condition, a_i is

$$a_i = \sum_{n' \in \mathcal{N}'_i} q_i(n', p)\,[d + (1-d)\,(\ell + .5(1-\ell))]. \quad (6)$$

This model was fit to the data by minimizing least-square error subject to the constraint that $\ell = .125$ as before. Figure 4 shows the predictions of this model, which fit the data very well. Based on the parameter estimates from the accuracy fits, we then generated adjusted chunk score predictions, which appear as the dashed line in Figure 4, Panel VIII. As shown, these scores were not far off from the model based directly on adjusted chunk scores. Finally, in Figure 3A, we show the pattern of accuracy data and plot the predictions of Model VIII, in order to show what the pattern looks like in these more familiar terms, both with parameter values obtained from the fit to adjusted chunk scores and slightly different parameter values obtained from the fit to accuracy. (The fact that the fits are slightly different is to be expected, given that the accuracy data do not allow a clear distinction between remembered items within and between chunks.)

Model IX. Unified memory for chunks With chunk decomposition—Models VII and VIII introduce a novel notion of incomplete chunks. This notion may be applied to models without capacity limits as well. We generalized Model II, the unified memory for chunks, by adding the same incomplete-chunk process as in Models VII and VIII. The chunk score is

$$y_i = \sum_{n' \in \mathcal{N}'_i} q_i(n', p) \times \left(n'_i u + (1-u) \sum_{j=1}^{n'_i} g_j \right). \quad (7)$$

Figure 4, Panel IX shows the mean prediction for adjusted chunk score, and the account still misses Trend B.

Parameter Estimates and Fit

Mean parameter values for all nine models are shown in Table 2. Estimates typically vary across the two experiments, and this makes some sense as they involve different people. Our main concern is the difference in capacity, which for Model VIII was estimated higher in Experiment 1 than in Experiment 2 (3.4 vs. 2.8, $t(47) = 1.73$, $B = 1.27$ in favor of the null). The Bayes factor indicates that the evidence for an effect of experiment is equivocal and that this difference should not be over-interpreted. In both experiments, the six-singleton condition was included and performance in that condition was different across experiments, commensurate with the capacity estimates. Thus, separate regressions for the two experiments comparing each individual's adjusted chunk score in the six-singleton condition with his or her capacity estimate yielded similar intercepts (3.42 and 3.58) at the point of overall mean capacity, and similar slopes (.81 and .56).

Table 2 also shows that the AIC value was best for Model VIII, the full model. Other values are reported relative to that value. The comparison of Model VIII to Model IX indicates that a WM component with a fixed capacity is necessary to obtain good fits, reinforcing the notion that we have provided a bridge between simpler uses of capacity estimates for change-detection procedures (e.g., Cowan, 2001; Rouder et al., 2008) and the much more complicated and varied recognition test situations used here.

Inspection of Figure 4 provides insight into what aspects of the model were needed for what aspects of the data. In Experiment 1, Models VII and VIII conform closely to the data. In Experiment 2, it is Models IV, V, VI, and VIII that conform to the data. What the models suiting Experiment 1 have in common is the possibility of incomplete chunks, whereas what the models suiting Experiment 2 have in common is the separate LTM factor. What all of these models have in common is the inclusion of a fixed WM capacity along with one other factor (unlike the single-factor models, I-III and IX). Given that the addition of chunk decomposition to a unified memory model did not improve the model, our data implicate a WM capacity limit, but one that must be supplemented in different ways depending on the nature of the stimuli. When the list can be coded as a limited number of chunks but not perfectly, then chunking decomposition must be considered; when there are long lists of singletons, an LTM storage factor must be considered. Conceptually, the incomplete nature of chunks that have been formed is quite related to the proportion of singletons stored in LTM but, in our modeling framework, these must be separately instantiated. In future work, we expect that it might well be possible to show a relation between chunk completion and LTM storage, in principle reducing the number of factors necessary. In short, then, we conclude that it is feasible to extend the widely-used capacity model of change detection (Cowan, 2001; Rouder et al., 2008) to complex verbal lists.

In summary, the main evidence supporting our conclusion includes the fact that (1) within-subject confidence intervals shown in Figure 4 overlap the predictions for all conditions in both experiments only in the case Model VIII; (2) Trends A, B, and C in adjusted chunk scores across conditions (Figure 3B) all go in the right direction only for this model; and (3) AIC values, shown in Table 2, are lowest for this model.

General Discussion

The type of formal model described by Cowan (2001) and refined by Rouder et al. (2008) has been widely used to estimate the number of items held in WM, and to understand WM capacity limits, in visual array probe-recognition tasks (for a review and tutorial see Rouder, Morey, Morey, & Cowan, 2011). This type of model, however, has been confined to situations in which one can assume massive proactive interference, reducing the usefulness of LTM, and little likelihood of combining items to form larger chunks. The present work extends this type of model to verbal lists, and to situations in which the same restrictive assumptions do not hold. Capacity limits may be observed cleanly only by disentangling these limits from other mnemonic processes, just as the law of gravity produces results that are observed cleanly only by disentangling gravity from such factors as air resistance and friction.

We found that constant capacity holds with two exceptions. First, as the lists were made longer by increasing the chunk length (Experiment 1), there was poorer performance than predicted by constant capacity (Trend B in Figure 3). Second, as lists were made longer by increasing the number of singletons (Experiment 2), there was better performance than predicted by constant capacity (Trend C in Figure 3). These trends are humbling because they clearly shows that the model of constant capacity as described by Cowan (2001), and well-confirmed by others for arrays of visual objects (D. Anderson et al., 2011; Rouder et al., 2008; Zhang & Luck, 2008), needs modification for lists of verbal chunks. The trends were fit well, though, by the addition to the model of (1) the partial decomposition of chunks, and (2) a form of LTM storage of list information (not just pre-stored chunk information).

It was already acknowledged that the acquisition and use of multi-word chunks could well be partial rather than total (Chen & Cowan, 2005; Cowan et al., 2004) and that the observation of a fixed chunk capacity limit should require articulatory suppression to prevent covert verbal rehearsal (Chen & Cowan, 2009a), as in the present data. Until now, however, it appeared that the number of items remembered might be accounted for solely by a chunk-capacity-limited WM, as in the array recognition procedures and the recall procedure of Chen and Cowan (2009a), in which the only known role of LTM was to define the chunks to be recalled. The present work shows an additional role of information in the activated portion of LTM for a recognition procedure, and shows that a coherent model incorporating this component can be formulated.

