Benchmarks for Models of Short Term and Working Memory

Klaus Oberauer

Stephan Lewandowsky

Edward Awh

Gordon D. A. Brown

Andrew Conway

Nelson Cowan

Christopher Donkin

Simon Farrell

Graham J. Hitch

Mark Hurlstone

Wei Ji Ma

Candice C. Morey

Derek Evan Nee

Judith Schweppe

Evie Vergauwe

Geoff Ward

Accepted for publication in *Psychological Bulletin*, March 9, 2018

Abstract

Any mature field of research in psychology – such as short-term/working memory – is characterized by a wealth of empirical findings. It is currently unrealistic to expect a theory to explain them all; theorists must satisfice with explaining a subset of findings. The aim of the present article is to make the choice of that subset less arbitrary and idiosyncratic than is current practice. We propose criteria for identifying benchmark findings that every theory in a field should be able to explain: Benchmarks should be reproducible, generalize across materials and methodological variations, and be theoretically informative. We propose a set of benchmarks for theories and computational models of short-term and working memory. The benchmarks are described in as theory-neutral a way as possible, so that they can serve as empirical common ground for competing theoretical approaches. Benchmarks are rated on three levels according to their priority for explanation. Selection and ratings of the benchmarks is based on consensus among the authors, who jointly represent a broad range of theoretical perspectives on working memory, and they are supported by a survey among other experts on working memory. The article is accompanied by a web page providing an open forum for discussion; a site for submitting proposals for new benchmarks; and a repository for reference data sets for each benchmark.

Keywords: Working memory; benchmarks; computational modelling

Public Significance Statement

Working memory – the system for holding information in mind and working on it – is central for cognition. The authors identify a set of findings about working memory that are well established, general, and theoretically informative. These benchmark findings should be explained with high priority by theories of working memory. The set of benchmark findings will facilitate building theories and comparing competing theories, and thereby advance our understanding of human cognition.

Benchmarks for Models of Short-Term and Working Memory

Since G. A. Miller, Galanter, and Pribram (1960) introduced the term "working memory" to refer to a temporary store for action-relevant information, about 11,600 articles have been written with "working memory" in their title[1]. Research on the topic has received a boost by the seminal paper entitled "working memory" by Baddeley and Hitch (1974), who proposed a multi-component system to replace the "short-term store" in earlier memory models. Today it is generally accepted in cognitive psychology that working memory (WM) plays a central role in all deliberative cognition, from language comprehension and mental arithmetic to reasoning and planning. A multitude of theories has emerged to characterize WM and explain phenomena related to it (for an early review see Miyake & Shah, 1999). Although there is no agreed-upon definition of "working memory", there is a core meaning of the term identifiable in most, if not all theories of it: WM refers to a system, or a set of processes, holding mental representations temporarily available for use in thought and action (Cowan, in press). We use this characterization as our working definition, chosen deliberately to be broad, including also what some researchers refer to as "short-term memory".

The extensive literature on WM reports a vast number of findings. Although this empirical richness means that, in some sense, we know a lot about WM, it raises an enormous challenge for theoretical progress: No theory can hope to get even close to explaining all existing findings. Therefore, theorists must decide which findings their theory should explain with highest priority. Our observation is that theorists —including some of us – have often made these decisions in an ad hoc and idiosyncratic fashion, prioritizing the findings they were most familiar with (often because they have emerged from their own empirical work), or those that their theory happens to be able to explain. The challenge of choosing findings to explain comes into particularly sharp focus when theories are formalized as computational models. Whereas verbally formulated theories can leave many details unspecified, giving them a high degree of flexibility for post-hoc adjustments to accommodate many findings, computational models are much less flexible: All assumptions must be made explicit, and predictions are derived from them through computation, leaving little ambiguity as to what a model

[1] Web of Science, search on November 18, 2017, using keyword „working memory" in the title, including Psychology, Neurosciences, and neighboring disciplines.

predicts (Farrell & Lewandowsky, 2010). Extending a computational model to a new finding requires adding assumptions, or at least changing parameter values, which risks destroying the model's previous explanatory success. Therefore, computational modelers are forced to acknowledge the limited scope of their models more explicitly than proponents of not-formalized theories, and they must decide more explicitly which findings to prioritize as targets for explanation.

The question of which findings to prioritize is even more pressing when it comes to theory competition: Given two theories *A* and *B* that explain the sets of findings *a* and *b*, respectively, how can we decide which theory is a better theory of "working memory" when the sets *a* and *b* overlap only partially, or not at all? It might be tempting at this point to reject the question as ill posed: Neither *A* nor *B* are complete theories of WM; rather, *A* is a theory of all empirical phenomena in set *a*, and B is a theory of all findings in set *b*. This approach is unsatisfactory when *A* and *B* make mutually incompatible assumptions. For instance, *A* might include the assumption that WM consists of multiple subsystems with separate stores, whereas *B* includes the assumption that the system has only a single store (and another theory *C* denies that there is any dedicated store of WM at all). In such a case, accepting both *A* and *B* as valid theories in their own domain is acceptable as a temporary solution at best, because holding mutually contradictory beliefs opens the door to mutually contradictory predictions.

In the field of computational modeling, much recent research has focused on the problem of model comparison. There is by now a highly sophisticated body of work on how to adjudicate between two formal models that make predictions for the same data set (Pitt, Myung, & Zhang, 2002). At the same time, hardly any consideration is given to the question of which data set (or sets) should be chosen for the comparison between two models. The decision on which data to test a model against involves "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011) that are not accounted for by model comparison techniques.

The aim of the present article is to offer a first proposal for addressing this question in one field of research, WM. As a stepping stone towards theoretical progress, researchers need an agreed-upon set of benchmarks for theories and models of WM, that is, a set of phenomena that every theory or model in that field should strive to explain. The purpose of a set of benchmarks is to provide common empirical ground for assessing theories and models: Any theory of WM can be measured against the benchmarks as a common yardstick to determine the theory's explanatory power and its limitations. Competing theories can be compared by examining how well they fare in explaining the benchmark phenomena. Modelers can use the benchmarks as a well-defined set of phenomena that they can aim to explain.

In this article we propose an initial set of benchmarks for theories and models of WM. We think of it as a starting point for a discussion that, we hope, will eventually lead to a consensus on a set of benchmark findings that are empirically robust and theoretically incisive, so that they can be relied upon as the common empirical constraints for competing theories and models. This would enable theorists to reduce their researcher degrees of freedom in selecting the findings that they evaluate their model against: They could start their efforts by aiming to model the benchmark findings with first priority.

Although considerably smaller than the set of all published findings, our proposed set of benchmarks is still large – probably too large for any model to explain them all in the foreseeable future. Therefore, we classified the benchmarks into three levels of priority (A, B, and C), so that modelers could initially pay more attention to explaining priority-A benchmarks before turning to those with B and C ratings.

To serve its purpose, the set of benchmarks needs to be as unbiased and theory-neutral as possible. This is why the present article is not a traditional review: We abstain as much as possible from theoretical interpretation of the phenomena we propose as benchmarks, and we will draw no theoretical conclusions from them. To be theory-neutral does not mean to be theory-free. There is arguably no theory-free language for describing empirical findings. Yet, we can reasonably aim for a description of each benchmark that is not biased in favor or against one contemporary theoretical view, so that theorists from different perspectives can agree on the description of a benchmark while disagreeing on its explanation. This aim is the reason why our endeavor could only be accomplished by a large team of researchers reflecting a broad range of theoretical perspectives and fields of specialization (as is reflected in the long list of authors of this article). To build the set of benchmarks on an even broader foundation, we invited scholars in the field to contribute to the development of the benchmarks through an online survey.

Methods

This section describes the process of collectively generating the set of benchmarks and of finding a consensus on their selection and their ratings.

Procedure

Initial Workshop. In October 2013 two of us (SL and KO) invited 26 researchers on human WM to a workshop with the purpose of developing benchmarks for models of WM. We selected invitees to represent the full diversity in the field with respect to theoretical views, areas of expertise, age and seniority, and geographical region. The majority of those invited – the present authors – accepted and formed the Benchmarks Team (those who declined did so primarily because of conflicting commitments). The organizers asked each member of the Team to be responsible for coverage of one subfield of WM research in which they specialized. At the workshop each Team member proposed between one and three phenomena as benchmarks, which we discussed in plenary and small-group meetings. During these discussions we developed a consensus on criteria for benchmarks, and on how to delimit the scope of findings that we regard as reflecting WM. The main result of the workshop was a preliminary set of benchmark candidates.

The candidate set was put together with a heuristic of inclusiveness: If the Team could not agree whether or not a finding should be a benchmark, it was included in the candidate list. The candidate list was a structured list that reflected the relations between phenomena: Groups of findings were clustered together because they pertain to a common theme, or appear to be instances of a more general phenomenon. Moreover, the candidate list consisted of main and subordinate findings. Main findings are those that we regarded as informative and important on their own, whereas subordinate findings derive their importance from their role in specifying the exact nature of a main finding, for instance by revealing its boundary conditions or by characterizing it in more detail.

We formulated each benchmark candidate in a way that is as theory neutral as possible, limiting the statement to a generalized description of the finding and avoiding potentially controversial interpretations. We did not aim for completely theory-free formulations, because we are not convinced that the description of observations can be completely divorced from theoretical concepts and ideas: Theoretical considerations influence which experiments we run, and they influence how we generalize findings across individual studies. We encountered the limits of theory-free descriptions particularly clearly in cases where the observation of benchmark findings depends on measurement procedures that rely on theoretical assumptions about (working) memory. This is the case, for instance, when measuring the maximum number of chunks that a person can hold in WM (benchmark 1.3): Efforts to obtain a pure estimate of the number of chunks held in WM involve methods to determine what is a chunk, and measures to control for extraneous factors influencing performance, both of which are informed by theoretical considerations. In these (few) cases we formulated a *conditional benchmark* of the form: "If measurement methods X, Y, and/or Z are applied, finding A is regularly obtained", making explicit the theoretical assumptions entering the choice of the measurement methods. Theorists who disagree with the premises of the measurement methods can and should still aim to explain the conditional benchmark, for instance by simulating data from a computational model and showing that, after applying the relevant measurement methods to the simulated data, the benchmark can be reproduced (for an example of this approach see van den Berg & Ma, 2014).

Expert Surveys. Subsequent to the workshop, all members of the Benchmarks Team independently rated every finding on the candidate list on a four-point scale: A (highest priority), B (intermediate priority), C (low priority), and “not a benchmark”. Contrary to our expectation, none of the candidates received a majority rating of “not a benchmark”, so we did not exclude any candidate at this stage.

To further reduce the chance of accidentally ignoring a potential benchmark, and to counteract any inadvertent bias in the selection and rating of benchmarks, in a further step we invited more than 200 researchers on human WM to take part in an online survey on the benchmark candidates; 81 of them responded. The set of researchers invited consisted of all participants of the International Conference on WM 2014 whose email addresses were available online, supplemented by authors of articles on WM in the literature data base of the first author. The survey was implemented with Qualtrics and hosted by the University of Western Australia. We also posted the link to the survey on the web page of the Psychonomic Society and circulated it through the newsletter of the European Society of Cognitive Psychology (ESCoP), with an open invitation to participate. The survey instruction included a brief description of the project's purpose and the three criteria for benchmarks explained in the next section. Because of the large number of benchmark candidates, each respondent was presented a randomly selected subset of 71 items. The random selection was performed for each participant by Qualtrics, subject to two constraints: (1) Subordinate findings were not presented without the corresponding main findings, because they would be difficult to interpret in the absence of that context, and (2) findings that received highly variable ratings from the members of the Benchmarks Team were always included in the survey, because we needed more information on these candidates than on the ones on which we had already achieved a high degree of consensus. In addition to rating the benchmark candidates, respondents had the opportunity to add further proposals in a free-text form at the end of the survey. Results from the expert survey are reported in Appendix A.

Follow-Up Workshop. The final set of benchmarks was developed by the Benchmarks Team during a follow-up workshop in the summer of 2015, by which time the survey results had become available. The ratings of benchmarks according to their priority (A, B, or C) was also agreed on at the workshop. The survey data were taken as one piece of information in these decisions, alongside the Team's judgments on the criteria for benchmarks, as outlined in the next section. Minor refinements of the selections and ratings were made through subsequent discussions.

Criteria and Scope of Benchmarks

We used the following criteria, agreed upon during the first workshop, to determine whether a finding represents a benchmark: (1) A benchmark must be *reproducible*, that is, there should be published replications, preferably from different labs. (2) A benchmark must *generalize*; the importance ranking that a benchmark deserves should increase with its breadth of generalization across several dimensions: Details of the testing method (e.g., presentation duration, presentation modality), experimental paradigm (e.g., serial recall, probed recall, recognition), material (e.g., words, digits, spatial locations), and population (e.g., individuals of different ages and educational backgrounds). (3) A benchmark must have *theoretical leverage*. That is, benchmarks must be informative for theoretical questions by distinguishing between theoretical proposals that are compatible with those benchmarks and others that are not. This third criterion is clearly the most difficult to ascertain because it is difficult to determine whether a theory or hypothesis is compatible with a finding, and because there is no way to map out the space of possible theories. Therefore, we had to rely on a criterion for theoretical leverage that is tied to the historical context of the theoretical discussion in the field: We regarded a finding as theoretically informative if it has been used to make a case in favor or against a theoretical proposal.

During the follow-up workshop we also refined our definition for the three priority levels as follows:

Rating A: The benchmark is general across paradigms and content domains. No theory must contradict it because it is a fundamental fact of WM. It should be addressed by theories with high priority insofar as it falls into the intended explanatory scope of the theory.

Rating B: The benchmark applies to a narrower set of tasks or paradigms than an A benchmark. It need not be addressed by general theories of WM with high priority, but theories focusing on the specific domain or paradigm for which the benchmark has been established must accommodate it.

Rating C: The benchmark finding is specialized, and diagnostic in a narrow domain for a specific theoretical question only. Robust findings that qualify a more general finding, for instance by an interaction, often received the C rating. We also assigned a C rating to relatively novel findings that are of high theoretical leverage but for which robustness or generality still need to be ascertained.

Our endeavor also necessitated a decision on the scope of the set of benchmarks: We had to determine which findings belong to the field of WM. One plausible way of delineating the scope of the field might be to start from a definition of WM. This path was closed to us because we decided to steer clear of theoretical commitments as much as possible, and definitions of scientific concepts are closely tied to theories of that concept. We therefore took the pragmatic route, including into the scope of our endeavor every finding that researchers in the field regard as being informative about WM. This pragmatic decision implies that in many regards we preferred to err on the over-inclusive side. For instance, some theories define WM as different from short-term memory, characterizing the latter as involving mere maintenance of information, whereas the former includes some form of processing (other than that required for a memory test). Other theories define WM in a more inclusive way, also encompassing mere maintenance (Cowan, in press). We included findings from tasks requiring only maintenance because excluding them would introduce a bias against the more inclusive theories and definitions of WM. In contrast, including them gives theorists the choice to define a broader or narrower scope for their model (an issue we will return to in the Discussion).

There are two instances where we could have made the scope even broader but decided against it. First, although many theorists see WM as closely aligned to executive functions, we did not include findings speaking primarily to executive functions (e.g., findings on Stroop interference, task switching, or verbal fluency) because research on executive functions has become a field of its own, with theories and models largely separate from theorizing on WM (Botvinick, Braver, Barch, Carter, & Cohen, 2001; J. W. Brown, Reynolds, & Braver, 2007). We maintain that a good theory of WM should not presently aim, with high priority, to explain the Stroop effect, task-switching phenomena, and other findings on executive control. Second, we did not include experimental findings on how WM contributes to a host of cognitive tasks, from language comprehension (Just & Carpenter, 1992; Lewis, Vasishth, & van Dyke, 2006) to mental arithmetic (Hecht, 2002; Logie, Gilhooly, & Wynn, 1994) to reasoning (Barrouillet & Lecas, 1999; Cho, Holyoak, & Cannon, 2007; Handley, Capon, Copp, & Harper, 2002), because explaining these findings relies at least as much on a model of the domain of application (e.g., a model of syntactic parsing, or of deductive reasoning) as on a model of WM. That said, we did include correlations between measures of WM capacity and performance in some other cognitive tasks insofar as these correlations are informative for theories about WM without requiring detailed theories about those other tasks (see Benchmarks 12.6 and 12.7).

Updating and Use of the Benchmarks

The present article can at best provide a snapshot of benchmarks for WM models at its time of writing. Empirical research in the field will make progress by which new important findings will emerge, and findings that we propose as benchmarks today might be viewed as much less important or general in light of future discoveries. Therefore, we put in place a mechanism for continuous discussion and periodical revision of the benchmarks. Specifically, we set up a web page[2] with (1) the current set of benchmarks, (2) a forum for general comments open to everyone, and (3) a second forum specifically dedicated to proposals of new benchmarks. We invite all researchers on WM to propose new benchmarks that meet the criteria outlined above.[3] We plan to prepare a revised version of the present set of benchmarks within 4 to 5 years. We want that revision to be as representative as possible of the perspective of all researchers in the field, and therefore we invite all scholars of WM to join the Benchmarks Team for preparing the revision.[4]

We are aware that there is typically more to a phenomenon than can be put into a brief verbal description. The ultimate aim of theories and computational models should not be to reproduce our verbal description of the benchmark phenomena but to reproduce data that reflect these phenomena. To facilitate that, we have started to put together a set of reference data for each of the phenomena on the benchmarks list. These data are available for download from a public repository.[5] We invite all researchers to contribute further data sets that are representative for one or several benchmark findings. Our long-term goal is to provide reference data sets for all benchmarks that meet the following criteria: (1) They use large samples of participants and trials to provide the basis for precise estimates of model parameters (i.e., effect sizes in statistical models, and estimates of latent variables in theoretical models). (2) They cover a broad range of methodological variants to establish the generality (or lack thereof) of the benchmark in question. (3) They are pre-registered replications of benchmark findings; such replications are desirable because, despite our efforts to ensure that all benchmarks are robust and replicable, we cannot rule out that the available evidence is compromised by publication bias. Some members of the Benchmark Team plan to carry out such pre-registered replications, and we encourage all researchers in the field to contribute to that effort.