Beyond the specific models tested, there are general lessons for what must be measured and taken into account in WM studies. According to the approach of Miller (1956), what is important is the length of a set of memoranda in chunks, and the appropriate measure is the number of chunks kept in memory. According to a simple view of memory in which all items are encoded and then retrieved with some probability, chunks may be important, but the items in memory should reflect a proportion of the chunks presented, not a fixed number of chunks. The present results and models suggest that one must consider both the number of chunks held in mind as one basis of recognition, and the proportion of chunks held in mind as another basis. In this sense, the results help to confirm the approach sketched out by Waugh and Norman (1965), which included WM and LTM processes in recall of the same

list. The role of those processes is extended here, to various multi-word chunks and list lengths and to capacity estimation.

In terms of the model of Cowan (1988, 1999, 2001), the focus of attention is limited to a fixed number of chunks, whereas the activated portion of long-term memory is not so limited, but is restricted by decay and interference factors. In both the original model and the present Model VIII, the parameter *k* reflects the capacity of the focus of attention but, in the present Model VIII, performance depends not only on *k* but also on the strength of an LTM contribution. The latter is presumably an activated portion of LTM, given that the chunks have all been made familiar but only a small subset of those are in an active state at any one time and therefore can help guide the response.

Relation to Other Theories of Working Memory Capacity

We make no claim to have ruled out all other potential competing models. What is key, however, is that no framework other than the one we adopt has been designed with the primary aim of estimating the number of units in WM as derived from recognition procedures, taking into account guessing processes. Nevertheless, many alternative theories at least have implications for capacity predictions. We will briefly discuss a few of them in order to describe the relation of our work to the broader theoretical landscape. We address multi-component models, the time-based resource sharing model, other resource models, a material-specific interference model, a general-plus-specific interference model, a two-store analysis of WM tasks, the general class of global memory models, and a general information processing theory. These treatments have provided perspectives on many other models, and each poses a different challenge to the present approach.

Multi-component working memory models—The traditional model in which there are separate verbal-phonological and visual-spatial buffers controlled by a central executive (Baddeley, 1986; Baddeley & Logie, 1999; Hulme & Tordoff, 1989; van der Meulen, Logie, & Della Sala, 2009; Schweickert et al., 1990) appears designed to handle certain kinds of memoranda: ordered sequences of verbal items, or spatial arrays of visual items. The emphasis in these models is on capacity limits measured not in terms of items, but in terms of the maximum time that it could take to rehearse or refresh the item set in memory. Chen and Cowan (2009a) showed that the approach works well to understand performance for serial recall of lists that are not too long, but that a chunk-capacity-limited approach works better in accounting for a wider range of list lengths, provided that the emphasis is on item information rather than order information and rehearsal is prevented. Then recall was limited to about 3 chunks in young adults.

Baddeley (2000) revised his model to include an episodic buffer, and Baddeley (2001) suggested that this buffer could be the locus of the limit of about 3 to 4 chunks in memory. The episodic buffer's purposes appear to be to retain, temporarily in an accessible form, semantic information and associations between items or their features regardless of their input modalities. As such, it is consistent with a chunk-capacity hypothesis.

Given that it is possible to derive both verbal and visual codes for the same materials, such as for kanji characters among Japanese participants (Saito, Logie, Morito, & Law, 2008), one can try to explain interference effects on the basis of recoding visual items to verbal form or vice versa, without any kind of central capacity-limited store. We believe, though, that capacity-limited WM is a needed concept because WM for information in one modality or code suffers interference from storage or processing of information in another modality or code, even when recoding is rigorously prevented. Tradeoffs occur between memory for color-location bindings on one hand and for digit-voice bindings on the other (Saults & Cowan, 2007); between colors and tones (Morey et al., 2011); between color array memory

and tone discrimination (Stevanovski & Jolicoeur, 2007); between novel character array memory and long-term verbal retrieval (Ricker, Cowan, & Morey, 2010); and between verbal digit list memory and a speeded spatial location task (Chen & Cowan, 2009b). (For other examples, see Cowan & Morey, 2007; Morey & Cowan, 2004; Vergauwe, Barrouillet, & Camos, 2009, 2010).

Time-based resource-sharing model—Barrouillet, Portrat, and Camos (2011) developed a model that has some similarities to the multicomponent model (in terms of the notion of a circulating refreshment of memory), but includes more central processing and attention dependence instead of multiple components. In it, the amount that can be recalled in complex WM span tasks is shown to be an inverse linear relation to the proportion of time between memoranda that is taken up with distraction, rather than being free for mnemonic activity (i.e., the *cognitive load*). Precise measurements have been made to show that this time relation depends not on the presentation times of distractors, but on the times used in processing them (e.g., Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007). The theory states that items decay from WM unless they are reactivated in time (Portrat, Barrouillet, & Camos, 2008), which can occur either through an attention-related process known as refreshing or through covert verbal rehearsal (Camos, Mora, & Oberauer, 2011; on refreshing see Raye, Johnson, Mitchell, Greene, & Johnson, 2007).[6]

The relation between the theory of Barrouillet et al. (2011) and the present work is an interesting one that has not yet been resolved. Barrouillet and his colleagues originally adopted a version of their model in which only a single item is held in the focus of attention, and the capacity limit arises from the speed at which items that have left the focus of attention can be refreshed (re-entered into the focus of attention momentarily) before they decay. According to that model, our capacity limit could be the result of the refreshing rate; assuming equal decay rates, individuals with faster refreshing rates would have higher capacities. That is not, however, a necessary interpretation, as Barrouillet et al. (2011) acknowledged. An alternative possibility is that multiple items are held in focus at once and that when there are more items to retain than the limited capacity will allow, extra ones can be retained through a process of rotating the focus among the items. Refreshment then would serve the purpose of re-entering items into the capacity-limited region. Gilchrist and Cowan (2011) showed that this kind of process involving a multi-item focus also can produce Barrouillet's inverse linear relation between cognitive load and WM span. Distraction might also cause one or more of the items in focus to be replaced with a distractor. The issue of mechanisms of capacity limits might be resolved in future work in which both articulatory suppression and an attention-demanding distractor are used, alone and in combination (cf. Oberauer & Lewandowsky, 2008), to determine whether chunk capacity depends on some sort of reactivation of information. No matter which interpretation is correct, the theory of Barrouillet et al. (2011) is compatible with the present work but largely orthogonal to it. Here we make no claims about the origin of the chunk capacity limit of WM (i.e., whether a speed or a space metaphor is more appropriate), only about its approximate size, its implementation in recognition tasks using items drawn from chunks, and its relation to LTM.

Resource models—We do not test an alternative to discrete slots or chunks, a continuously divisible resource (e.g., Bays & Husain, 2008, 2009; Gorgoraptis, Catalao, Bays, & Husain, 2011; Just & Carpenter, 1992; Wilken & Ma, 2004). We believe that some of the evidence in favor of the resources approach is flawed, and this point has been

[6]Some authors disagree about the existence of decay as a mechanism of forgetting (Lewandowsky & Oberauer, 2009; Lewandowsky, Oberauer, & Brown, 2009) but that issue is not germane to the present theoretical treatment and is not pursued further.

addressed elsewhere (D. Anderson et al., 2011; Cowan & Rouder, 2009; Thiele, Pratte, & Rouder, 2011). The debate cannot be definitively settled here. Given that slots are more restricted, though, there is value in showing what must be added to make the discrete capacity model sufficient for complex stimulus situations, such as the lists of variable length and mixed chunk size of the present Experiments 1 and 2.

It is worth noting that the line between slot models and resource models was blurred within the theory of Zhang and Luck (2008). They proposed that, when the available slots outnumber the items to be retained in WM, two or more slots can be assigned to a given item concurrently, increasing the precision of its WM representation. It remains to be seen whether this is the correct interpretation but it brings up the important point that there are intermediate theories between fixed slots and fluid resources. For example, it is possible that attention is a fluid resource that can be distributed unevenly to multiple items, but to no more than k items in all. Also, the resource might be distributed to an entire field of items (Chong & Treisman, 2005), but perhaps with some limited k abstract characteristics of that field that can be extracted and retained at once. In sum, there is a friendly bridge between slot models and resource models as there are characteristics of both types of model that might be justifiable in an updated theory of WM.

Oberauer and Kliegl's interference-based approach—This model is taken to represent the premise that no fixed-capacity system is needed to account for memory results. Oberauer and Kliegl (2001, 2006) compared hypotheses based on a resource that is shared between items (e.g., J.R. Anderson, 1983; Daily, Lovett, & Reder, 2001; Just & Carpenter, 1992), limited slots in working memory (Cowan, 2001; Miller, 1956), decay and time-limited reactivation of memory (Baddeley, 1986; Salthouse, 1996), and interference from cross-talk between items (e.g., Brown, Neath, & Chater, 2007; Nairne, 1990; Tehan & Humphreys, 1995; Oberauer and Kliegl's own model).

Given the systematic nature of this comparison among models, it seems particularly useful to comment on this comparison as it pertains to capacity models. In a key experiment by Oberauer and Kliegl (2001), the participant began each series of trials with a set of digits in n specific box locations; $1 \leq n \leq 6$. Eight updating arithmetic operations were presented in a limited amount of time (e.g., add 2 to the digit found in the second box) and the task was to report each updated digit in turn, remember the updated set, and finally report the updated set in the order in which the items were queried. Clearly some speed-dependent processes were needed to explain how performance is limited when the presentation time is limited, but we will consider that speed limit outside of the scope of the present treatment, and will discuss asymptotic performance when the presentation time is sufficiently long. The resulting asymptotic functions for young and older adults decreased across set sizes, slowly at first and at an accelerating rate at larger set sizes. The issue is how many items can be updated and held when time is sufficient for maximal encoding.

Among the models that were rejected was a limited-capacity model that worked in a manner essentially the same as a resource-limitation model. Oberaur and Kliegl (2001) show very poor fits for this model, but we suspect that the poor fit was the result of ancillary assumptions about the relationship between activation and performance rather than the core construct of capacity. Indeed, the young adults examined by Oberauer and Kliegl (2001) show near-ceiling-level performance up to 4 items and then a steep drop with further increases in memory load, similar to the benchmark array recognition data of Luck and Vogel (1997); the older adults show a function that would be expected with lower capacities in some individuals and not others.

Oberauer and Kliegl (2006) presented a model of capacity limits based on the difficulty in discriminating among memory traces, claiming (p. 606) that \The central assumption of the interference model is that items in working memory interfere with each other through interactions of their features." We believe that this kind of approach falls short because of the tradeoffs that occur between information in different codes and modalities, reviewed above under multicomponent models. The existence of specific interference between items presented together in a working memory task seems likely as an additional constraining factor, but not the defining factor of capacity as Oberauer and Kliegl (2006) suggest.

The existence of capacity independent of specific-interference factors is seen in Awh, Barton, and Vogel (2007, Experiment 2). The performance on arrays of mixed sets, which included both cubes of different orientations and Chinese characters, was compared to performance on arrays of colored squares. When one examines only mixed-set trials in which the change was from an object in one category to an object in another category, the level of performance was almost exactly the same as for colored squares. If the overlap between items in the set was the basis of capacity limits, detection of the cross-category changes should have greatly exceeded detection of color changes, which was not the case. Yet, performance was much poorer for within-category changes of cube orientations or between one Chinese character and another. This suggests that the overlap between items is a factor that can restrict performance within a capacity limit, but does not actually define the maximal capacity limit.

Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, and Usher (2005) general-plus-specific interference model—In the model of Davelaar et al. (2005), a kind of general interference is used as a mechanism to create the capacity limit. This is interference with the activation of any item in the WM buffer, caused by the entry of another item, regardless of their similarity to one another. Given that we do not investigate the basis of the capacity limit, this model is not necessarily incompatible with our approach but serves as a fine-grained analysis that has not yet been applied to the broad range of stimulus conditions we used to test our model. We present it to illustrate the complementary nature of different levels of analysis and modeling.

Davelaar et al. (2005) provided one possible instantiation of a general model that includes both capacity limits and material-specific interference within those limits. Their model was designed to account for lexical-semantic item information in free and cued recall; it was suggested that working memory for serial order information involves additional processes such as phonological encoding and rehearsal. They proposed that the there are four sources of activation of an item: (1) self-excitatory activation that leads to persistence of the information after stimulus offset, (2) inhibition between different lexical-semantic units, \causing them to compete for activation and resulting in displacement from the buffer when too many units are active at the same time (i.e., when the capacity of the system is surpassed) (p. 8)," (3) weak connections to semantic associates, and (4) input from the contextual representation. In most of the simulations, however, the stimuli were lists of semantically unrelated words and the parameter for semantic associates was set to zero.

Because the main source of interference between items was independent of the nature of the items, this formulation is consistent with a slots model. When there are semantic associates that serve as competitors, the effective capacity does not change but the interfering competition results in an increased possibility that the capacity will not be filled with as many of the correct items, as opposed to incorrect associates. Thus, capacity limits exist and similarity interference can degrade performance within those limits, in keeping with our interpretation of the data from Oberauer and Kliegl (2001, 2006).