[2] URL: https://wmbenchmarks.wordpress.com/
[3] Proposals should be backed by references and a representative, ideally pre-registered, data set.
[4] For an expression of interest, send an email to the first or the second author: k.oberauer@psychologie.uzh.ch, or stephan.lewandowsky@bristol.ac.uk
[5] URL: https://github.com/oberauer/BenchmarksWM.git, and https://osf.io/g49c6/

In the remainder of this article we describe each benchmark, and justify its selection and its rating. To limit the article's length, we describe benchmarks with A and B ratings in the main text, and those with C ratings in Appendix B. Where we describe individual studies in detail, or plot illustrative data, we chose them to be representative (i.e., using a typical experimental paradigm and typical materials) and comprehensive (i.e., covering a broad range of experimental conditions relevant to the benchmark). When we illustrate benchmark findings with figures, we produced them, where possible, from raw data available to us. In these cases, error bars reflect 95% confidence intervals for within-subjects comparisons (Bakeman & McArthur, 1996), which give an indication of the variability of within-subjects effects (but not of individual data points).

Table 1: Brief Descriptions of Experimental Paradigms for Studying Working Memory

Name	Description
Serial Recall (SR)	Reproduction of a sequentially presented list of items in the order of presentation
Free Recall (FR)	Reproduction of a list of items in free order
Probed Recall (PR)	Recall of items in response to a retrieval cue uniquely identifying that item (e.g., its ordinal list position, or its spatial location)
Reconstruction of order (ROO)	Reproduction of the order of presentation of a list of items by placing each item in its correct ordinal list position (e.g., by moving the item with the mouse into a spatial place-holder for its list position)
Complex Span (CS)	Presentation of a list of items is interleaved with brief episodes of a distractor processing task; at the end the items have to be recalled (usually in serial order)
Brown-Peterson (BP)	Recall of a short list of items after a retention interval filled with a distractor processing task
Change Detection (CD)	An array of objects with simple visual features (e.g., colors, orientations) is presented briefly; after a brief delay the entire array is presented again, and

	the person decides whether or not one feature has been changed. Sometimes only one object is presented at test in the location of one original array object.
Continuous reproduction (a.k.a. delayed estimation) (CR)	An array of objects with simple visual features (e.g., colors, orientations) is presented briefly. After a brief delay one object is marked (usually by its location), and the person reproduces its feature on a continuous response scale (e.g., selecting its color on a color wheel), enabling the measure of memory precision on a continuous scale
N-Back (NB)	A long series of stimuli is presented sequentially, and the person decides for each stimulus whether it matches the one presented *n* steps back.
Running memory span (RM)	A series of stimuli of unpredictable length is remembered, and when it stops, the person is asked to recall the last *n* list elements, or as many list elements as possible from the end of the list.
Item Recognition (IRec)	A sequentially presented list, or simultaneously presented array of items is remembered briefly, followed by a single probe; the person decides whether that probe was contained in the memory set.
Relational Recognition (a.k.a. Local Recognition) (RRec)	Each item of a memory set is presented in a unique location, or in a relation with another unique stimulus, such as color. At test, a probe is presented in one of the locations (or in conjunction with one of the unique stimuli), and the person decides whether the probe matches the memory item in that location (or with that unique stimulus).
Memory Updating (MU)	Presentation of a set of initial items (e.g., digits, spatial locations of objects) is followed by instructions to update individual items, either through transformation (e.g., adding or subtracting some value from a digit, or shifting an object to a new spatial location) or through replacement (e.g., presenting a new digit, or presenting an object in a new location).

Note: Each paradigm name is accompanied by the acronym that this paradigm is given in the reference tables in Appendices C and D.

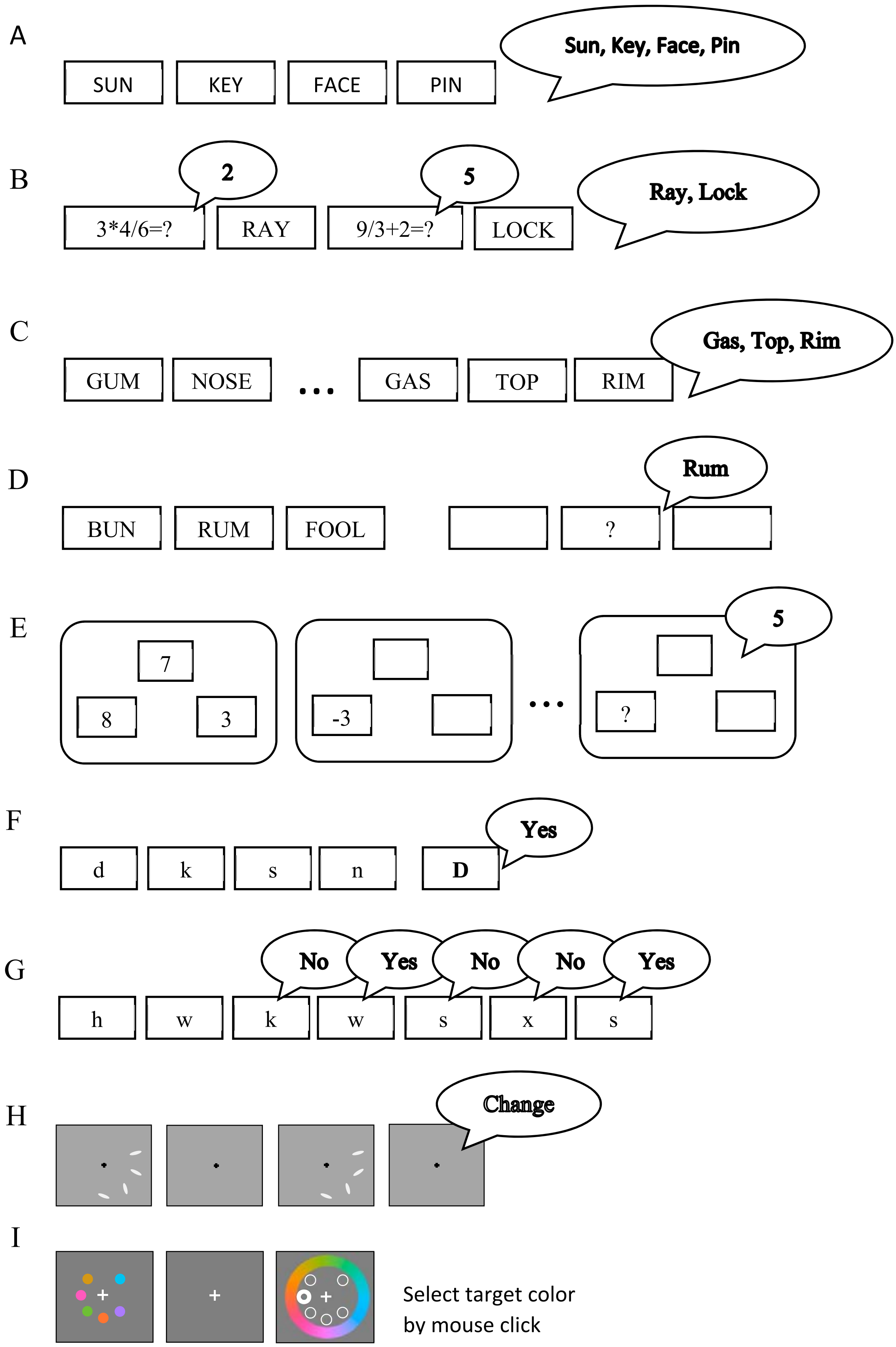

Figure 1: Schematic illustrations of paradigms for investigating WM, with flow of events from left to right. The examples show visual presentation and mostly oral responses, but the modality of

presentation and response varies across experiments. A: Immediate serial recall (a.k.a. simple span): A list of items (e.g., digits, letters, words, spatial locations) is presented sequentially, typically at a rate of 0.5 to 1 s per item. Immediately after the last item, participants attempt to recall the list in forward order of presentation. A variant of this paradigm, backward recall (not shown) requires recall in the reverse order of presentation. B: Complex span: Brief episodes of distractor processing are interleaved with presentation of items for immediate recall. C: Running Memory Span: A list of unpredictable length is presented sequentially. When the list stops, participants try to recall the last N items. D: Probed recall: After sequential presentation of a list of items, one item selected at random is probed for recall, for instance by a cue to its spatial position in a row from left to right, or by presenting one list item and asking participants to recall the next item. E: WM updating: Starting values (e.g., digits) are presented across a set of boxes, and are updated according to a series of operations (e.g., additions and subtractions) displayed in individual boxes. After several updating operations the final values in each box are tested. F: Item recognition: After presentation of a list of items, a single item probe is presented, and participants decide whether that item is an element of the list. G: N-back: A series of stimuli is presented, and participants decide for each stimulus whether it matches the one presented N steps back (the example is for N=2). H: Change-detection: An observer reports whether or not a change occurred between study and test, either in a single probed item or in any item in a whole array. Variants of this paradigm (not shown) ask for identifying the direction of change (e.g., whether the changed bar was rotated clockwise or counter-clockwise) or the location of change (i.e., which of the items has changed). I: Continuous reproduction (a.k.a. delayed estimation): An observer reproduces the feature of a target item in the array – marked here by the thick white outline in the test display – on a continuous response scale, for instance by selecting the target color on a color wheel.

We document the generality of each benchmark across experimental paradigms, content domains, and populations through two reference tables. Appendix C presents a table with all benchmarks, with references to supporting studies, organized by paradigm and content domain. Appendix D presents an analogous table, organized by paradigm and age group. As studies with children and older adults typically enroll a broader range of educational levels than studies with young adults (i.e., mostly university students), generality across age groups also goes some way towards demonstrating generality across educational background.

1: Set-Size Effects

Benchmarks 1.1: Set-Size Effects on Accuracy (Rating: A).

On every test of WM, accuracy declines as the set size increases. The set size refers to the number of elements in the set that participants are asked to hold in WM. These elements can be linguistic units such as digits, letters, and words (Crannell & Parrish, 1957; Guilford & Dallenbach, 1925), spatial locations (Cortis, Dent, Kennett, & Ward, 2015; Glahn et al., 2002), features or conjunctions of features of visual stimuli (Luck & Vogel, 1997; Shah & Miyake, 1996), and many others. The set-size effect on accuracy has been observed across a broad range of paradigms for studying WM (see Table 1 and Figure 1 for an overview of paradigms): Serial and free recall of verbal and spatial lists (Cortis et al., 2015; Grenfell-Essam & Ward, 2012; Smyth & Scholey, 1994, 1996; Woods, Wyma, Herron, & Yund, 2016), complex span for verbal and visual-spatial materials (Shah & Miyake, 1996; Unsworth & Engle, 2006a), probed recall (Murdock, 1968a), running memory span (Bunting, Cowan, & Saults, 2006; Pollack, Johnson, & Knaff, 1959), working-memory updating (Oberauer & Kliegl, 2001), item recognition (McElree & Dosher, 1989; R. E. Morin, DeRosa, & Ulm, 1967), n-back (Jonides et al., 1997; Verhaeghen, Cerella, & Basak, 2004), change detection (Luck & Vogel, 1997; Morey, Morey, van der Reijden, & Holweg, 2013), and change discrimination (J. Palmer, 1990). When participants are asked to reproduce a visual feature varying on a continuous dimension (e.g., its orientation or color), their mean absolute deviation from the true feature value increases with set size (Wilken & Ma, 2004; c.f. Benchmark 4.5). Representative data from six paradigms are shown in Figure 2.

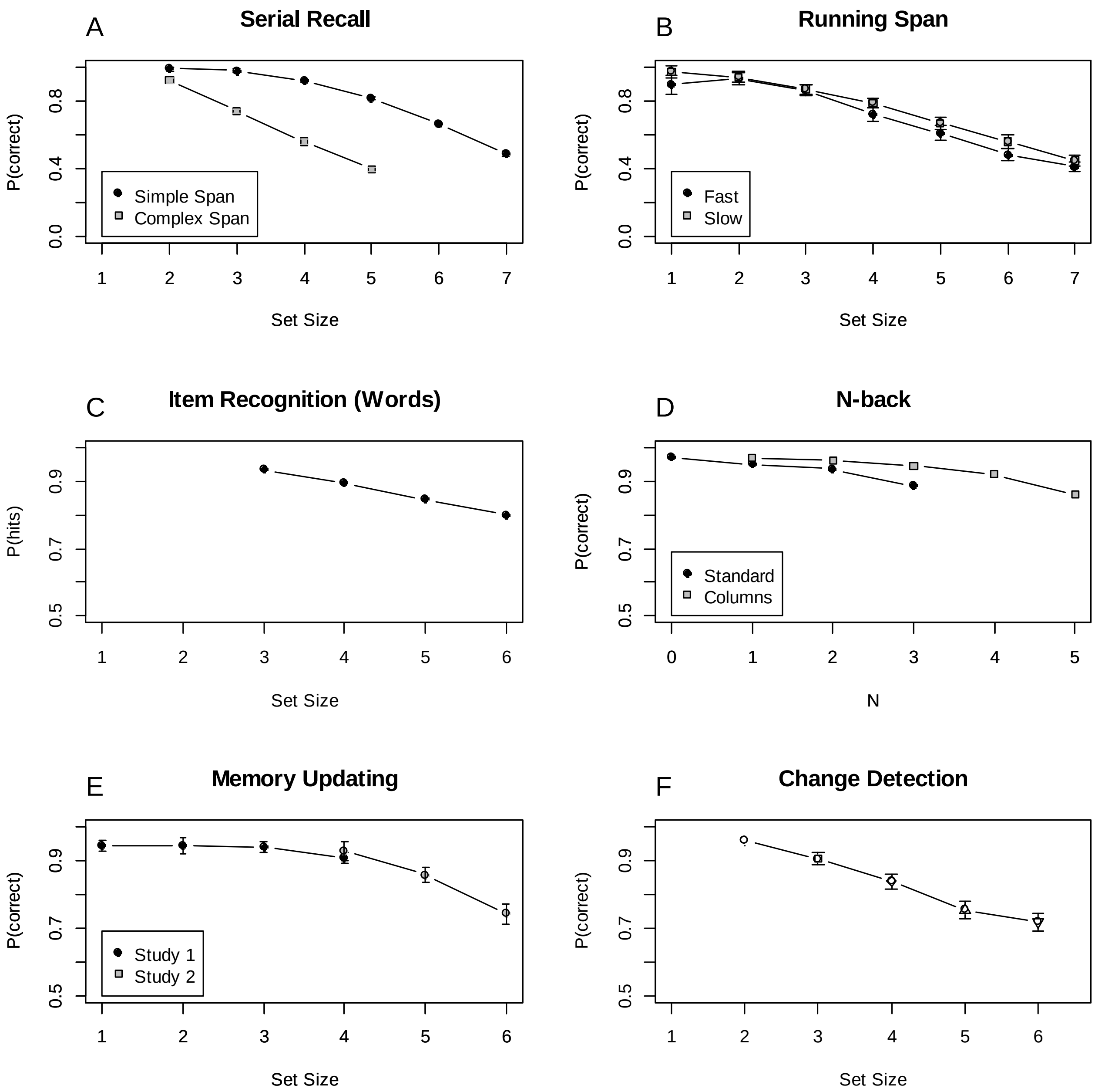

Figure 2: Representative findings demonstrating the set-size effect on accuracy. A: Serial recall in simple and complex span tests with verbal materials (Unsworth & Engle, 2006a). B: Running memory span (Bunting et al., 2006). C: Item recognition (McElree & Dosher, 1989). D: Standard N-back (Jonides et al., 1997), and a version of N-back in which subsequent stimuli are presented across N columns, such that each stimulus appears in the same column as the one N steps back (Verhaeghen & Basak, 2005), E: WM updating with digits and arithmetic operations (Oberauer & Kliegl, 2001), F: Change detection with arrays of colored squares (Adam, Mance, Fukuda, & Vogel, 2015).

The set size effect on accuracy is a highly robust and general phenomenon. Moreover, it reflects a core feature of WM: people's ability to hold information available for processing is severely limited. Therefore, we assign this benchmark the highest level of priority (A), in agreement with the majority of survey respondents (50/81) (see Appendix A for the full survey data).

Boundary Conditions. There is no paradigm used for studying WM that does not show a set-size effect on accuracy, but there are conditions under which the set-size effect is strongly mitigated, such that people can remember much larger sets than is typically observed. This occurs when the person can bring to bear long-term knowledge of relations between elements in a memory set. For example, a rich retrieval structure permits people to remember a much larger set of items over the short term than when the memory set does not afford applying that specialized knowledge (see Benchmark 1.3). These are mostly conditions giving rise to chunking (see Benchmark 11.1). The set-size effect was also much reduced in an item-recognition task when stimuli were used only once throughout the experiment (Endress & Potter, 2014).

Benchmark 1.2: Set-Size Effects on Retrieval Latency (A)

Across all types of WM tasks, it is generally true that responses drawing on information in WM become slower as there are more things to remember. The benchmark we propose here is a monotonic increase in response time as the size of the memory set increases. This pattern emerges across a range of WM tasks, including recognition, serial and free recall, WM updating, as well as change-detection tasks. The result is robust, has been replicated often, and has been informative for theorizing about access to WM since the early days of cognitive psychology (Sternberg, 1966, 1969), and hence we give it an A rating. This rating is in line with the most frequent rating in the survey (39/77).

Recognition. Perhaps the most famous early demonstration of the effect of set size on response latency was Sternberg’s (1966) work using the short-term memory scanning task. Sternberg presented participants with up to six digits (0-9), each for 1.2 s, followed by a delay of 2 s. At test, participants were asked to indicate whether a digit was either part of the study list, or *old*, or was not studied, or *new*. The seminal result was that the time to recognize both new and old items increased with the size of the study list. Though Sternberg found that set-size was linearly related to response time, this result is not observed in all experiments, and therefore our benchmark is that the relationship is monotonic. Moreover, to obtain a more complete picture, the set-size effect needs to be further decomposed. Monsell (1978) found that there was no impact of set-size on the time to recognize old items, above and beyond the effect of serial position. Rather, old items in earlier serial positions – with a longer lag between study and test – were recognized more slowly than those in later positions. Longer lists yielded slower responses on average because they contained more items with a larger study-test lag. Donkin and Nosofsky (2012b) found that the timing of the task moderates whether set size has an influence on the time to recognize old items. The average response times from Donkin and Nosofsky (2012b) are plotted in Figure 3. When study items were presented relatively quickly (500 ms per item), and with relatively little delay between study and test (500 ms), recognition response times for old items were driven primarily by serial position. However, when Sternberg's slower timings were used, there was a distinct influence of set size, and almost no role of serial position.

One boundary condition of the set-size effect on recognition latency has been observed in a local-recognition task, in which probes are presented in the locations of list items, to be compared only to the item in the same location. When each location is tested one by one in the order of presentation, the set-size effect disappears (Lange, Cerella, & Verhaeghen, 2011).

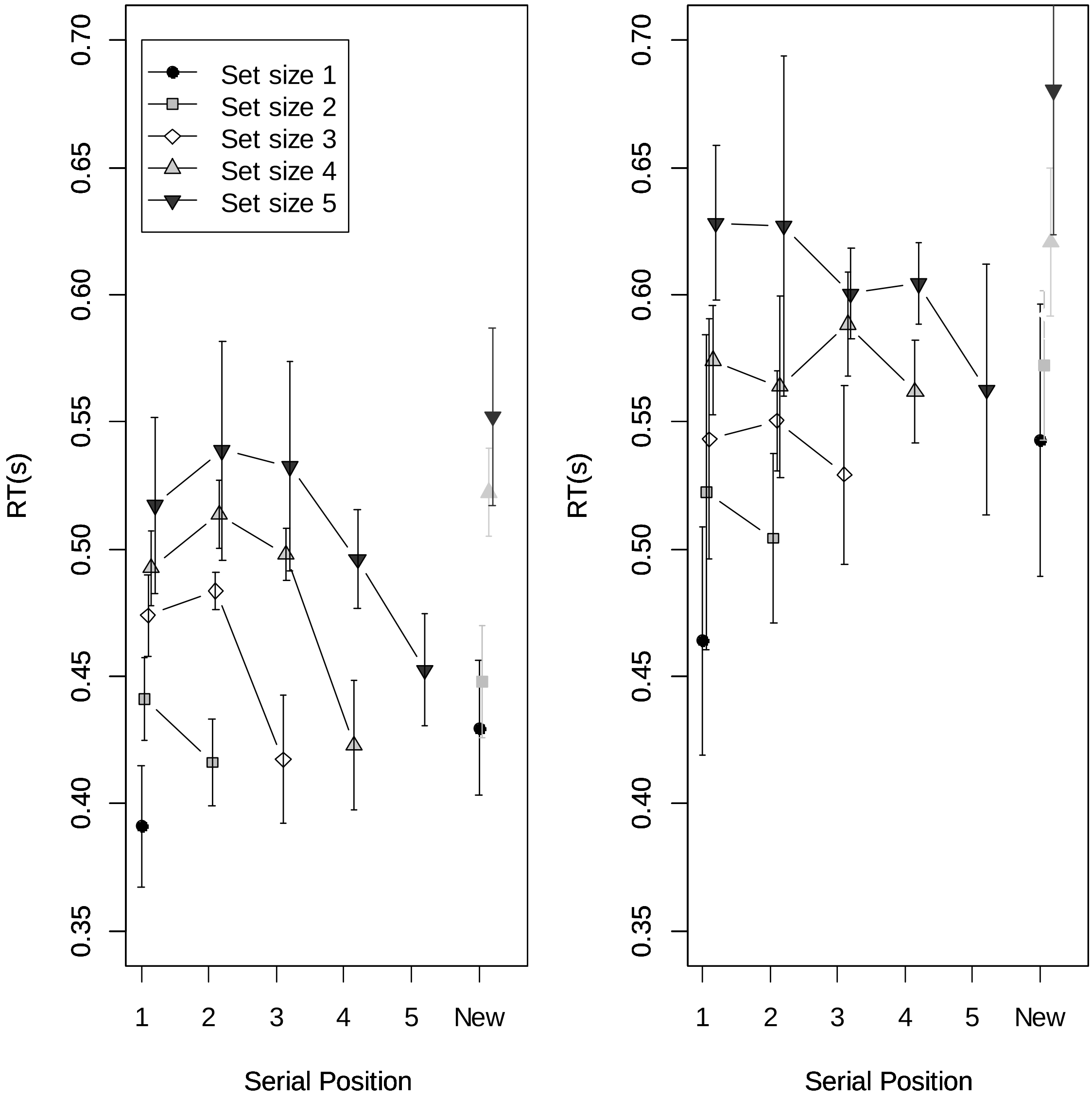

Figure 3: The effects of set size and serial position on response times in the Sternberg item-recognition paradigm, for fast presentation rate and short retention interval (left), and for slower presentation rate, longer retention interval, and serial recall of the list following the recognition decision (right) (Donkin & Nosofsky, 2012b). Response times for positive probes are plotted by their list position; those for new probes have no list position. Error bars are 95% confidence intervals for within-subjects comparisons (Bakeman & McArthur, 1996).

Recall. Effects of set size on response latencies have also been observed in recall tasks. For example, using a serial recall task, Maybery, Parmentier, and Jones (2002) had participants remember lists of between 3 to 6 items, and found that the time taken to produce the first response increased with set size. The time between each response also increased with set size. That is, when there were more items to remember, all responses were slowed by a roughly constant amount. This pattern has also been observed for complex span tasks (Towse, Cowan, Hitch, & Horton, 2008). In a free recall task, Rohrer (1996) showed that the time taken to recall items increased substantially when the number of items in the study list increased from 8 to 16 words. Further, participants were slower to recall words from longer lists, even when the same total number of words were recalled (i.e., the time to recall 5 words from an 8-item list was shorter than recalling 5 words from a 16-item list).

Other Tasks. Increasing the number of items to be remembered also increases response time in other WM tasks. In a working-memory updating task, updating an individual item is slower when participants have to hold more items in WM (Oberauer & Kliegl, 2001; Oberauer, Wendland, & Kliegl, 2003). Finally, more recent work in the area of visual WM has reported a monotonic increase in response time as a function of set size in change-detection tasks (Donkin, Nosofsky, Gold, & Shiffrin, 2013; Gilchrist & Cowan, 2014).

Benchmark 1.3: Number of Items Recalled or Recognized (B)

Probably the most basic folk question about WM concerns the number of items that can be remembered and then immediately reproduced, often referred to as *memory span*. We next consider the evidence speaking to this question, and what benchmarks can be distilled from it and from related recognition procedures.

Upper bounds on performance. A first benchmark states that in any test of WM using simple, highly discriminable stimuli, young adults can reliably recall or recognize across many trials *no more than 3 to 5 separate units* (Broadbent, 1975; Cowan, 2001). This is the number of units up to which accuracy remains very close to ceiling. Cardozo and Leopold (1963) showed virtually error-free serial recall of digits and letters up to 5 items. Crannell and Parrish (1957) showed nearly perfect serial recall of up to 5 digits, and 4 letters. Oberauer and Kliegl (2001) found perfect performance in arithmetically updating boxes containing numbers for a memory demand of no more than three boxes (Figure 2E). When recognition of simple visual objects is the measure, for brief arrays to be recognized there is ceiling-level performance with 3 or fewer items (Luck & Vogel, 1997; cf. Figure

2F). A similar limit on the number of items recalled is observed in free recall (Ward, 2002): With increasing list length, the number of words recalled increases less and less, reaching a level within Miller's range for span.

A related second benchmark pertains to the number of separate units that participants can reproduce on 50% of their trials – the classic definition of memory span. The famous paper by G. A. Miller (1956) suggested that healthy young adults have a span of about 7 separate, familiar items (e.g., letters, digits, or words), give or take a few. This generalization is based on a large number of memory span tasks administered across the years. Although there is variability in the span for different materials and testing procedures, the observed span limits for serial recall are sufficiently informative to qualify as a benchmark: Virtually all healthy adults have a span of more than 4 items in the serial recall of familiar verbal materials, and almost none have a span of more than 10 items.

Under many circumstances estimates of span fall considerably below 7 to 10. Inspired by Broadbent (1975), Cowan (2001) reviewed a large variety of situations in which the smaller limit of 3 to 5 items is the mean rather than the point of perfect performance. This is the case, for example, when the endpoint of the list is unpredictable, as in running memory span (Pollack et al., 1959), or when items are presented in a brief array (Sperling, 1960). Cowan's proposal was that there is a limit in span of 3 or 4 that can be overcome with the use of strategies such as grouping, chunking, and rehearsal unless the task prohibits such strategies. This limit was shown to range in adults between 2 and 6 items on average, and between 3 and 5 items in most of them. For one example, memory for spoken digit lists that were unattended when presented, see Cowan (2001, Figures 3-4, pp. 97-98); for another example, running span with a fast, 4/s presentation rate, see Figure 2B. Based on that proposal, we propose a further, conditional benchmark, outlined next.

A stronger benchmark of items in WM? Whereas the observed item limits vary substantially between materials and testing procedures, it is possible that there is an underlying invariant (cf. Rouder, Speckman, Sun, & Morey, 2009) – a relatively constant item limit that generalizes across materials and situations and predicts lower as well as upper bounds. To identify such an invariant, one must make several critical theoretical assumptions about cognitive processes that influence the observed item limits for specific materials and test situations, thereby explaining the variability on the surface. The rationale of this endeavor is analogous to the identification of constants in natural sciences. For instance, the constant of gravity holds despite variability in the observed rates of acceleration of falling bodies because the latter can be explained through auxiliary assumptions (e.g., assumptions about variability in air resistance; see Cowan, 2001). Following this rationale, we propose as a *conditional* benchmark that young adults can hold in WM about 3 to 4 chunks (Cowan, 2001). This benchmark is conditional because observation of the supporting evidence depends on measurement processes that rely on substantive assumptions, most notably about the nature of chunks in WM. We next make these assumptions explicit.

Chunking and structuring assumptions and their application to apparent exceptions to the item-limit benchmarks. As Miller (1956) proposed, the memory of sets of items is assisted by the recognition of item groups with strong inter-item associations, or chunks. This point is probably generally accepted in cognitive psychology (see Benchmark 11.1). For example, it is easier to remember a sequence of nine random letters if they can be grouped to form three known acronyms (which serve as chunks) such as *IRS, CIA*, and *FBI*. A more controversial tenet is that it is possible to identify chunks in a wide variety of situations, based on the following assumptions:

(1) Boundaries between chunks can sometimes be identified by long temporal gaps in recall sequences or changes in intonation at the end of a recalled chunk. With the help of this assumption, several studies have observed that people can recall about 3 to 4 chunks when reproducing the placement of chess pieces on a board (Chase & Simon, 1973b; Gobet & Simon, 1996).

(2) When people are free to rehearse verbal materials, or when they have a chance to group materials (see Benchmark 9.2), then they can use these processes to integrate several familiar items (e.g., several digits or words) into a chunk. Therefore, researchers must prevent rehearsal and grouping processes so that each presented, familiar item remains an individual chunk (Broadbent, 1975; Cowan, 2001). One way of doing so is to present spoken items and ask for serial recall during concurrent articulation. With this procedure, young adults can recall about 3 or 4 items (for a review see Cowan, 2001, Table 2).

(3) Conversely, for some complex materials a single nominal item might have to be represented by more than one chunk if the elements cannot be easily integrated. For example, a multi-syllable non-word is arguably represented not as one chunk, but as several chunks (e.g., one for each syllable). Consequently, for complex materials the number of items that can be recalled or recognized may fall substantially below 3. Therefore, to measure the chunk limit, researchers must use simple items that are unambiguously familiar to participants as single units, so that they can encode each item as a single chunk. Alternatively, researchers can use some means to ascertain that separate elements have been successfully combined to form larger chunks, as explained next.

(4) Chunks can be identified by presenting materials consisting of several elements that people have learned to integrate into units, such as known proverbs, so that each long unit can be assumed or shown to act as a single chunk (Glanzer & Razel, 1974; Simon, 1974; Tulving & Patkau, 1962).

(5) Chunks can also be identified by creating them experimentally, for instance by teaching inter-item associations to create new chunks, possibly assessing the integrity of these chunks with a cued-recall test. With this method it was again found that young adults can recall or recognize about 3 to 4 chunks (Chen & Cowan, 2009; concurrent-articulation condition shown in Figure 4; Cowan, Chen, & Rouder, 2004; Cowan, Rouder, Blume, & Saults, 2012).

The item limit we propose as conditional Benchmark 1.3 builds on evidence from several paradigms and materials, and it is of central importance for theorizing about the nature of the capacity limit of WM. At the same time, because of its conditional nature its status is less certain than that of other benchmarks that are less dependent on measurement assumptions. Therefore, in agreement with the modal rating of survey respondents for most of the instances of this benchmark, we rate it as B.

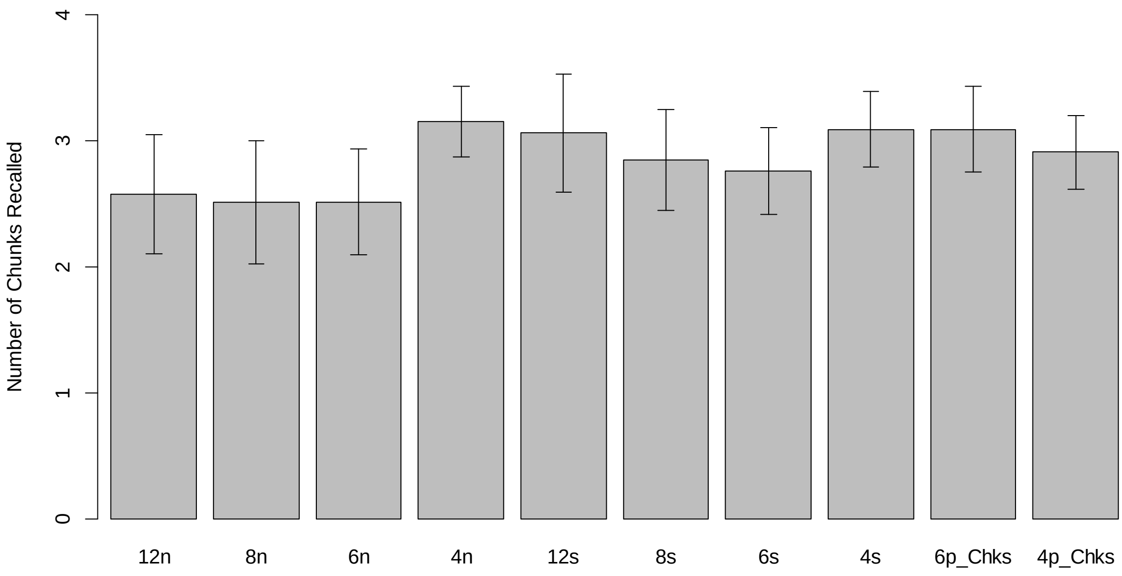

Figure 4: Number of chunks recalled (regardless of serial order) in the experiment of Chen and Cowan (2009). Condition labels indicate the number of chunks presented, followed by the condition: New single words (n), single words presented during pre-training (s), and word-pairs learned as chunks during pre-training (p_Chks); in the word-pairs condition each pair counts as one chunk. Error bars are 95% confidence intervals for within-subjects comparisons (Bakeman & McArthur, 1996).

Boundary Conditions. Experts in forming chunks of a certain kind can learn to recall lists of many more items than the usual 7 or so, e.g., 80 digits or more in a list (Ericsson et al., 1980, 2004; Wilding 2001). This kind of finding has been explained not completely on the basis of simple chunking, but with the further assumption of a hierarchical organization in which chunks of 3 to 5 items are organized within about 3 to 5 super-chunks, and so on (see also Benchmark 11.1). There is good support for the first-order chunks, in the form of pauses in recall between chunks, and super-chunks, in the form of falling intonation at the end of a super-chunk (Ericsson, Chase, & Faloon, 1980). However, whereas the formation of chunks of digits (e.g., by replacing short sequences of digits by known athletic record times) is well explained, it is not clear what long-term memory information allowed the formation of super-chunks.