Unsworth and Engle (2007) Dual-store model—This model is related to our approach as applied to a very different set of procedures. Unsworth and Engle (2007) suggested that individual differences in standard WM tasks stem from two sources in combination: differences in the ability to maintain items in primary memory, and differences in the ability to search for items in secondary memory in a controlled manner. If primary memory is essentially a matter of keeping items in a capacity-limited store in a way that demands attention, it makes sense for individuals with better attentional capabilities to excel at both primary and secondary memory functions. In order to observe capacity limits clearly, it is therefore necessary to avoid tasks that elicit extensive primary and secondary memory components together. In running memory span with digits, for example (e.g., Cowan et al., 2005), it is assumed that the buildup of proactive interference from digits used over and over soon prevents the use of secondary memory to carry out the task. The same applies to an instantaneous array of random visual items (Luck & Vogel, 1997). When a list of semantically unrelated words is used on each trial at a fairly rapid presentation rate, there is no clear pattern that would assist secondary memory search and, if rehearsal also is prevented, a capacity-limited mechanism becomes the key storage mechanism. Thus, the present work helps to show that the general framework suggested by Unsworth and Engle can be usefully implemented as a mathematical model (elaborating on the earlier framework by Cowan, 1988), within which the contributions of WM capacity and LTM can be distinguished.

On the other hand, because articulatory suppression to prevent phonological rehearsal was not used by Unsworth and Engle (2007), it would be unwise to make too much of the similarity in capacity estimates obtained here and by Unsworth and Engle. Further research is needed to find a common theoretical framework that applies across the different test situations used by them and in our study.

Global memory models—There is a large class of models termed global memory models, examples of which include MINERVA (Hintzman, 1988), TODAM (Murdock, 1993), CHARM (Eich, 1982), SAM (Gillund & Shiffrin, 1984) and REM (Shiffrin & Steyvers, 1997). (For reviews of the models and their properties see Clark & Gronlund, 1996; Murnane, Phelps, & Malmberg, 1999). One common paradigm used to assess global memory models is yes/no recognition memory, in which the participant indicates whether a probe was studied or is novel. The defining feature of this class is that each item in memory is activated to some degree by a probe, and the total activation across all items supports subsequent old/new decisions. The more items in the memory set, the more diluted will be the proportion difference between the amounts of activation resulting from target-present versus target-absent sets. Thus, global memory models most likely can account for list length effects on performance. Our task, however, discourages decisions based on global activation. We provide two probes and ask participants to identify the studied one (two-alternative forced choice). Consequently, the model should base decisions on the activation of two items without recourse to the whole of memory. The development of global memory models is largely orthogonal to the development here as the distinctive elements comprising global activation are not applicable. Global memory models have multiple parameters and detailed specification designed to capture phenomena that, by design, are not present in our studies. In order to account for our finding that there must also be an LTM contribution that is not capacity-limited, a different parameter is needed, and many of the models include such parameters (e.g., Gillund & Shiffrin, 1984). It would be a difficult but important task to determine how these models could account for our results and estimate WM capacity.

ACT-R—As perhaps the best example of a general information processing framework, the current ACT-R model (J.R. Anderson, 2007; Borst, Taatgen, & Van Rijn, 2010) includes an imaginal or problem-state module that holds a\chunk" with 3 or 4 slots, within a larger

system in which the recognition of list items is based on temporary memory activation. We believe that Anderson's slots could be equivalent to our chunks except that in our conception, the chunks need not contribute to any single, well-defined problem. It is noteworthy that the parietal region that Anderson designated for the imaginal module is quite similar to the left intraparietal sulcus region that Cowan, Li et al. (2011) found as the area that responds to an increase in memory load regardless of the input domain (spatial or verbal).

Conclusion

No concept seems more important to the foundation of cognitive psychology than limited capacity, but this concept has been elusive. Capacity clearly governs performance in simple situations designed to eliminate or control for complicating factors (e.g., D. Anderson et al., 2011; Chen & Cowan, 2009a; Rouder et al., 2008; Zhang & Luck, 2008). We have taken a step toward the more difficult issue of the application of capacity limits in complex situations in which supplementary processes are used, with encouraging results.

The extent to which LTM contributes to immediate memory appears to depend on the measure. Chen and Cowan (2009a) did not find a role of LTM in recall of lists up to 12 items forming up to 6 chunks. Provided that participants engaged in articulatory suppression and the serial order of recall was ignored, Chen and Cowan found a good fit by a constant-capacity store. We, in contrast, have explored how well a WM capacity model might hold up under articulatory suppression when the lists varied widely in phonological length, the number and size of chunks varied widely, sub-chunk items were tested in random order, chunks to be recognized could break apart or be imperfectly retained in WM, and there were up to 18 singletons in the list. The chunk capacity concept withstood these assaults extremely well, though the need to take into account both chunk decomposition and LTM also was striking.

On one hand, fixed storage capacity, which has been one of the most fundamental concepts at the heart of cognitive psychology at least since Miller (1956), is shown here to be what many modern researchers take it to be: an impossibly simple concept. On the other hand, the concept is redeemed here in that a reasonable, straightforward set of ameliorating factors was shown to be sufficient to make the concept work, accounting for a much broader range of evidence than would be possible without the capacity concept.[7]

Acknowledgments

This work was carried out with assistance from NIH Grant R01-HD21338 and NSF Grant SES-1024080. We thank Pierre Barrouillet, Alice Healy, Stephan Lewandowsky, and two anonymous reviewers for helpful comments.

References

Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; 19:716–723.

Anderson D, Vogel E, Awh E. Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. Journal of Neuroscience. 2011; 31:1128–1138. [PubMed: 21248137]

Anderson, JR. The architecture of cognition. Cambridge, MA: Harvard University Press; 1983.

[7]In the visual domain, chunks composed of multiple visual items can be created (Jiang, Chun, & Olson, 2004), but models of capacity in that case have not been investigated to our knowledge. Our research opens up the possibility of applying capacity-estimation formulae to a much wider variety of real-world visual information that includes multi-item chunks and LTM information.

Anderson, JR. How can the human mind occur in the physical universe?. Oxford University Press; 2007.

Awh E, Barton B, Vogel EK. Visual working memory represents a fixed number of items regardless of complexity. Psychological Science. 2007; 18:622–628. [PubMed: 17614871]

Baddeley, A. Working memory. Oxford, England: Clarendon Press; 1986.

Baddeley A. The episodic buffer: A new component of working memory? Trends in Cognitive Sciences. 2000; 4(11):417–423. [PubMed: 11058819]

Baddeley A. The magic number and the episodic buffer. Behavioral and Brain Sciences. 2001; 24:117–118.

Baddeley, AD.; Hitch, GJ. Working memory. In: Bower, GH., editor. The psychology of learning and motivation: Advances in research and theory. New York: Academic Press; 1974. p. 47-89.

Baddeley, AD.; Logie, RH. Working memory: The multiple-component model. In: Miyake, A.; Shah, P., editors. Models of working memory: Mechanisms of active maintenance and executive control. Cambridge, UK: Cambridge University Press; 1999. p. 28-61.

Baddeley AD, Thomson N, Buchanan M. Word length and the structure of short-term memory. Journal of Verbal Learning & Verbal Behavior. 1975; 14(6):575–589.

Barrouillet P, Bernardin S, Portrat S, Vergauwe E, Camos V. Time and cognitive load in working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2007; 33(3): 570–585.