One additional condition for observing an upper bound of 3 to 4 chunks reliably across trials is that the items should be from a common category; the proposed benchmark might not apply, for example, in the recall or recognition of stimulus sets that include verbal and nonverbal items together. Memory for these sets might exceed the limit for more uniform sets, presumably because some mechanism helps to keep the subsets of items separate and limit interference between different item types (Baddeley & Hitch, 1974; Cowan, Saults, & Blume, 2014). This situation might be considered another one in which chunking and structure contribute to performance by allowing fewer chunks than the number of stimuli.

2: The Effects of Retention Interval and Presentation Duration

Most memories are eventually forgotten. The benchmarks in this section pertain to the time course of forgetting over varied retention intervals (Benchmarks 2.1, 2.2, 2.3) and to another time-related effect, that of varying the presentation duration or presentation rate at encoding (Benchmark 2.4).

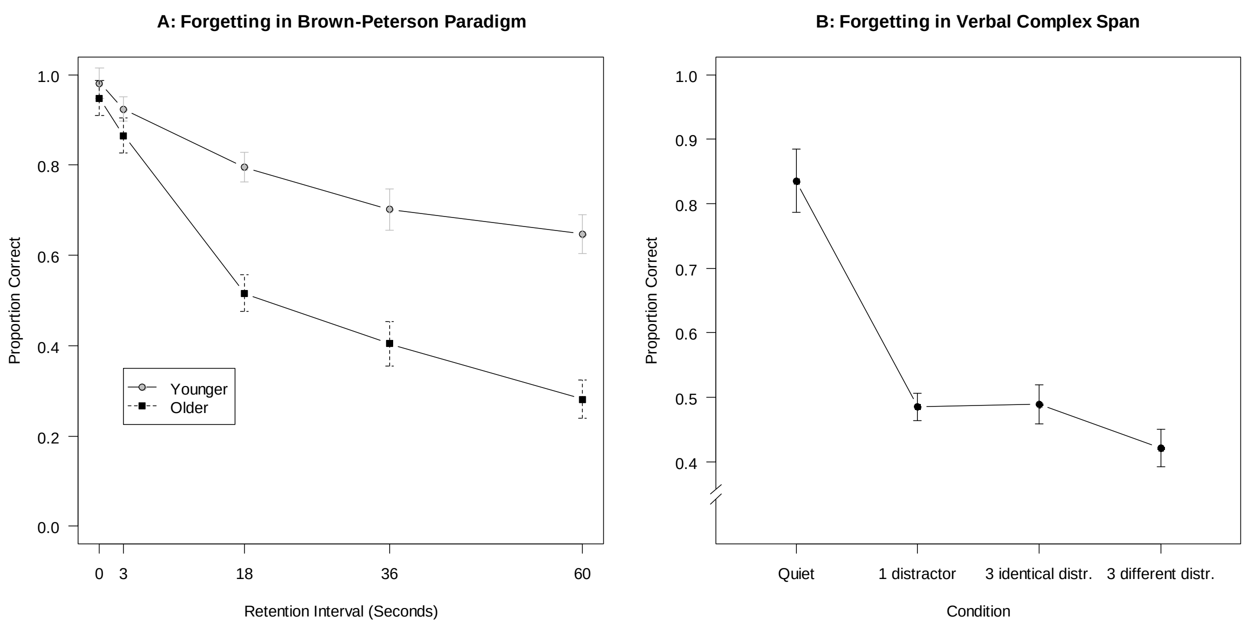

Figure 5: Forgetting (i.e., loss of memory accuracy) as a function of the duration of, and the events in, the retention interval. A: Forgetting as a function of the duration of a filled retention interval in the Brown-Peterson paradigm; data for a group of younger and a group of older adults (Floden, Stuss, & Craik, 2000, Exp. 2) B: Forgetting in serial recall as a function of distractor processes interleaved

between encoding of items: none (Quiet), reading a single word (1 distractor), reading the same word three times (3 identical distr.) or reading three different words (3 different distr.) (Lewandowsky, Geiger, Morrell, & Oberauer, 2010, Exp. 2).

Benchmark 2.1: The effects of filled retention intervals (A)

When list presentation is followed by retention intervals of varying lengths during which people engage in a distracting activity that prevents rehearsal, performance typically declines as the retention interval is increased. The effect is particularly robust and pervasive for verbal materials.

The classic studies of J. Brown (1958) and Peterson and Peterson (1959) show that recall performance declines over time when people engage in a distractor task that involves constantly-changing materials, such as counting backwards from a random number. In most instances, steep forgetting is observed for the first 15-18 seconds of distracting activity, followed by a performance plateau during which forgetting is considerably slower or even absent (see Figure 5A). A decline of memory over time is also observed when the time between list items at encoding or the time between retrievals of individual items is increased, as long as that time is filled with a distractor task involving changing materials (Lewandowsky et al., 2010; Lewandowsky, Geiger, & Oberauer, 2008). This effect is a classic, well-replicated finding that has played a major role in the discussion about the causes of forgetting in working memory (for reviews see Oberauer, Farrell, Jarrold, & Lewandowsky, 2016; Ricker, Vergauwe, & Cowan, 2016).

For visual-spatial materials, there is also evidence that memory performance declines as a distractor- filled retention interval is increased. For example, Ricker and Cowan (2010) showed that memory for unconventional symbols (e.g., ϡ, ℑ) that defy verbalization decreases as people spend more time on a verbal distracting activity (e.g., determining whether a spoken digit is odd or even) after list presentation. Similarly, Kopelman and Stanhope (1997) found a decline in Corsi-block performance when a distractor-filled retention interval was extended from 5s to 15s, and Meudell (1977) found that memory for matrix patterns declines over a retention interval filled with backward counting. There is, however, at least one exception to this general pattern: Christie and Phillips (1979) reported a study in which participants had to memorize random matrix patterns. Although a distractor task during the retention interval (counting backwards by 3's from a random number) lowered performance overall compared to an unfilled control, the duration of the distractor task had no effect on performance. However, we are not aware of any replications of this result.

When considered across both stimulus domains and across the preponderance of results, the effects of filled retention intervals are sufficiently clear and robust for us to rate it as A, in agreement with the majority of survey respondents (19/34).

By contrast, we do not consider the results with *unfilled* retention intervals (i.e., intervals during which the participant is not engaged in any experimenter-directed activity) sufficiently consistent and unambiguous to warrant a benchmark. Although extending an unfilled retention interval after study of verbal material generally does not lead to a reduction in performance (e.g., Oberauer & Lewandowsky, 2016; Ricker & Cowan, 2010; Vallar & Baddeley, 1982; but see Ricker, Spiegel, & Cowan, 2014), the pattern is considerably more ambiguous with visual and spatial information (e.g., colored shapes). On the one hand, extending an unfilled retention interval leads to a further decline in performance (e.g., Mercer & Duffy, 2015; Ricker & Cowan, 2010; Ricker, Spiegel, & Cowan, 2014; Sakai & Inui, 2002). On the other hand, there is a substantial number of reports in which no decline of performance with additional unfilled retention time is observed (Burke, Poyser, & Schiessl, 2015; Gorgoraptis, Catalao, Bays, & Husain, 2011; Kahana & Sekuler, 2002; Smyth, Hay, Hitch, & Horton, 2005), or a decline is observed in one condition and not in another (Lilienthal, Hale, & Myerson, 2014). Because there is still much uncertainty about the conditions under which memory for visual and spatial materials does or does not decline over an unfilled retention interval, we do not believe that the effects of unfilled retention intervals can be described as a benchmark result.

Benchmark 2.2: The interaction of retention interval with proactive interference (B)

The effects of retention interval are much attenuated when proactive interference (PI)[6] is

[6] Proactive interference refers to the finding that memory declines over successive trials in which materials from the same category are studied; after a change of category, release from proactive interference is observed (Gardiner, Craik, & Birtwistle, 1972)

absent, as for example on a participant's first trial in a distractor task (e.g., Keppel & Underwood, 1962; Loess, 1964; Meudell, 1977). However, when studies are sufficiently powerful, small but significant forgetting during a retention interval exceeding 3 seconds can be observed even in the absence of proactive interference (Baddeley & Scott, 1971). As the attenuation of time-based forgetting in the absence of PI qualifies Benchmark 2.1, and has been observed only with one paradigm and primarily with verbal materials, we rate it as B. This was also the modal rating in the survey (14/30).

Benchmark 2.3: The interaction of forgetting with the type of distractor material (B)

There is considerable evidence that extending the duration of a retention interval has little or no effect on performance when no new information is processed during the retention interval. That is, the time-dependent forgetting captured by Benchmarks 2.1 and 2.2 only applies when the distractor task involves changing-state material such as counting backwards from a random number. When the distractor-task material remains unchanged, distraction still impairs verbal memory (see Benchmark 5.2), but the amount of forgetting does not increase with a longer retention interval. For example, Vallar and Baddeley (1982) showed that recitation of the word "the" for 15 s after presentation of a trigram of consonants had little effect on memory. Similar results have been reported by Longoni, Richardson, and Aiello (1993), Phaf and Wolters (1993), and Humphreys et al. (2010). Likewise, repetition aloud of the same word in between recall attempts does not impair performance appreciably, even if an additional 12 s has elapsed during retrieval of a list (Lewandowsky, Duncan, & Brown, 2004; Lewandowsky, Geiger, et al., 2008). Performance remains unaffected by retention interval even when a speeded choice task has to be performed in addition to repeating a constant distractor out loud (Oberauer & Lewandowsky, 2008). A parallel pattern of results is obtained when distractor tasks are inserted in between items at encoding (Lewandowsky et al., 2010; Oberauer & Lewandowsky, 2008, 2013). This finding, illustrated in Figure 5B, qualifies Benchmark 2.1. It has been replicated with several variants of the serial-recall paradigm, but is currently limited to verbal materials; hence we rate it as B. Survey respondents mostly agreed with this assessment (21/63 B ratings, and about equally many A and C ratings).

Benchmark 2.4: The effects of presentation duration (B)

For simple visual materials such as colors or orientations, performance increases across a very narrow and brief range of presentation durations. No further increases of accuracy are observed beyond a presentation duration of around 50-100 ms per item (Bays, Gorgoraptis, Wee, Marshall, & Husain, 2011; Vogel, Woodman, & Luck, 2006); see Figure 6 (left panel).

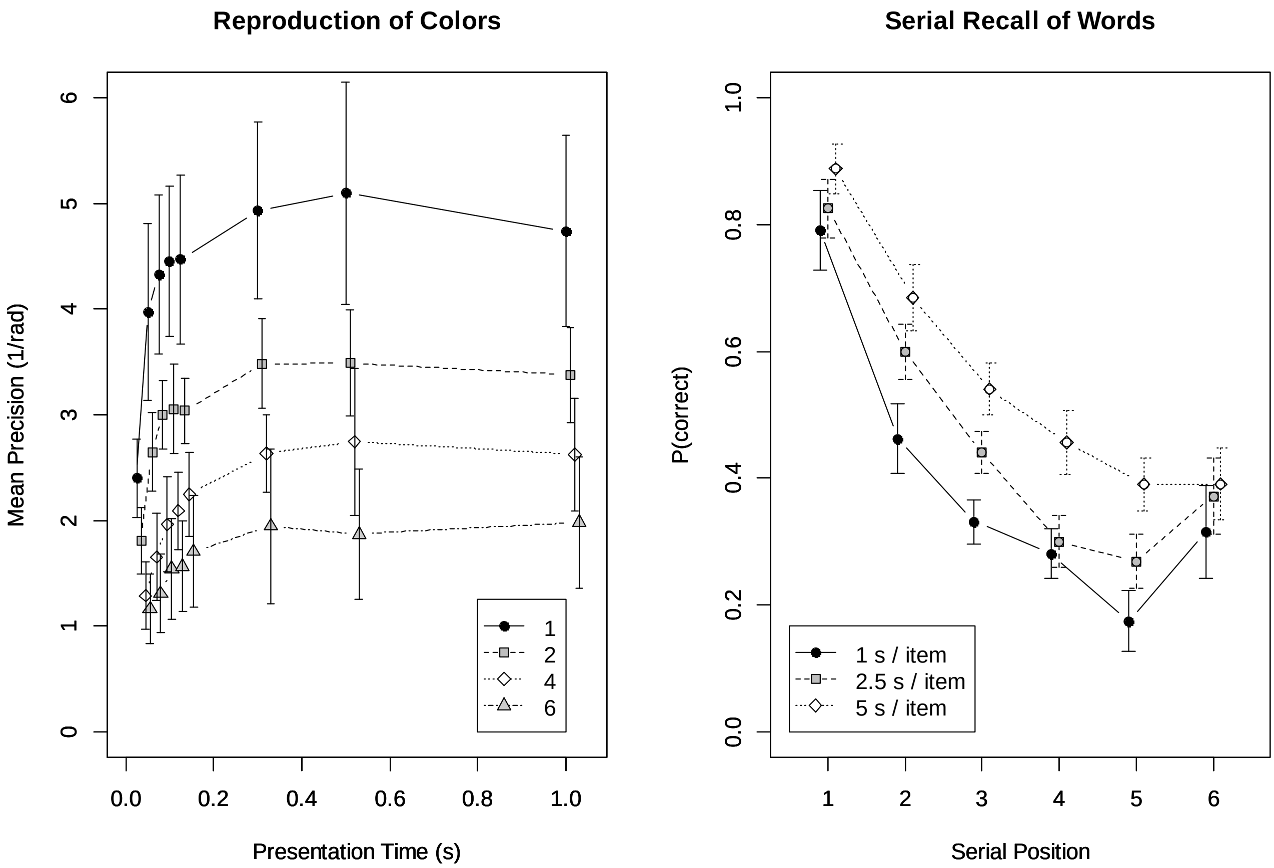

Figure 6: Effects of presentation time on short-term retention. Left: Precision of reproduction of colors from an array as a function of presentation time between onset of the array and onset of a mask, and set size (Bays et al., 2011). Right: Accuracy of serial recall of words as a function of presentation rate and serial position (Tan & Ward, 2008).

With verbal material, the effects of presentation duration are also positive for free recall (Glanzer & Cunitz, 1966; Roberts, 1972), albeit over a longer range of times than for visual material. With serial recall, the presentation modality matters. There is ample evidence that slower presentations give rise to better serial-recall performance for visually presented lists of verbal items (Dornbush, 1968a, 1968b, 1969; Fell & Laugherty, 1969; Mackworth, 1962, 1964; D. J. Murray, 1965, 1966; Norman, 1966; Tan & Ward, 2008). For example, Tan and Ward (2008) observed that when presentation duration is extended from 1 s/word to 5 s/word, serial-recall performance for a 6-item list increased by about 20 percentage points for all but the first item (Figure 6, right panel).

With auditory presentation, by contrast, no clear summary is possible because every possible outcome has been observed. In some cases, slower presentation has improved performance (e.g., Fell & Laugherty, 1969; Gerver, 1969), in some cases a null effect of presentation duration has been observed (e.g.,D. J. Murray & Roberts, 1968), and in other instances performance declined as presentation was slowed (e.g.,Dornbush, 1969; Mackworth, 1964).

In sum, the beneficial effect of increasing presentation duration is sufficiently robust and general to deserve benchmark status; at the same time the effect varies considerably across materials and is qualified by presentation modality; therefore, we assigned it only intermediate priority (B).

3: Effects of Serial Position in Lists

Benchmark 3.1: Primacy and Recency Effects on Accuracy (A)

The relationship between accuracy of retrieval for an item and its position in the experimenter's list is known as the *serial position curve*. For lists long enough so that accuracy is less than perfect, there are recall advantages for those items presented at the start of the list, called the *primacy effect*, and at the end of the list, called the *recency effect*.

Primacy effects and recency effects can be found in most, if not all, immediate memory tasks (for some examples see Figure 7). Considering first the recall of verbal lists, both effects are observed in immediate free recall (Murdock, 1962), immediate serial recall (Drewnowski & Murdock, 1980; Murdock, 1968a), backward serial recall (Madigan, 1971) and probed serial recall using either the serial position or the prior item as probe (Murdock, 1968a). Primacy and recency effects are also observed in recognition tests (Oberauer, 2003b) when participants are asked to identify whether a probe item has been presented in a particular list position (local recognition) or in a particular list (global recognition). Finally, they are observed in the reconstruction of order task (see Table 1), in which participants are simultaneously re-presented at test with all the list items (in a new visuo-spatial arrangement), and must select the items in the correct serial order (Lewandowsky, Nimmo, & Brown, 2008; Nairne, Neath, & Serra, 1997).