Barrouillet P, Portrat S, Camos V. On the law relating processing to storage in working memory. Psychological Review. 2011; 118(2):175–192. [PubMed: 21480738]

Bays PM, Husain M. Dynamic Shifts of Limited Working Memory Resources in Human Vision. Science. 2008; 321(5890):851–854. Available from http://www.sciencemag.org/cgi/content/abstract/321/5890/851. [PubMed: 18687968]

Bays PM, Husain M. Response to Comment on \Dynamic Shifts of Limited Working Memory Resources in Human Vision". Science. 2009; 323(5916):877d. Available from http://www.sciencemag.org/cgi/content/abstract/323/5916/877d. [PubMed: 22822271]

Berger JO, Berry DA. Statistical analysis and the illusion of objectivity. American Scientist. 1988; 76:159–165.

Borst JP, Taatgen NA, Rijn H van. The problem state: A cognitive bottleneck in multitasking. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2010; 36(2):363–382.

Broadbent, D. The magic number seven after fifteen years. In: Kennedy, A.; Wilkes, A., editors. Studies in long term memory. Oxford, England: John Wiley & Sons; 1975. p. 3-18.

Brown GDA, Neath I, Chater N. A temporal ratio model of memory. Psychological Review. 2007; 114(3):539–576. [PubMed: 17638496]

Camos V, Mora G, Oberauer K. Adaptive choice between articulatory rehearsal and attentional refreshing in verbal working memory. Memory & Cognition. 2011; 39(2):231–244.

Chen Z, Cowan N. Chunk limits and length limits in immediate recall: A reconciliation. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2005; 31(6):1235–1249.

Chen Z, Cowan N. Core verbal working-memory capacity: The limit in words retained without covert articulation. The Quarterly Journal of Experimental Psychology. 2009a; 62(7):1420–1429. [PubMed: 19048451]

Chen Z, Cowan N. How verbal memory loads consume attention. Memory & Cognition. 2009b; 37(6): 829–836.

Chong S, Treisman A. Statistical processing: computing the average size in perceptual groups. Vision Research. 2005; 45:891900.

Clark S, Gronlund S. Global matching models of recognition memory: How the models match the data. Psychonomic Bulletin and Review. 1996; 3:37–60.

Cowan N. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. Psychological Bulletin. 1988; 104(2):163–191. [PubMed: 3054993]

Cowan, N. An embedded-processes model of working memory. In: Miyake, A.; Shah, P., editors. Models of working memory: Mechanisms of active maintenance and executive control. Cambridge, UK: Cambridge University Press; 1999.

Cowan N. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences. 2001; 24:87–114. [PubMed: 11515286]

Cowan, N. Working memory capacity. Psychology Press; 2005.

Cowan N, AuBuchon AM, Gilchrist AL, Ricker TJ, Saults JS. Age differences in visual working memory capacity: Not based on encoding limitations. Developmental Science. 2011; 14(5):1066–1074. [PubMed: 21884322]

Cowan N, Chen Z, Rouder JN. Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). Psychological Science. 2004; 15:634–640. [PubMed: 15327636]

Cowan N, Elliott EM, Saults JS, Morey CC, Mattox S, Hismjatullina A, et al. On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. Cognitive Psychology. 2005; 51:42–100. [PubMed: 16039935]

Cowan N, Li D, Moffitt A, Becker TM, Martin EA, Saults JS, et al. A neural region of abstract working memory. Journal of Cognitive Neuroscience. 2011; 23(10):2552–2563.

Cowan N, Morey CC. How can dual-task working memory retention limits be investigated? Psychological Science. 2007; 18:686–688. [PubMed: 17680938]

Cowan N, Rouder JN. Comment on “Dynamic Shifts of Limited Working Memory Resources in Human Vision”. Science. 2009; 323(5916):877c. Available from http://www.sciencemag.org/cgi/content/abstract/323/5916/877c. [PubMed: 19213899]

Daily LZ, Lovett MC, Reder LM. Modeling individual differences in working memory performance: A source activation account. Cognitive Science: A Multidisciplinary Journal. 2001; 25(3):315–353.

Davelaar EJ, Goshen-Gottstein Y, Ashkenazi A, Haarmann HJ, Usher M. The demise of short-term memory revisited: Empirical and computational investigations of recency effects. Psychological Review. 2005; 112(1):3–42. [PubMed: 15631586]

Dyne AM, Humphreys MS, Bain JD, Pike R. Associative interference effects in recognition and recall. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1990; 16(5):813–824.

Edwards W, Lindman H, Savage LJ. Bayesian statistical inference for psychological research. Psychological Review. 1963; 70:193–242.

Eich J. A composite holographic associative recall model. Psychological Review. 1982; 89:627–661.

Gilchrist AL, Cowan N. Can the focus of attention accommodate multiple separate items? Journal of Experimental Psychology: Learning, Memory, and Cognition. 2011; 37:1484–1502.

Gilchrist AL, Cowan N, Naveh-Benjamin M. Working memory capacity for spoken sentences decreases with adult ageing: Recall of fewer but not smaller chunks in older adults. Memory. 2008; 16(7):773–787. [PubMed: 18671167]

Gilchrist AL, Cowan N, Naveh-Benjamin M. Investigating the childhood development of working memory using sentences: New evidence for the growth of chunk capacity. Journal of Experimental Child Psychology. 2009; 104(2):252–265. [PubMed: 19539305]

Gillund G, Shiffrin R. A retrieval model for both recognition and recall. Psychological Review. 1984; 91:97–123.

Glanzer M, Razel M. The size of the unit in short-term storage. Journal of Verbal Learning & Verbal Behavior. 1974; 13(1):114–131.

Gorgoraptis N, Catalao R, Bays P, Husain M. Dynamic updating of working memory resources for visual objects. The Journal of Neuroscience. 2011; 31(23):8502–8511. [PubMed: 21653854]

Healy AF, Fendrich DW, Cunningham TF, Till RE. Effects of cuing on short-term retention of order information. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1987; 13(3): 413–425.

Hintzman DL. Judgments of frequency and recognition memory in a multiple-trace model. Psychological Review. 1988; 95:528–551.

Hulme C, Tordoff V. Working memory development: The effects of speech rate, word length, and acoustic similarity on serial recall. Journal of Experimental Child Psychology. 1989; 47(1):72–87.

Jacoby LL. A process dissociation framework: Separating automatic from intentional uses of memory. Journal of Memory and Language. 1991; 30:513–541.

Jeffreys, H. Theory of probability. 3. New York: Oxford University Press; 1961.

Jiang Y, Chun MM, Olson IR. Perceptual grouping in change detection. Perception & Psychophysics. 2004; 66:446–453. [PubMed: 15283069]

Johnson NF. The memorial structure of organized sequences. Memory & Cognition. 1978; 6(3):233–239.