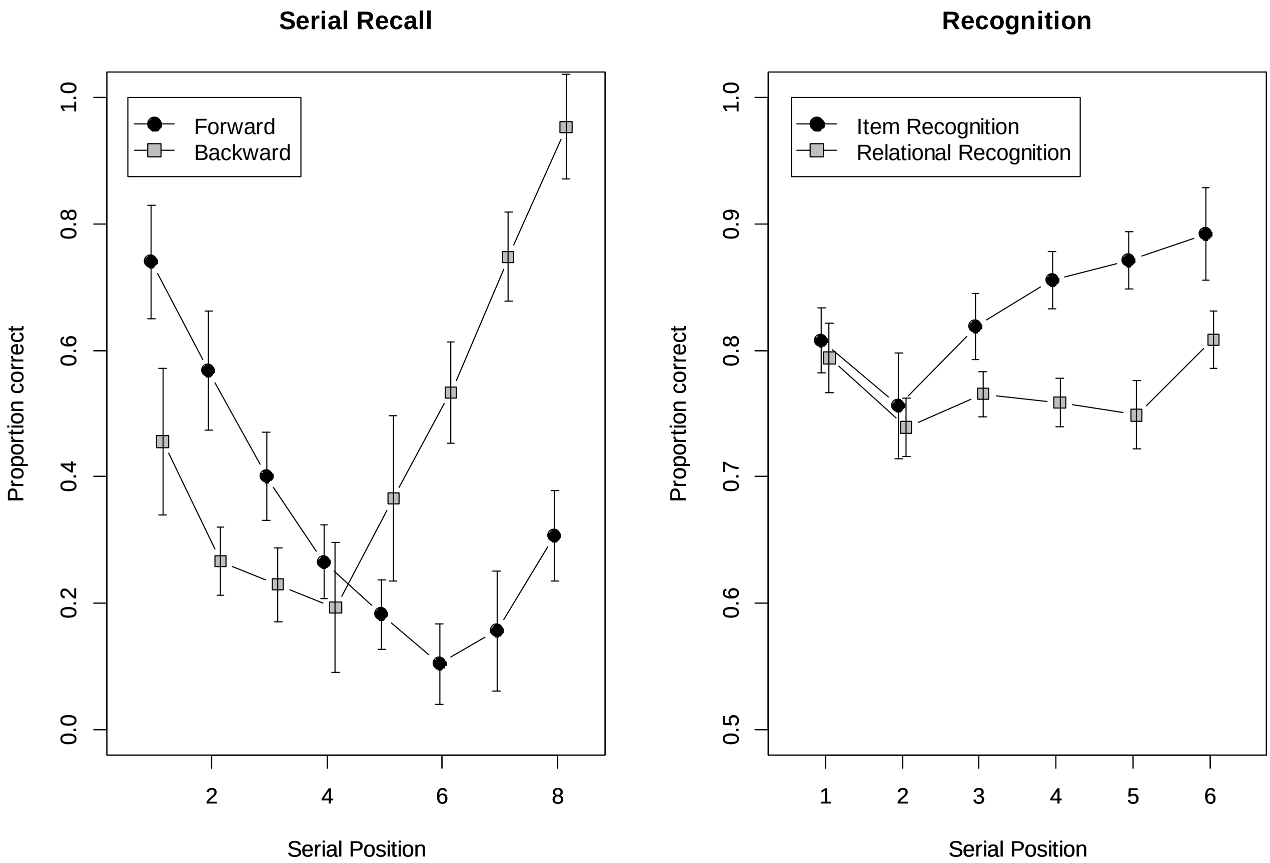

Figure 7: Serial position curves from forward and backward serial recall (Madigan, 1971; data for visual presentation modality), and from item and relational recognition (Oberauer, 2003b; random presentation order)

The relative strength of primacy and recency is strongly modulated by output order (Grenfell-Essam & Ward, 2012; Madigan, 1971; Ward, Tan, & Grenfell-Essam, 2010). Items are more likely to be recalled correctly if their retrieval is attempted early in the recall period (see Benchmark 3.4.1 on output order effects). Yet, primacy and recency effects in recall and recognition are found across input positions even when the output order is controlled (Cowan, Saults, Elliott, & Moreno, 2002; Oberauer, 2003b).

Recency but not primacy in free recall is eliminated by a filled distractor at the end of the list (Glanzer & Cunitz, 1966; Postman & Phillips, 1965). However, both primacy and recency effects are observed when a period of distractor activity is interleaved between each and every list item in the continual distractor free recall task (Bjork & Whitten, 1974) or the complex span task (Unsworth & Engle, 2006b).

Primacy and recency effects are also observed with non-verbal materials (for a review see Hurlstone, Hitch, & Baddeley, 2014): They have been obtained in the immediate serial recall (Jones, Farrand, Stuart, & Morris, 1995) and immediate free recall (Cortis et al., 2015) of visuo-spatial locations, and in the reconstruction-of-order task for non-verbal items such as block matrices or unfamiliar faces (Avons, 1998; Smyth et al., 2005).

We rated this benchmark as A because its high degree of generality, and because it is highly diagnostic for theories of WM. Serial-position effects have long been accepted as benchmarks by theorists. Because all contemporary models of list recall explain it at least qualitatively (Farrell & Lewandowsky, 2004), it does not serve to adjudicate between competing models but rather represents a minimum requirement: Any comprehensive model that fails to predict primacy and recency effects is not viable from the start. The Benchmark A status for primacy and recency effects is echoed by the majority (13) of the 19 ratings in the survey.

Moderators of Primacy and Recency. In addition to distractor tasks (Benchmark 3.1), output order (Benchmark 3.4), and presentation modality (Benchmark 3.2), a number of other variables have been observed to affect the relative strength of primacy and recency: Instructions inducing test expectancy, list length, and scoring method. Because none of these findings is sufficiently general, well established, and theoretically informative, we do not assign them benchmark status but rather mention them as moderators of Benchmark 3.1.

Instructing participants before list presentation about how they will be tested has sometimes been found to affect the serial-position curve (for forward recall vs. recognition instructions see Duncan & Murdock, 2000; for forward vs. backward recall instructions see Neath & Crowder, 1996). In contrast, the serial-position curves of forward serial recall and free recall are hardly affected by whether participants are informed before list presentation which of these two recall tasks they will have to do on a given list (Bhatarah, Ward, & Tan, 2008).

The effects of list length have been examined in a series of experiments using immediate free recall, variants of immediate serial recall, and the reconstruction of order tasks (Grenfell-Essam & Ward, 2012; Ward et al., 2010). Increasing the list length reduced the magnitude of the primacy effect but had little effect on the magnitude of the recency effect.

Benchmark 3.2: The Modality Effect and its Interaction with Recency (B)

There are enhanced recency effects in many immediate verbal memory tasks with spoken stimuli relative to silently read visual stimuli (Conrad & Hull, 1968; Gardiner & Gregg, 1979; Murdock & Walker, 1969; Watkins, Watkins, & Crowder, 1974), a finding known as the modality effect (see Figure 8). The modality effect is mostly limited to verbal materials, although there is one study showing elevated recency (and poorer memory in earlier list positions) in serial-order memory for spatial locations (Tremblay, Parmentier, Guérard, Nicholls, & Jones, 2006). The modality effect is observed across a wide range of immediate memory tasks (Penney, 1989), including immediate free recall (Murdock & Walker, 1969), reconstruction of order (Neath, 1997), and continuous-distractor free recall (Gardiner & Gregg, 1979; Glenberg, 1984). The modality effect is reduced when recall is spoken compared to when it is written (Harvey & Beaman, 2007).

Modality effects have long played an important role in some, though not all, theories of working memory. Some theorists attribute them to the output of a separate, pre-categorical acoustic store (Crowder & Morton, 1969; C. Frankish, 2008) or an echoic memory (Cowan, 1999). Other theorists attribute modality effects to differences in item coding (Nairne, 1990; Neath, 2000) or differences in perceptual, attentional, and speech motor processing (Jones, 1993; Macken, Taylor, Kozlov, Hughes, & Jones, 2016). Still others argue that modality effects lie outside of the scope of working memory (Baddeley, 1986). It is perhaps not surprising that survey respondents were split between rating it as A (8/19) or C (6/19). Because the modality effect has substantial theoretical leverage for many, though not all, theories of working memory, and it is limited mostly, though not exclusively, to verbal materials, we consider it as a benchmark B.

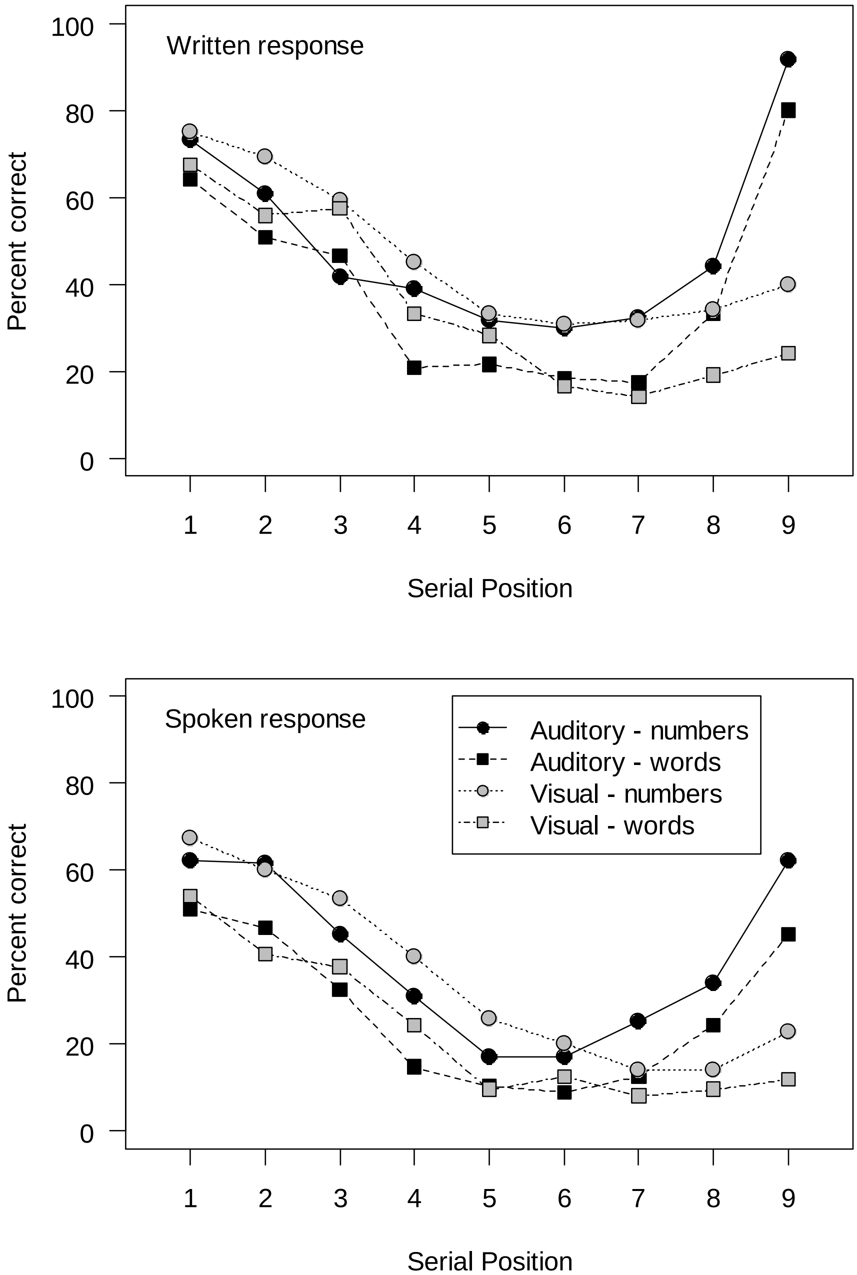

Figure 8: Modality Effect: Serial position curves for forward serial recall of words or numbers presented visually or auditorily, with written or spoken response modality (Harvey & Beaman, 2007)

3.3: Effects of Serial Position on Retrieval Latency

Benchmarks 3.3.1. Serial-Position Effects on Recognition Latencies (B). Recency effects are strong in short-term recognition tasks, such that observers are quicker to identify a match between more recently presented study items and a test item. For this reason, response times are often plotted as a function of the *lag* between the study and test item, where the most recently presented item has a lag of 1. In general, response times become slower with increasing lag between study and test items (Corballis, 1967; Forrin & Morin, 1969; McElree & Dosher, 1989; Monsell, 1978). However, the benchmark result is not a monotonic decrease in response time with study-test lag. Rather, there are usually small primacy effects, such that the first and second study items are usually responded to more quickly than would be expected if only recency were operating (Donkin & Nosofsky, 2012a). The left panel of Figure 3 is representative of the typical serial-position (or study-test lag) effect on mean response time. The serial-position effects on recognition latencies are well replicated, and they have theoretical leverage because they question the serial-scanning model of Sternberg (1969). At the same time, they have so far been found only with verbal materials, and they are observed only for relatively fast timing conditions (see Benchmark 1.2); due to their limited generality we rate them as B.

Benchmark 3.3.2: Particularly Fast Access to the Last List Item (C). This benchmark is described in Appendix B.

Benchmark 3.3.3. Serial-Position Effects on Recall Latencies (C). This benchmark is described in Appendix B.

3.4: Effects of Output Order

We rated this set of benchmarks as B throughout. Although we consider these benchmarks to be potentially diagnostic, they have not received the extensive theoretical treatment we felt was necessary to be classed as an A benchmark. Our rating agrees with the majority ratings in the survey (see Appendix A).

Benchmark 3.4.1: Effects of Output Order on Accuracy (B). Complementary to the effect of input serial position on accuracy (Benchmark 3.1), output order is also found to systematically affect accuracy. Generally, accuracy monotonically declines across output positions. In order to deconfound input and output position in serial recall, Cowan et al. (2002) used a 'wrap-around' procedure, where participants were retrospectively cued to serially recall starting at position *N*, and then to serially recall from the beginning of the list up to position *N*-1. For a fixed input position, performance was found to decline as a function of the number of items already recalled, particularly for visually presented items. A decrease in accuracy across output position has also been observed when dissociating input and output order in probed recall (Oberauer, 2003b), cued recall using paired associates (Tulving & Arbuckle, 1966), and item recognition (Oberauer, 2003b). In free recall, asking people to begin their recall with a target portion of the list impairs later free recall of the remaining list items (Dalezman, 1976).

Benchmark 3.4.2: Effects of Output Order on Retrieval Latency (B). In probed recall, latencies decrease over output position (Oberauer, 2003b). The same trend is found in local recognition (Lange et al., 2011; Oberauer, 2003b). In contrast, free recall typically produces latencies that increase in an accelerating fashion across output positions (Murdock & Okada, 1970), well fit by a hyperbolic function (Rohrer & Wixted, 1994).

Benchmark 3.4.3: Effects of Output Contiguity (B). Many WM tasks have revealed a benefit of probing items in the order in which they have been encoded. In probed recall, the recall of item *N* is facilitated by the preceding recall of an item at an input position preceding *N*, particularly item *N-1* (Nairne, Ceo, & Reysen, 2007), and recall is more accurate if items are probed in the same order as the original presentation (Oberauer, 2003b). In local recognition, RTs are faster when items are probed in forward order than when probed in random order, and in addition, probing in forward order eliminates the set-size effect on recognition latencies (Lange et al., 2011).

Free recall shows contiguity effects on people's preferred output order. The *lag-recency* effect (Kahana, 1996) refers to the observation that having recalled an item from serial position *N*, people are most likely to recall a nearby item next, typically item *N*+1. Analysis of the lag-recency effect takes into account the more numerous opportunities to make transitions over smaller distances. This effect is also observed in free reconstruction of order (Lewandowsky, Brown, & Thomas, 2009). In addition, latencies in free recall are shortest for transitions from item *N* to *N*+1 (Kahana, 1996).

3.5: Self-Chosen Output Order in Free Recall

Benchmark 3.5.1: First-Recall Probability and Its Interaction with Serial-Position Effects (B). When participants are free to recall a list in any order, they tend to initiate recall of short lists with the first list item (Grenfell-Essam & Ward, 2012; Ward et al., 2010). However, as the list length increases, participants tend to initiate recall with one of the last four list items (Hogan, 1975; Howard & Kahana, 1999; Laming, 1999); see Figure 9. The choice of where to initiate recall has consequences for the serial-position curve: When recall is initiated with the first list word, there is elevated recall of the early list items and a reduced recency effect, whereas when recall is initiated with one of the last four words, there is an extended recency effect and a greatly reduced primacy effect (Grenfell-Essam & Ward, 2012; Ward et al., 2010). The relationship between first recall, list length, and serial position was ranked as B because it is a robust finding that qualifies the theoretically highly informative serial-position effects, but at the same time it is limited to the free-recall paradigm. Survey data on this benchmark are sparse but supportive (4/7 for A).

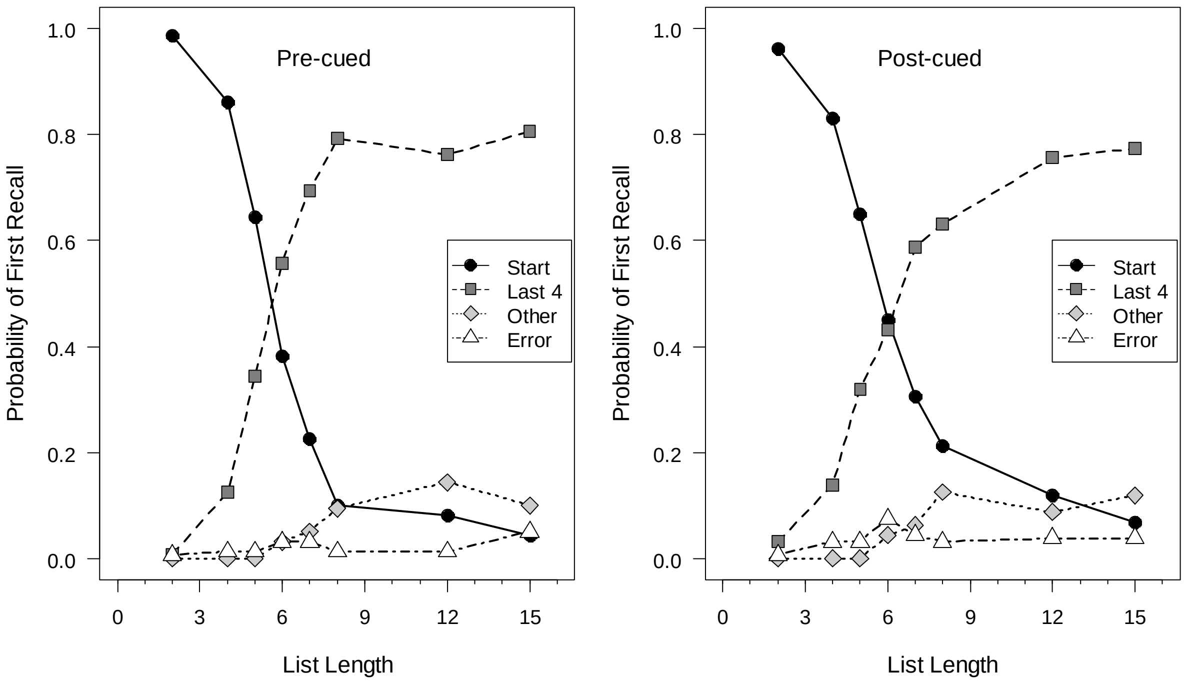

Figure 9: Probability of first recall in free recall of words: The probability that the first recalled word comes from the beginning of the list (Start), from the last four presented words (Last 4), from another list position (Other), or of being an extra-list intrusion (Error), as a function of list length. Recall method (free vs. serial) was cued before list presentation (left panel) or after list presentation (right panel) (Grenfell-Essam & Ward, 2012)

Benchmark 3.5.2: Semantic Clustering in Free Recall (C). This benchmark is described in Appendix B.

4: Characteristics of Errors

The analysis of recall errors has long been recognized as a particularly diagnostic approach. For example, in serial recall error analyses have proven fruitful in adjudicating between competing mechanisms for the representation of serial order, which are prone to generating different types of error profiles (e.g., Henson, 1998b).

Benchmark 4.1. Confusions of target item with other items in memory set (A)

In various WM tasks, errors often involve the confusion of the target item with other items in the memory set. In serial recall (Aaronson, 1968; Guérard & Tremblay, 2008; Henson, Norris, Page, & Baddeley, 1996; Smyth et al., 2005) and probed recall (Fuchs, 1969), these errors take the form of transpositions, which are items from the study sequence recalled in wrong positions. In local recognition (Oberauer, 2003b) and change detection (Donkin, Tran, & Pelley, 2015; Wilken & Ma, 2004) tasks (see Table 1), these errors occur in the form of increased false alarm rates, and slowed rejection latencies, to lure probes that match non-target items from the memory set, compared to lures not matching any item in the memory set. In the continuous-reproduction task confusion errors are represented by a tendency to respond with the feature of a non-target item from the current array (Bays, Catalao, & Husain, 2009).

Benchmark 4.1.1. Locality constraint on transpositions (A). In serial recall, most transpositions are movements of items to positions adjacent to their target position. The left panel of Figure 10 plots the proportion of transpositions as a function of absolute distance between an item's correct position and the position in which it was erroneously reported. The figure shows the typical transposition gradient, characterized by a decrease in transpositions as distance between output position and input position of the recalled item increases. This tendency for transpositions to cluster around their correct positions is known as the *locality constraint* (Henson et al., 1996; Hurlstone & Hitch, 2015; Lee & Estes, 1977; Nairne, 1991; Smyth et al., 2005). The locality constraint also manifests in the probed recall task (Fuchs, 1969), and in the *n*-back task in the form of increased false alarms to *n*-1 and *n*+1 lure probes (Szmalec, Verbruggen, Kemps, & Vandierendonck, 2011). A locality constraint over the spatial—as opposed to temporal—distance between items has been witnessed in visual WM tasks (Bays, 2016; Emrich & Ferber, 2012; Rerko, Oberauer, & Lin, 2014), a probed recall task (Hitch, 1974), and a reconstruction of order task for sound-specified locations (Groeger, Banks, & Simpson, 2008); the spatial transposition gradient is illustrated in the right panel of Figure 10. In sum, the locality constraint generalizes across several paradigms and materials, and has been identified as a key finding that any model of serial-order memory must predict (G. D. A. Brown, Preece, & Hulme, 2000; Henson, 1998b; Page & Norris, 1998). Therefore, we rate it as A. Survey responses were broadly distributed (10/30 for A, 9/30 for B, 7/30 for C). This benchmark is qualified by Benchmark 4.1.2 (Appendix B).

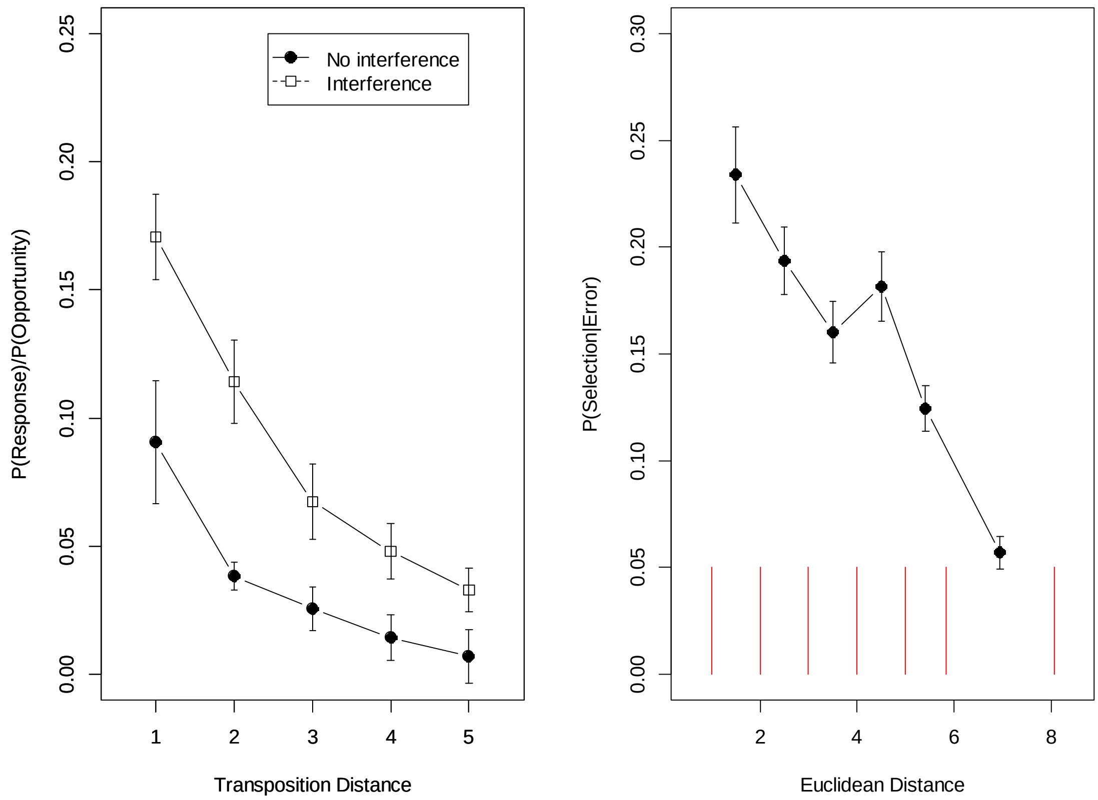

Figure 10: The locality constraint. Left: Transposition errors in serial recall are more likely for shorter transposition distances, after correcting for the different numbers of opportunities for different transposition distances (Farrell & Lewandowsky, 2004; No interference = immediate recall; Interference = recall after reading aloud four distractor digits). Right: The probability of selecting a non-target color in probed recall of color arrays decreases with the Euclidean distance between the target and the non-target in the array. Euclidean distances were sorted into six bins, with bin boundaries

(red vertical bars) chosen such that error frequencies would be equal across bins if non-targets were selected at random. Observed error probabilities are plotted for each bin. (Rerko et al., 2014)

Benchmark 4.1.2. Fill-in effect in serial recall (C). This benchmark is described in Appendix B.

Benchmark 4.2. Serial position effects on error-types in serial recall (C).

This benchmark is described in Appendix B.

Benchmark 4.3. Intrusions from previous memory sets (B)

Tests of immediate memory are subject to intrusions from previous memory sets. In recognition tests negative probes that have been memory items on recent trials (so-called recent negative probes) are slower and harder to reject than negative probes that have last been included in less recent trials (Atkinson, Herrmann, & Wescourt, 1974; Berman, Jonides, & Lewis, 2009; Jonides, Smith, Marshuetz, Koeppe, & Reuter-Lorenz, 1998). The effect is strongest for negative probes from the immediately preceding trial and gradually decreases in strength as negative probes are taken from temporally more distant trials (Berman et al., 2009; Hartshorne, 2008).

In immediate serial recall there is a tendency of erroneously recalling items from lists of recent earlier trials. These so-called *protrusions* tend to be recalled in the same list position in which they had been presented on the earlier list, or in a close-by list position (Drewnowski & Murdock, 1980; Fischer-Baum & McCloskey, 2015; Quinlan, Neath, & Surprenant, 2015).

Intrusion errors from previous trials are a benchmark because they are robust, generalize across paradigms, and are theoretically informative: They speak to the mechanisms of interference in WM, and because they reflect how long information encoded into WM remains in some form of memory even when no longer needed. That said, with the exception of Hartshorne (2008), the effect has only been demonstrated for verbal materials. Therefore, we rate this benchmark as B, in agreement with the survey ratings (recent-negative probes: 2/8; protrusions: 21/60).

Benchmark 4.4. Ranschburg effect in serial recall (C).

This benchmark is described in Appendix B.

Benchmark 4.5. Error distributions on continuous response scales (B)

The adaptation by Wilken and Ma (2004) of the psychophysical method of adjustment (Gescheider, 1997; Prinzmetal, Amiri, Allen, & Edwards, 1998) to the study of WM is known as *continuous reproduction* or *delayed estimation* (see Table 1 and Figure 1I). In this paradigm, the observer adjusts a feature value of an item until it matches the corresponding feature value of a remembered item.

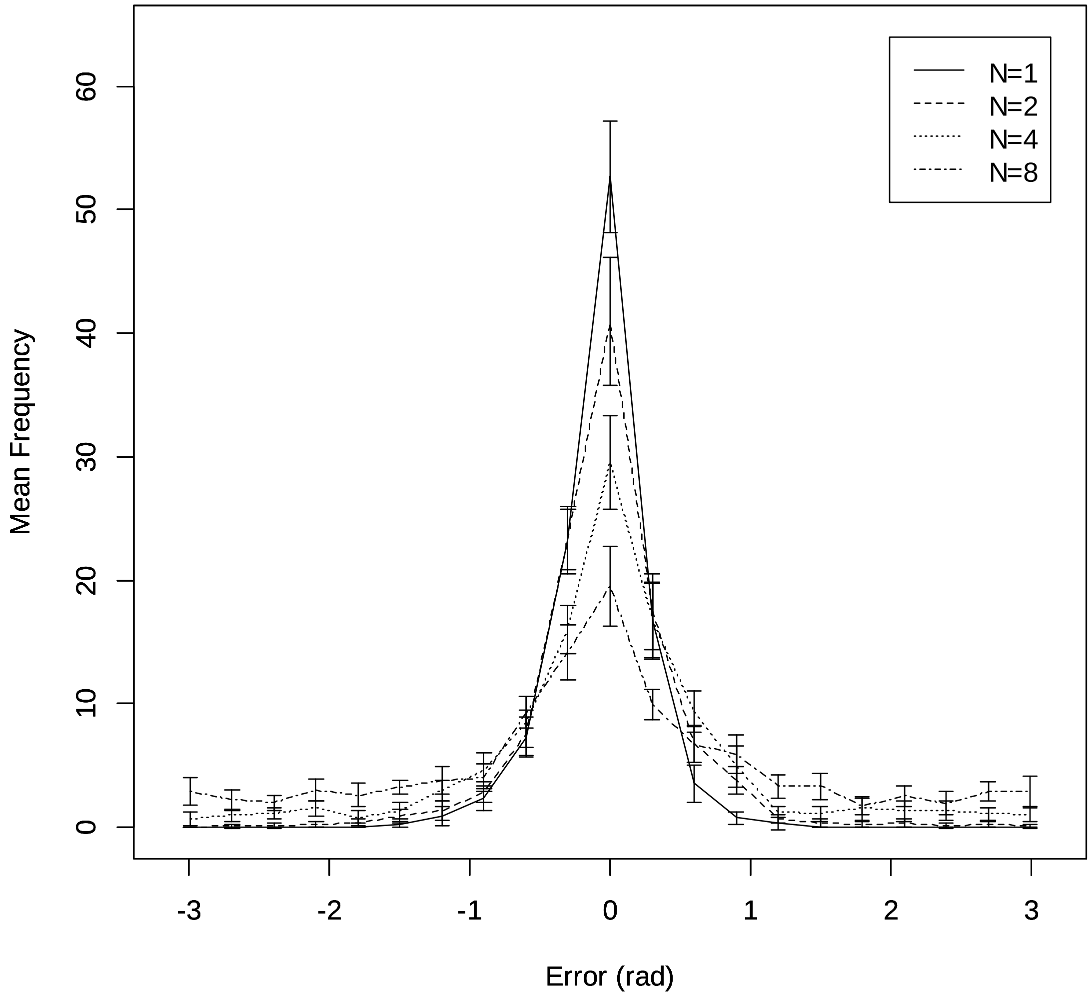

Figure 11: Error distributions in the continuous-reproduction (or delayed-estimation) paradigm for four levels of memory set size (Van den Berg et al., 2012).

Data from delayed-estimation experiments consist of empirical distributions of the estimation error (measured, for example, as the circular difference between an estimated orientation and the true orientation), and their moments. Many studies have fitted quantitative models to these error distributions (Fougnie, Suchow, & Alvarez, 2012; Sims, 2015; Van den Berg, Awh, & Ma, 2014; Van den Berg, Shin, Chou, George, & Ma, 2012; Zhang & Luck, 2008). These studies agree that the error distribution (a) is more heavy-tailed than a Von Mises (circular normal) distribution; and (b) is wider at higher set sizes (see Figure 11). These findings are theoretically important because they have played a role in attempts to distinguish theories that assume an upper limit on the number of items stored in WM (Zhang & Luck, 2008) from theories based on the notion that memory precision is variable (Fougnie et al., 2012; Van den Berg et al., 2014; Van den Berg et al., 2012). At the same time, the finding is specific to one paradigm in one domain (visual WM). Therefore, we rated it as B, which also reflects the broad distribution of survey responses to this benchmark.

5: Effects of Combining Multiple Demands

Many studies have asked participants to meet multiple demands on their WM simultaneously, such as holding in memory two or more sets of items, or carrying out a processing task while maintaining a memory set. These multiple-demand studies have revealed patterns of mutual impairment of the simultaneous demands, which have been highly informative for theories of WM.

5.1: Multiple Memory-Set Effects

Benchmark 5.1.1. Effects Within and Across Domains (A). Increased variability among the to-be-remembered items increases overall memory for these items. If one must hold in mind two memory sets of items from differing domains (e.g., one set of verbal and one set of spatial items), memory performance is superior compared to a situation in which two sets of items from the same domain are held. Regardless of the contents of the two sets, however, memory for two sets is generally poorer than memory for a single set when measured by the proportion of items recalled (see Figure 12). This benchmark has been rated A because it has been observed consistently and has appeared with a variety of task combinations, including verbal serial recall combined with visual recognition (Cowan & Morey, 2007), verbal item recognition with visual recognition (Fougnie & Marois, 2006), verbal recall with spatial recall (Sanders & Schroots, 1969), spatial pattern memory when combined with verbal serial recall that does or does not make use of a visual-spatial heuristic (Logie, Zucco, & Baddeley, 1990), and memory for action sequences combined with spatial locations (Smyth & Pendleton, 1990). The diversity of stimulus combinations that show this effect and its replicability justify it as a benchmark for WM models, and indeed, this particular finding has already been a key driver of WM theory (Baddeley, 2012; Barrouillet & Camos, 2015; Logie, 2011). Most survey respondents who rated this benchmark assigned it A (4/16) or B (8/16). This benchmark is further qualified by Benchmark 5.1.3 (Appendix B).

Benchmark 5.1.2. Effects of Heterogeneity Within a Domain (B). Notably, it is not necessary for memory sets to rely on different domains for this pattern to emerge: the same principle likewise applies to within-domain similarity. Memory for two sets is superior when they come from different categories of stimuli within either the verbal domain (Sanders & Schroots, 1969) or the visual domain (e.g., Delvenne & Bruyer, 2004). Sanders and Schroots required participants to simultaneously maintain two sets of consonants, or alternatively two sets of decreasing similarity: for instance, a set of consonants and a set of digits, a set of consonants and a set of tones, or a set of consonants and a set of spatial positions. Better recall was observed for sets as their similarity decreased, even for within-domain sets. A parallel finding occurs in visual recognition memory. Delvenne and Bruyer observed better recognition accuracy for mixed displays including elements from two visual feature dimensions than for displays including the same number of elements from a single feature dimension. A majority of survey respondents ranked this benchmark as one of the top two priority categories (5/14 for A, 6/14 for B). We rated this benchmark as a “B” because it is a modifier of the primary benchmark that cross-domain sets are maintained with less cost than within-domain sets (5.1.1).