Jones GV. A fragmentation hypothesis of memory: Cued recall of pictures and of sequential position. Journal of Experimental Psychology: General. 1976; 105(3):277–293.

Just MA, Carpenter PA. A capacity theory of comprehension: Individual differences in working memory. Psychological Review. 1992; 99(1):122–149. [PubMed: 1546114]

Lewandowsky, S.; Farrell, S. Computational modeling in cognition: Principles and practice. Thousand Oaks, CA: Sage; 2011.

Lewandowsky S, Oberauer K. No evidence for temporal decay in working memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2009; 35(6):1545–1551.

Lewandowsky S, Oberauer K, Brown GDA. Response to Barrouillet and Camos: Interference or decay in working memory? Trends in Cognitive Sciences. 2009; 13(4):146–147.

Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g-priors for Bayesian variable selection. Journal of the American Statistical Association. 2008; 103:410–423. Available from http://pubs.amstat.org/doi/pdf/10.1198/016214507000001337.

Luck SJ, Vogel EK. The capacity of visual working memory for features and conjunctions. Nature. 1997; 390:279–281. [PubMed: 9384378]

Meulen, M van der; Logie, RH.; Della Sala, S. Selective interference with image retention and generation: Evidence for the workspace model. The Quarterly Journal of Experimental Psychology. 2009; 62(8):1568–1580. [PubMed: 19096989]

Miller GA. The magical number seven plus or minus two: Some limits on our capacity for processing information. Psychological Review. 1956; 63:81–97. [PubMed: 13310704]

Morey CC, Cowan N. When visual and verbal memories compete: Evidence of cross-domain limits in working memory. Psychonomic Bulletin & Review. 2004; 11:296–301. [PubMed: 15260196]

Morey CC, Cowan N, Morey RD, Rouder JN. Flexible attention allocation to visual and auditory working memory tasks: Manipulating reward induces a trade-off. Attention, Perception & Psychophysics. 2011; 73:458–472.

Murdock BB. Todam2: A model for the storage and retrieval of item, association, and serial order information. Psychological Review. 1993; 100:183–203. [PubMed: 8483981]

Murnane K, Phelps MP, Malmberg K. Context-dependent recognition memory: the ICE theory. Journal of Experimental Psychology: General. 1999; 128:403–415. [PubMed: 10650581]

Nairne JS. A feature model of immediate memory. Memory & Cognition. 1990; 18(3):251–269.

Nairne JS. Remembering over the short-term: The case against the standard model. Annual Review of Psychology. 2002; 53:53–81.

Nelder JA, Mead R. A simplex method for function minimization. Computer Journal. 1965; 7:308–313.

Oberauer K, Kliegl R. Beyond resources: Formal models of complexity effects and age differences in working memory. European Journal of Cognitive Psychology. 2001; 13(1–2):187–215.

Oberauer K, Kliegl R. A formal model of capacity limits in working memory. Journal of Memory and Language. 2006; 55(4):601–626.

Oberauer K, Lewandowsky S. Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? Psychological Review. 2008; 115(3):544–576. [PubMed: 18729591]

Penney CG. Modality effects and the structure of short-term verbal memory. Memory & Cognition. 1989; 17(4):398–422.

Portrat S, Barrouillet P, Camos V. Time-related decay or interference-based forgetting in working memory? Journal of Experimental Psychology: Learning, Memory, and Cognition. 2008; 34(6):1561–1564.

Potter MC, Lombardi L. Regeneration in the short-term recall of sentences. Journal of Memory and Language. 1990; 29(6):633–654.

R Development Core Team. R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria: 2009. Available from http://www.R-project.org

Raye C, Johnson M, Mitchell K, Greene E, Johnson M. Refreshing: A minimal executive function. Cortex. 2007; 43:135–145. [PubMed: 17334213]

Ricker T, Cowan N, Morey C. Visual working memory is disrupted by covert verbal retrieval. Psychonomic Bulletin & Review. 2010; 17(4):516–521. [PubMed: 20702871]

Rouder JN, Morey RD, Cowan N, Zwilling CE, Morey CC, Pratte MS. An assessment of fixed-capacity models of visual working memory. Proceedings of the National Academy of Sciences. 2008; 105:5976–5979.

Rouder JN, Morey RD, Morey CC, Cowan N. How to measure working-memory capacity in the change-detection paradigm. Psychonomic Bulletin & Review. 2011; 18:324–330. [PubMed: 21331668]

Rouder, JN.; Morey, RD.; Speckman, PL.; Province, JM. Default Bayes factors for ANOVA designs. (submitted)

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. Bayesian *t*-tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin and Review. 2009; 16:225–237. [PubMed: 19293088]

Saito S, Logie RH, Morita A, Law A. Visual and phonological similarity effects in verbal immediate serial recall: A test with kanji materials. Journal of Memory and Language. 2008; 59(1):1–17.

Salthouse TA. The processing speed theory of adult age differences in cognition. Psychological Review. 1996; 103:403–428. [PubMed: 8759042]

Saults JS, Cowan N. A central capacity limit to the simultaneous storage of visual and auditory arrays in working memory. Journal of Experimental Psychology: General. 2007; 136:663–684. [PubMed: 17999578]

Schwartz G. Estimating the dimension of a model. Annals of Statistics. 1978; 6:461–464.

Schweickert R, Guentert L, Hersberger L. Phonological similarity, pronunciation rate, and memory span. Psychological Science. 1990; 1(1):74–77.

Sellke T, Bayarri MJ, Berger JO. Calibration of *p* values for testing precise null hypotheses. American Statistician. 2001; 55:62–71. Available from http://www.jstor.org/stable/2685531.

Shiffrin RM, Nosofsky RM. Seven plus or minus two: A commentary on capacity limitations. Psychological Review. 1994; 101:357–361. [PubMed: 8022968]

Shiffrin RM, Steyvers M. A model for recognition memory: REM{retrieving effectively from memory. Psychonomic Bulletin and Review. 1997; 4:145–166. [PubMed: 21331823]

Simon HA. How big is a chunk? Science. 1974; 183(4124):482–488. [PubMed: 17773029]

Stevanovski B, Jolicœur P. Visual short-term memory: Central capacity limitations in short-term consolidation. Visual Cognition. 2007; 15(5):532–563.

Tehan G, Humphreys MS. Transient phonemic codes and immunity to proactive interference. Memory & Cognition. 1995; 23(2):181–191.

Thiele J, Pratte M, Rouder J. On perfect working-memory performance with large numbers of items. Psychonomic Bulletin & Review. 2011

Towse JN, Cowan N, Hitch GJ, Horton NJ. The recall of information from working memory: Insights from behavioural and chronometric perspectives. Experimental Psychology. 2008; 55(6):371–383. [PubMed: 19130763]

Tulving E, Patkau JE. Concurrent effects of contextual constraint and word frequency on immediate recall and learning of verbal material. Canadian Journal of Psychology/Revue canadienne de psychologie. 1962; 16(2):83–95.