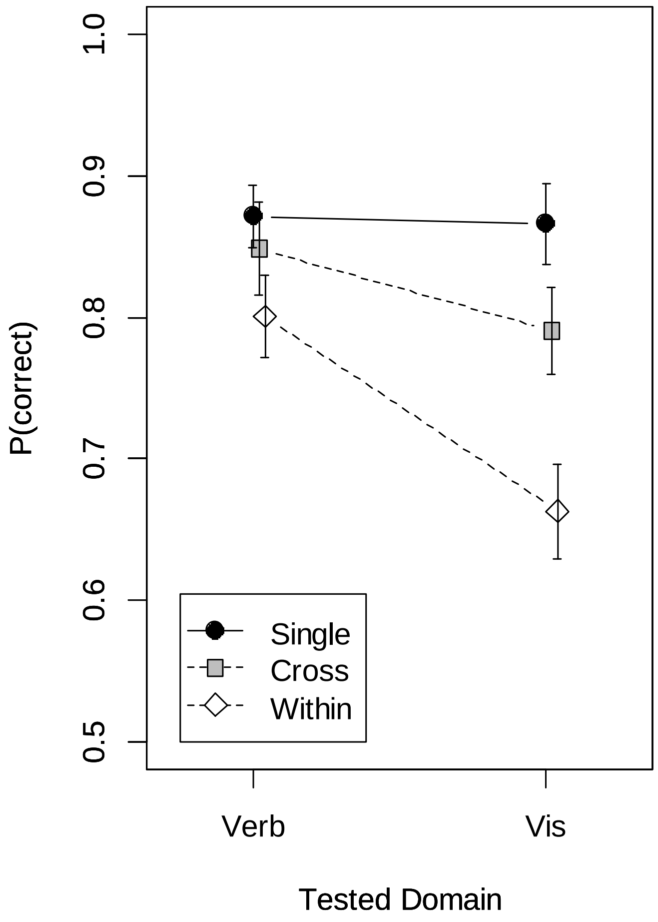

Figure 12: Accuracy of remembering a set of verbal or visual stimuli when holding in memory only the tested set (Single), or together with a second set from the other content domain (Cross) or a second set from the same domain (Within) (Cowan & Morey, 2007; data collapsed across post-cue conditions)

Benchmark 5.1.3. Asymmetric Effects Between Verbal and Spatial Sets (C).

This benchmark is described in Appendix B.

5.2 Multiple-Task Effects

Performing a secondary processing task during retention impairs memory. Memory is disrupted when a processing task needs to be performed throughout study and/or test, in the delay between study and test, or in between the presentation of the memoranda.

Benchmark 5.2.1. Disruption of memory by processing in the same domain (A). It is typically observed that the disruption of memory caused by concurrent processing of distractors is substantial when the memory items and distractors come from the same content domain (see Figure 13). Distractor processing can involve concurrent articulation of unrelated verbal material, often referred to as "articulatory suppression" (e.g., Baddeley, Lewis, & Vallar, 1984, in a serial-recall paradigm; Camos, Lagner, & Barrouillet, 2009, in a complex-span paradigm; Meiser & Klauer, 1999, in a Brown-Peterson paradigm), or another form of processing, such as mental arithmetic (e.g., Barrouillet, Bernardin, & Camos, 2004) or spatial judgments (e.g., Vergauwe, Barrouillet, & Camos, 2009). This benchmark has been rated A because it has been observed consistently in different WM domains and across a variety of paradigms. This benchmark has been observed when memory and processing items come from the verbal domain (Chein, Moore, & Conway, 2011; Hale, Myerson, Rhee, Weiss, & Abrams, 1996; Jarrold, Tam, Baddeley, & Harvey, 2010; Logie et al., 1990; Shah & Miyake, 1996), the spatial domain (Chein et al., 2011; Hale et al., 1996; Klauer & Zhao, 2004; Shah & Miyake, 1996) or the visual domain (Klauer & Zhao, 2004; Tresch, Sinnamon, & Seamon, 1993). Furthermore, it has been observed in the following paradigms (see Table 1): complex span (Chein et al., 2011; Shah & Miyake, 1996), serial recall (Hale et al., 1996) and Brown-Peterson (Jarrold et al., 2010; Klauer & Zhao, 2004; Logie et al., 1990; Tresch et al., 1993).

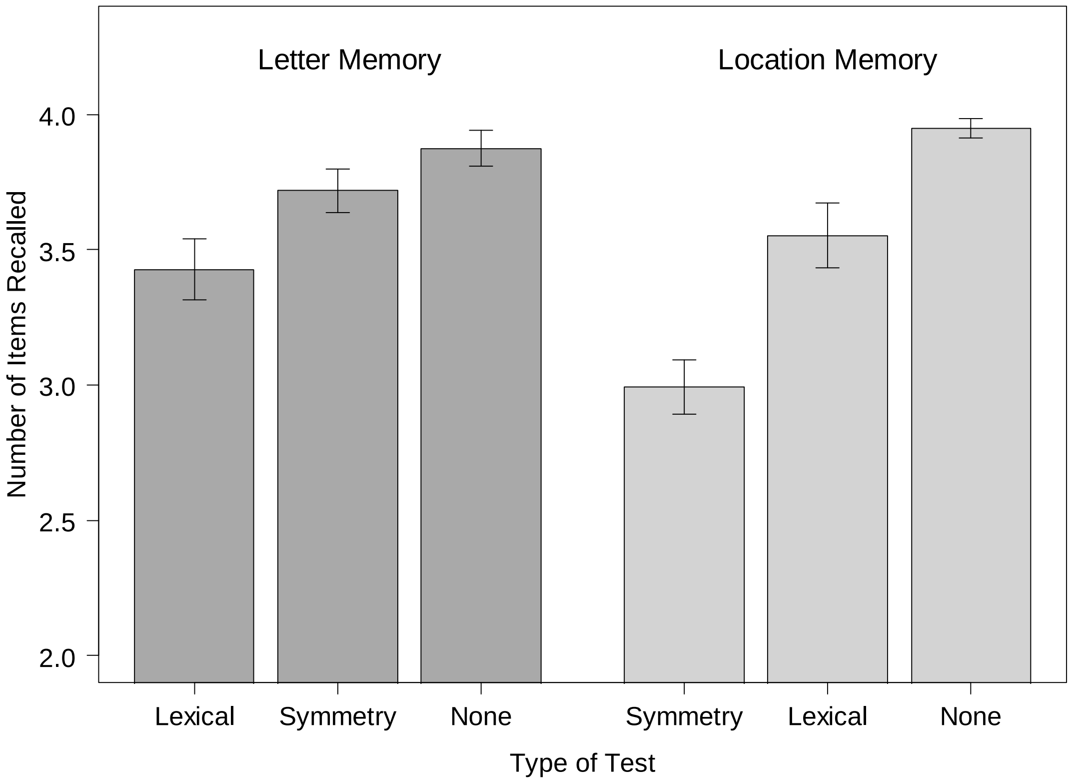

Figure 13: Accuracy of serial recall of letters or of spatial locations without concurrent distractor processing, or in conjunction with a verbal (lexical decision) or spatial (symmetry decision) distractor task (Chein et al., 2011)

Benchmark 5.2.2. Disruption of memory by processing in another domain (A). The disruption of memory caused by concurrent processing of distractors is less serious, but still present, when the memory items and distractors come from different domains (see Figure 13). This has been observed across the verbal and visuo-spatial domains (Chein et al., 2011; Jarrold, Tam, Baddeley, & Harvey, 2011; Makovski, 2012; Vergauwe, Barrouillet, & Camos, 2010; Vergauwe, Dewaele, Langerock, & Barrouillet, 2012) as well as across the visual and spatial domains (Vergauwe et al., 2009). Whereas this benchmark has mainly been observed in complex-span tasks (Chein et al., 2011; Jarrold et al., 2011; Vergauwe et al., 2010, 2012), it has also been observed in Brown-Peterson tasks (Jarrold et al., 2011; Vergauwe et al., 2009) and in a change detection task (Makovski, 2012). The observation of more severe memory disruption when memory and processing items pertain to the same domain, together with the observation of memory disruption even when processing pertains to another domain, is theoretically relevant for resource theories of WM (e.g., supporting the existence of domain-specific vs. domain-general resources) and for competing explanations of the mutual impairment of concurrent storage and processing (e.g., competition for central attention vs. representation-based interference between processing and storage; Baddeley & Logie, 1999; Barrouillet & Camos, 2015; Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012). Most survey respondents (87.5%) rated this benchmark A (6/16) or B (8/16). Because of its theoretical leverage and because it has been observed across different domains of WM and across several experimental paradigms, we assigned it an A rating.

Benchmark 5.2.3. Processing of material from same or different category as the memory materials (C).

This benchmark is described in Appendix B.

Benchmark 5.2.4. Effect of cognitive load of the processing demand (A). Another important benchmark is the cognitive-load effect of concurrent processing on memory performance, illustrated in Figure 14. Memory performance decreases with the increasing ratio between the time needed for attention-demanding processing and the time available for processing. This benchmark has been observed across a variety of task combinations: verbal storage and verbal processing (Barrouillet et al., 2004; Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007; Barrouillet, Portrat, & Camos, 2011; Hudjetz & Oberauer, 2007), verbal storage and visuo-spatial processing (Barrouillet et al., 2007; Vergauwe et al., 2010; Vergauwe et al., 2012), visuo-spatial storage and visuo-spatial processing (Vergauwe et al., 2009, 2010), and visuo-spatial storage and verbal processing (Ricker & Cowan, 2010; Vergauwe et al., 2010; Vergauwe et al., 2012). Increasing the cognitive load of a more domain-neutral processing task such as tone discrimination also disrupts verbal and visuo-spatial memory performance (Langerock, Vergauwe, & Barrouillet, 2014; Vergauwe, Langerock, & Barrouillet, 2014). In the same vein, processing tasks that require more executive control lead to poorer memory performance (Kiyonaga & Egner, 2014; Szmalec, Vandierendonck, & Kemps, 2005). In particular, the additional need for response selection (Barrouillet et al., 2007; Barrouillet et al., 2011), task switching (Liefooghe, Barrouillet, Vandierendonck, & Camos, 2008), inhibition (Barrouillet et al., 2011), or updating (Barrouillet et al., 2011) in such processing tasks yields a drop in memory performance that is commensurate with the respective increase in the time needed for processing.

While the cognitive-load effect has mainly appeared in serial recall tasks such as complex span tasks (Barrouillet et al., 2004, 2007, 2011; Vergauwe et al., 2009, 2010, 2012) and Brown-Peterson tasks (Liefooghe et al., 2008), it has also been obtained in single-item recall or local recognition tests (Ricker & Cowan, 2010; Vergauwe et al., 2009; Vergauwe, Hartstra, Barrouillet, & Brass, 2015) and change detection tasks (Vergauwe et al., 2014).

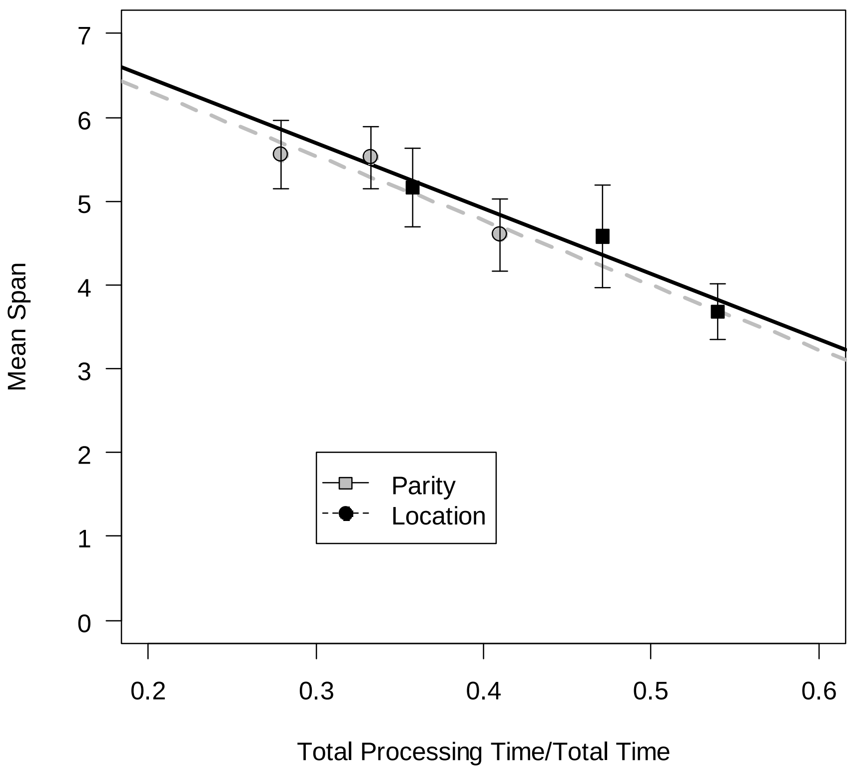

Figure 14: The effect of cognitive load on performance in a complex span task. Cognitive load was estimated by the ratio of total processing time on the parity or location judgments (i.e., the sum of response times) to total time (i.e., the sum of the time intervals for each judgment). (Barrouillet et al., 2007)

The majority of survey respondents (66%) rated this benchmark A (20/60) or B (19/60). The Benchmarks team rated it as a top priority (A) because it generalizes across WM domains and experimental paradigms. This rating is further justified by its theoretical leverage. Indeed, the cognitive load effect is directly relevant for the theoretical debate concerning the causes of forgetting from WM (i.e., decay vs. interference) and the mechanisms that can counteract that forgetting (i.e., rehearsal, refreshing, consolidation, removal of distractors) (Barrouillet & Camos, 2015; Oberauer et al., 2012).

Benchmark 5.2.5. Effect of concurrent processing on memory for features and bindings (B). Finally, memory for individual features (e.g., color or shape) and memory for bindings between features (e.g., information about which color is associated with which shape in a set of colored shapes) are equally impaired by attention-demanding secondary tasks (Allen, Baddeley, & Hitch, 2006; Allen, Hitch, & Baddeley, 2009; Morey & Bieler, 2013; Vergauwe et al., 2014); though see L. A. Brown and Brockmole (2010) for an exception. The question of whether memory for bindings is more impaired by concurrent attention-demanding processing than memory for features is central for the question of whether attention is required to establish and/or maintain bindings in WM. We rated this benchmark as B because, even though it is not general enough to be rated as A, it is one of the most general and well-replicated findings from research on binding, and must be addressed with quite high priority by theories of binding in WM.

6: Auditory distraction effects

On tests of verbal WM, performance declines as a function of auditory distraction during study, retention, or retrieval. Two types of auditory distraction can be distinguished: (1) the negative impact of to-be-ignored speech or sound on serial recall of (mostly) visually presented verbal items, known as the *irrelevant speech effect* (Salamé & Baddeley, 1982) or *irrelevant sound effect* (Beaman & Jones, 1998) and (2) the negative impact of a *deviant auditory distractor* during visual presentation of verbal lists (Hughes, Vachon, & Jones, 2007; Sörqvist, 2010).

Benchmark 6.1 Irrelevant sound effect (B).

The irrelevant sound effect is observed when participants read and memorize a list of items (e.g., digits, letters, or words). Serial recall of the list is poorer when during encoding or during the retention interval participants are exposed to auditory material that they are supposed to ignore (e.g., Miles, Jones, & Madden, 1991). Since the effect was first shown with to-be-ignored spoken language (e.g., Salamé & Baddeley, 1982), it is also known as irrelevant speech effect, but it has also been observed with non-speech sounds such as tones (Jones & Macken, 1993) or instrumental music (e.g., Klatte, Kilcher, & Hellbrück, 1995; Salamé & Baddeley, 1989; Schlittmeier, Weißgerber, Kerber, Fastl, & Hellbrück, 2012). The irrelevant sound effect has been predominantly observed with verbal memory materials, and of these mostly with visually presented ones, although there are also a few studies demonstrating it with auditory presentation of verbal items (Campbell, Beaman, & Berry, 2002) and in a visuo-spatial WM task (Jones et al., 1995; Tremblay, Macken, & Jones, 2001).

This classic finding has been rated as a benchmark because it has been observed consistently and has stimulated theoretical discussion about models of verbal WM (e.g., Baddeley, 2000b; Jones & Tremblay, 2000; Neath, 2000). In the survey, this benchmark was mostly rated as A (10/24) or B (6/24). The Benchmarks Team rated it as a moderate-priority benchmark (B) because its replicability and robustness hold, so far, only for verbal WM tasks with a strong serial-order component.

Boundary conditions. There are strong irrelevant-sound effects on tasks that rely heavily on memory for serial order, such as serial recall and serial-order reconstruction, whereas the effect on free recall is smaller (Salamé & Baddeley, 1990) – unless participants are encouraged to use a serial rehearsal strategy (Beaman & Jones, 1998). It is also smaller when participants' task is to identify an item missing from a well-known set (Beaman & Jones, 1997).