Unsworth N, Engle RW. The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. Psychological Review. 2007; 114(1):104–132. [PubMed: 17227183]

Vergauwe E, Barrouillet P, Camos V. Visual and spatial working memory are not that dissociated after all: A time-based resource-sharing account. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2009; 35(4):1012–1028.

Vergauwe E, Barrouillet P, Camos V. Do mental processes share a domain-general resource? Psychological Science. 2010; 21(3):384–390. [PubMed: 20424075]

Wagenmakers EJ. A practical solution to the pervasive problem of p values. Psychonomic Bulletin and Review. 2007; 14:779–804. [PubMed: 18087943]

Waugh NC, Norman DA. Primary memory. Psychological Review. 1965; 72(2):89–104. [PubMed: 14282677]

Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson G, Wagenmakers EJ. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. Perspectives on Psychological Science. 2011; 6:291–298.

Wheeler ME, Treisman AM. Binding in short-term visual memory. Journal of Experimental Psychology: General. 2002; 131:48–64. [PubMed: 11900102]

Wilken P, Ma WJ. A detection theory account of change detection. Journal of Vision. 2004; 4:1120–1135. [PubMed: 15669916]

Zhang W, Luck SJ. Discrete fixed-resolution representations in visual working memory. Nature. 2008; 453:233–235. [PubMed: 18385672]

Appendix A: Base Word Triads

For a given participant in Experiment 1, a third of base word triads were used in full, a third as just the final pair (e.g., *door knob*) and a third as just the final singleton (e.g., *knob*). For Experiments 2 and 3, only the final singletons of each triad were used. The following triads were used:

brass door knob, college game day, curly pig tail, dark stone cavern, finger nail polish, giant horse fly, gold fish bowl, goose down pillow, graham cracker crust, heavy house work, helpful sales woman, honey bee sting, ice cold drink, leather brief case, little black book, live jazz band, loose knit shirt, magic school bus, metal hack saw, neon head light, orange lady bug, plaid table cloth, plush velvet rope, pulled leg muscle, red fire truck, silk night gown, silly monkey business, small sail boat, steamed egg roll, striped neck tie, swiss army knife, tin oil can, very nice person, wicked step mother, wooden mouse trap, yellow rain coat.

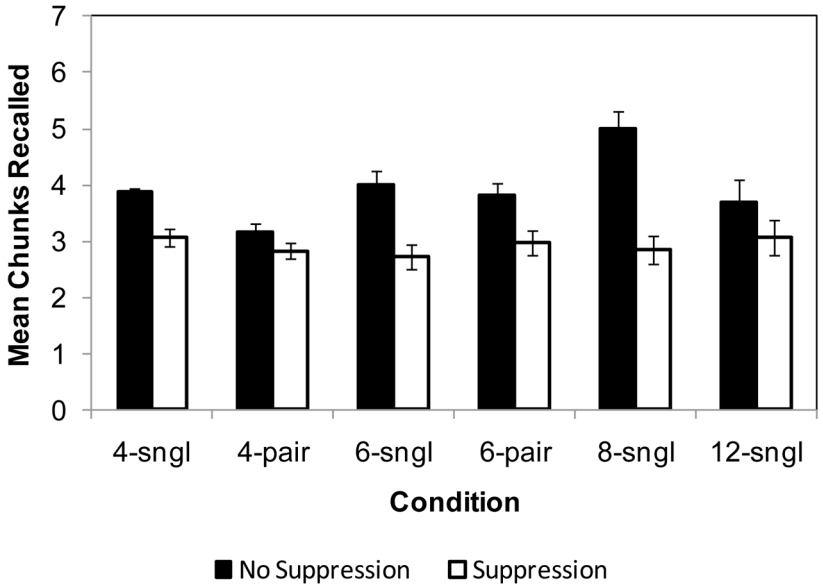

Figure 1.

Mean number of chunks recalled in several conditions from Chen & Cowan (2009a). Recalled chunks refer to recalled words in the singleton condition and word pairs in the pair condition. Articulatory suppression was used during list presentation in one of two groups. Chunks were scored without regard to serial position in the response. Sngl = singleton condition; pair = learned word pair condition. For example, *6-pair* refers to lists with 6 learned pairs. Error bars are standard errors.

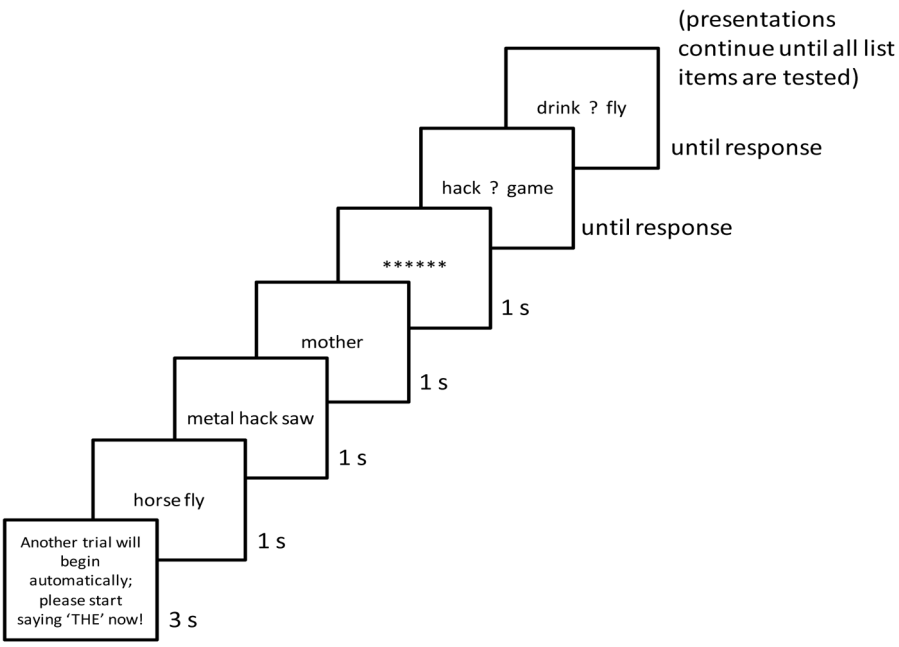

Figure 2.
Example of a list-recognition trial with a singleton, a pair, and a triplet (*123* condition).

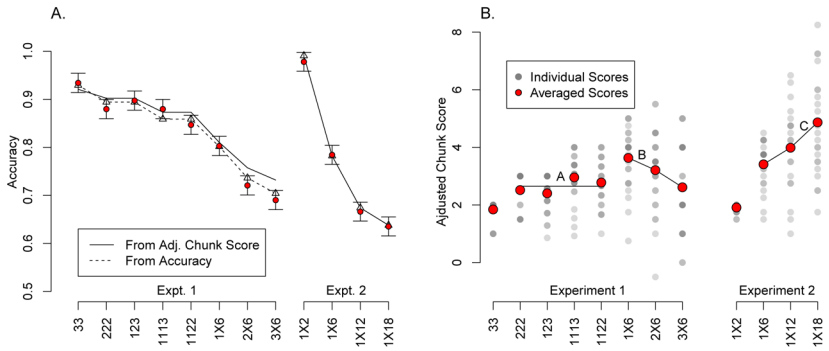

Figure 3.

Performance in both experiments. **A**. Accuracy in both experiments is denoted by points with error bars, where the error bars denote within-subject 95% confidence intervals. Lines show predictions from Model VIII when the model is fit to accuracy (dashed lines) or to adjusted chunk score (solid lines). Within-subject confidence intervals were calculated with the *1x2* condition excluded because the variability in this condition is attenuated from ceiling effects. **B.** Individual and average chunk score for both experiments is denoted by smaller translucent dots and larger dots, respectively. Trend A shows constant capacity; Trend B shows capacity decreases with increasing presented-chunk length; Trend C show capacity increases with increasing list length in chunks.

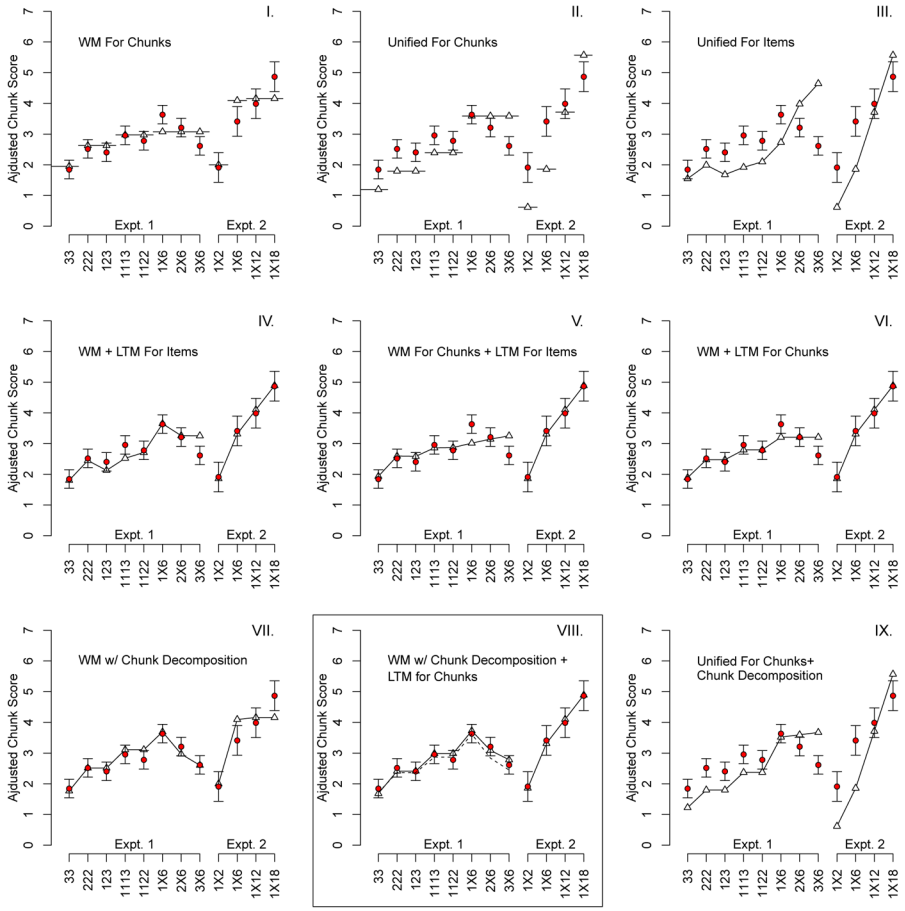

Figure 4.

Data and model predictions. The dots with error bars denote observed mean adjusted chunk scores and 95% within-subject confidence intervals, respectively. The triangles denote model predictions. The nine panels are for the nine models, one model per panel as noted in graph labels. The dashed line in Model VIII shows model predictions when the model was fit to the accuracy data rather than to the adjusted chunk score. Within-subject confidence intervals were calculated with the *33* and *1x2* conditions excluded because the variability in these conditions is attenuated from ceiling effects. The box around Model VIII signifies that it is the best-fitting model.

Table 1

List Conditions in Two Experiments

	Experiment 1		Experiment 2	
List Condition	**# of Words**	**# of Chunks**	**# of Words**	**# of Chunks**
33	6	2	-	-
222	6	3	-	-
123	6	3	-	-
1113	6	4	-	-
1122	6	4	-	-
1x6	6	6	6	6
2x6	12	6	-	-
3x6	18	6	-	-
1x2	-	-	2	2
1x12	-	-	12	12
1x18	-	-	18	18

Note. Naming convention: *33* stands for two chunks of three words; *2x6* stands for six chunks of two words; and other conditions are named by one of these two conventions.

Psychol Rev. Author manuscript; available in PMC 2013 July 01.

Table 2

Mean Parameter Estimates For Each Model In Each Experiment and AIC Increase Above Model VIII Value for Each Model

	Experiment 1				Experiment 2				
	k	***ℓ***	***p***	***u***	***k***	***ℓ***	***p***	***u***	**AIC**
I. WM for Chunks	3.1	-	-	-	4.1	-	-	-	68.7
II. Unified for Chunks	-	-	-	.60	-	-	-	.31	152.7
III. Unified for Items	-	-	-	.46	-	-	-	.31	272.7
IV. WM+LTM for Items	3.5	.06	-	-	2.8	.13	-	-	47.8
V. WM for Chunks + LTM for Items	2.8	.05	-	-	2.8	.13	-	-	56.8
VI. WM + LTM For Chunks	2.8	.20	-	-	2.8	.13	-	-	49.4
VII. WM w/decomposition	3.7	-	.33	-	4.1	-	n/a	-	91.4
VIII. WM w/decomposition + LTM	3.4	.13*	.37	-	2.8	.13	n/a	-	0
XI. Unified for Chunks w/decomposition	-	-	.03	.59	-	-	n/a	.31	203.0

*This value was fixed to be the same as that in Experiment 2 rather than estimated.

Note. Parameters: *k*=WM capacity; *ℓ*=LTM success probability; *p*=probability of failure of a within-chunk association, leading to chunk decomposition; and *u*=probability within a unified model that a unit (item or chunk, depending on the model) is stored.