Irrelevant sound effects are particularly large with speech or speech-like sounds characterized by changes in temporal and spectral structure (e.g., backward speech or music, for an overview see Ellermeier & Zimmer, 2014). However, meaningfulness of the distractors as well as similarity between the to-be-recalled and the to-be-ignored materials do not affect the magnitude of the effect (e.g., phonologically similar distractors interfere as much as phonologically dissimilar ones, Jones & Macken, 1995; speech in a language unknown to the participants interferes as much as speech in the same language as the to-be-recalled materials, Jones, Miles, & Page, 1990). The irrelevant sound effect is, however, stronger when the distractor items are identical to the target items but occur in a different order (Bell, Mund, & Buchner, 2011; Salamé & Baddeley, 1982).

Benchmark 6.2 Changing-state effect (B)

The most investigated moderator of the irrelevant sound effect concerns variation in the acoustic characteristics of the to-be-ignored sound. Sound that changes acoustically from one token to the next (i.e., changing-state sound) is more disruptive to serial recall than repetitive, steady-state sound (Jones & Macken, 1993; Jones, Madden, & Miles, 1992; Meiser & Klauer, 1999). This is the case both for speech as irrelevant sound (e.g., a sequence of different digits vs. repetition of the same digit) and for non-speech sounds (e.g., staccato music vs. legato music; Klatte et al., 1995). This finding is illustrated in Figure 15.

Like the irrelevant sound effect, the changing-state effect has been observed repeatedly and has driven theory development in the domain of verbal WM. In the survey, most participants rated it as B (7/28) or C (11/28). We assigned it a B rating because of its strong theoretical leverage but under-specification regarding the acoustic factors that do or do not result in a changing-state effect.

Boundary conditions. In contrast to serial recall and serial order reconstruction, changing-state sound is no more disruptive than steady-state sound for free recall (Jones & Macken, 1993). Concerning serial recall, not every acoustic variation causes stronger impairment than steady-state sound. Whereas there are robust effects of pitch variation (Jones, Alford, Bridges, Tremblay, & Macken, 1999), sounds changing in intensity do not impair serial recall more than sounds of the same intensity (Tremblay & Jones, 1999).

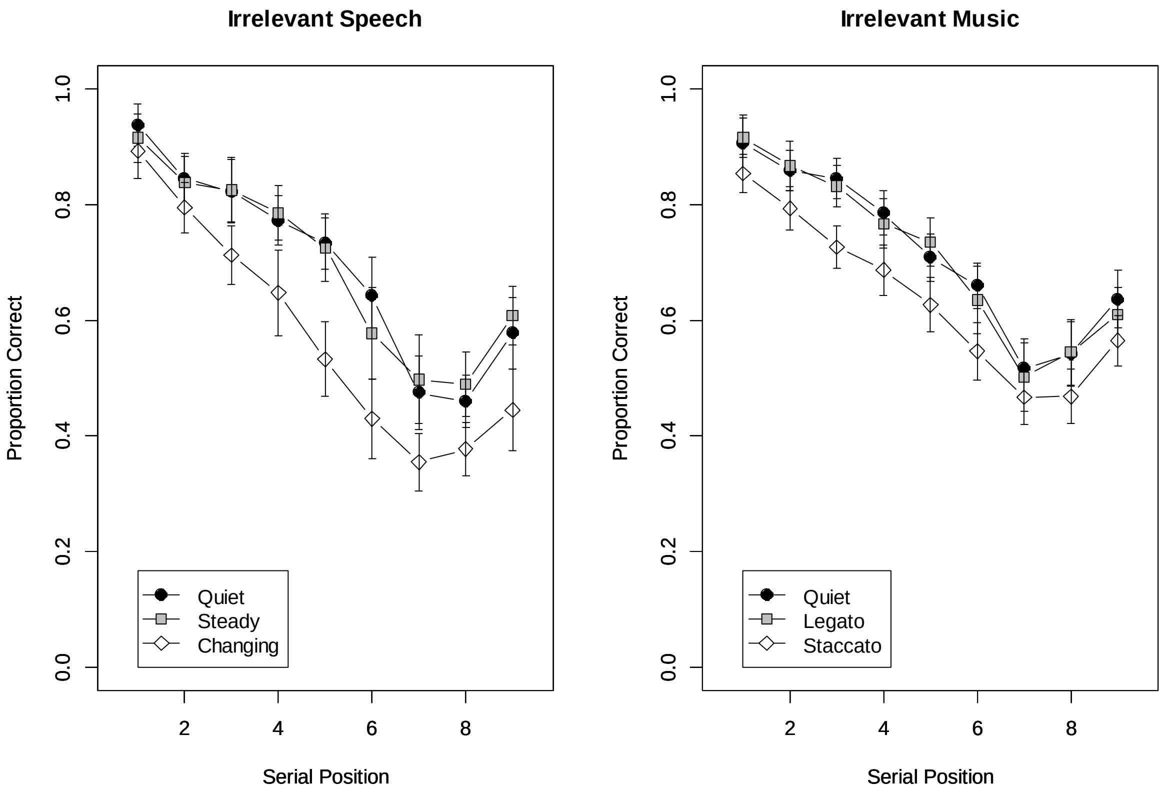

Figure 15: Irrelevant-speech and irrelevant-sound effects: Serial-position curves for serial recall of nine-digit lists presented visually, in quiet or accompanied by to-be-ignored acoustic input. Left: Irrelevant speech sounds are presented in steady-state or changing-state mode. Right: Irrelevant legato or staccato music is played (data from two studies reported in Schlittmeier et al., 2012)

Benchmark 6.3 Auditory deviant effect (C)

This benchmark is described in Appendix B.

Benchmark 7. Syllable-Based Word Length Effect in Serial and Free Recall (B).

Performance in verbal WM tasks decreases with increasing length of the items on a list. This word-length effect is robust when word length is manipulated through the number of syllables. For instance, participants correctly recall more items when a list consists of monosyllabic words compared to a list of three-syllable words. This effect is observed in serial recall (Baddeley, Thomson, & Buchanan, 1975; Mackworth, 1963) and in free recall (Bhatarah, Ward, Smith, & Hayes, 2009;

Watkins, 1972) as well as – of smaller magnitude – in probed recall (Avons, Wright, & Pammer, 1994). The syllable-based word-length effect is a key finding for theories of verbal WM and has been observed consistently in various languages. Given its restriction to the verbal domain, the rating proposed by the Benchmarks Team is a B. In the survey most of the respondents rated it as A (24/53) or B (14/53).

Boundary conditions. Word length interacts with presentation modality and articulatory suppression. For one, the word length effect is stronger for visually presented than for auditory lists (Baddeley et al., 1975; Watkins & Watkins, 1973). In addition, articulatory suppression during presentation reduces or eliminates the word length effect with visual presentation but not with auditory presentation (Baddeley et al., 1984; Baddeley et al., 1975; Bhatarah et al., 2009; LaPointe & Engle, 1990) – unless articulatory suppression also occurs during recall (Baddeley et al., 1984).

8: Effects of Similarity

8.1 Effects of within-set similarity

Benchmark 8.1.1. Phonological similarity (A). Increasing the phonological similarity between memoranda leads to worse serial recall performance (Conrad & Hull, 1964; Farrell & Lewandowsky, 2003). An example is shown Figure 16A. The phonological similarity effect is observed regardless of whether memoranda are presented visually or auditorily (Peterson & Johnson, 1971). This suggests that phonology is a preferred form of representation for verbal WM. The finding that the effect is abolished by instructing participants to encode list items semantically (Campoy & Baddeley, 2008) argues against the necessity of phonological coding. Phonological similarity also affects whole and partial report from arrays of visually presented letters (Sperling & Speelman, 1970), and recognition of serial order (Nimmo & Roodenrys, 2005). The phonological-similarity effect generalizes beyond "short-term memory tasks": A detrimental effect of phonological similarity has been observed in a memory-updating paradigm (Oberauer & Kliegl, 2006). However, with complex-span tasks, a beneficial effect of phonological similarity (i.e., rhyming words) is often observed (Chow, Macnamara, & Conway, 2016; Copeland & Radvansky, 2001; Macnamara, Moore, & Conway, 2011).

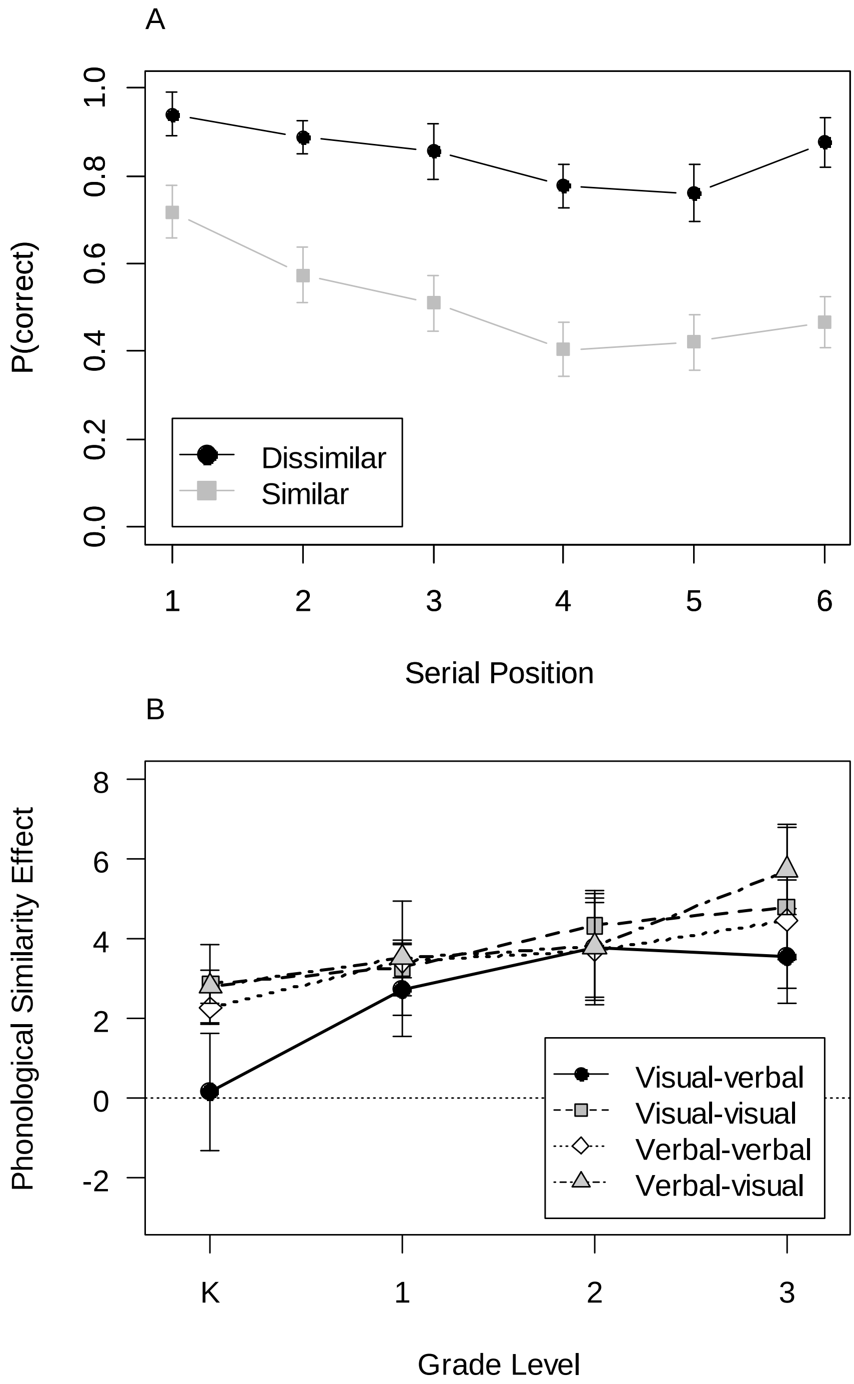

Figure 16: Phonological similarity effect. A: Serial-position curves from serial recall of lists of dissimilar or similar letters (Farrell & Lewandowsky, 2003, Exp. 1). B: Phonological similarity effect (i.e., difference in recall accuracy between dissimilar and similar condition) as a function of children's age, and combinations of modality of encoding with modality of retrieval (Jarrold & Citroën, 2013).

Closer examination has revealed that although phonological similarity detrimentally affects memory for the order of items (as measured by scoring order accuracy for those items from the list that were recalled), it can sometimes benefit recall of the identity of items (Wickelgren, 1965). This benefit applies to lists of rhyming items, whereas an item memory benefit is not observed for lists of similar items that do not rhyme (Fallon, Groves, & Tehan, 1999).

The detrimental effect of phonological similarity on order memory has been well replicated. It is a cornerstone of the phonological loop theory (Baddeley & Hitch, 1974), and it has informed several computational models of serial recall (Burgess & Hitch, 1999; Henson, 1998b; Lewandowsky & Farrell, 2008). Its limitation to verbal materials is partially overcome by the observation of analogous effects of visual similarity (Benchmark 8.1.4). For these reasons, we regard the effect of phonological similarity on order memory as a category A benchmark.

Benchmark 8.1.2. Mixed list effect of phonological similarity (B). Some studies have examined the consequences of mixing together phonologically similar and dissimilar items on memory lists (Baddeley, 1968; Farrell & Lewandowsky, 2003; Henson et al., 1996), for example, by alternating rhyming and non-rhyming letters (e.g., R*B*L*V*K*C*). The detrimental effect of similarity is restricted to the similar items, with dissimilar items being remembered as well as items on control lists composed only of dissimilar items (Baddelely, 1968; Henson et al., 1996). Indeed, when examining ordering accuracy independent of item memory, Farrell and Lewandowsky, (2003) observed that the presence of similar items improved ordering accuracy of dissimilar items on the same list. The mixed-list similarity effect is a benchmark because of its strong theoretical leverage: Mixed list effects have provided a strong test of item chaining models of serial-order memory (Lewandowsky & Murdock, 1989): The “immunity” of dissimilar items to surrounding similar items implies that the similar items are not acting as recall cues to the dissimilar items. At the same time, the effect qualifies benchmark 8.1.1, and it is limited to the serial-recall paradigm; hence we assigned it a B rating, in agreement with the survey results (22/55 B ratings).

Benchmark 8.1.3. Phonological similarity interacts with concurrent articulation (B). Although the phonological similarity effect is usually observed irrespective of whether items are presented visually or auditorily (Benchmark 8.1.1), a modulating factor is concurrent articulation. Having participants perform concurrent articulation of irrelevant material reduces or abolishes the phonological similarity effect for visually presented materials, but not for information presented auditorily (Baddeley et al., 1984; Peterson & Johnson, 1971). This three-way interaction has been interpreted as evidence supporting the phonological loop model, according to which rehearsal is needed to recode visual information into a phonological representation, and this rehearsal is blocked by concurrent articulation (Baddeley et al., 1984). As such, we found it of sufficient theoretical importance to propose it as a category B benchmark. Most survey responses were split between A and C (5/15 each).

Benchmark 8.1.4. Development of phonological similarity effect (C). This benchmark is described in Appendix B.

Benchmark 8.1.5. Effect of visual similarity on serial recall (C). This benchmark is described in Appendix B.

Benchmark 8.2. Effects of Item-Probe Similarity in Recognition and Change Detection (C)

This benchmark is described in Appendix B.

9: Effects of Distinctiveness and of Grouping

According to Hunt and Worthen (2006), more than 2000 articles on distinctiveness effects in memory have been published. Distinctiveness can be viewed as the inverse of similarity; "distinctive" memories are ones that are different from, and hence unlikely to get confused with, other memories (see G. D. A. Brown, 2015, for different notions of distinctiveness). We refer to grouping and isolation effects to follow existing literature, while noting that the same effects may be taken as evidence for either grouping or temporal isolation mechanisms by different authors (Farrell, 2012; Farrell, Wise, & Lelièvre, 2011; Hartley, Hurlstone, & Hitch, 2016).

9.1. Effects of Distinctiveness

Benchmark 9.1.1: Temporal isolation effect (B). Distinctiveness effects are often studied by varying the temporal isolation of items. Items are isolated by surrounding them with relatively large temporal gaps at the time of learning; this leads to better memory for the isolated item. This temporal isolation effect is found in both recognition and free recall tasks (C. Morin, Brown, & Lewandowsky, 2010), and in running memory tasks when list length is unpredictable (Geiger & Lewandowsky, 2008). The temporal-isolation effect is also found in tasks that measure memory for serial order when order of item recall is unconstrained (Lewandowsky, Nimmo, et al., 2008).

We note two boundary conditions. First, Polyn, Kragel, McClurey, and Burke (2016) isolated items using longer temporal intervals than Morin et al. (2010) did, and found that temporally isolated items were *less* well remembered in free recall. Second, temporal isolation experiments have typically used verbal material. When non-verbal material is used, the temporal isolation effect is sometimes found (Guérard, Neath, Surprenant, & Tremblay, 2010; Shipstead & Engle, 2013) and sometimes absent (e.g., Parmentier, King, & Dennis, 2006).

We consider the temporal-isolation effect a benchmark because of its generality across many paradigms, and because it informs the debate on whether items in working memory are distinguished on a psychological dimension of time (G. D. A. Brown et al., 2000; Lewandowsky et al., 2004). Models of WM focusing on memory for sequentially experienced events should account for the temporal-isolation effect. At the same time, there are important boundary conditions (see also 9.1.2), so that the success of a model cannot be said to stand or fall with whether it predicts this effect. On balance, we rated this benchmark as B. This agrees with the survey, where responses were about evenly distributed between A, B, and C. This benchmark is qualified by Benchmark 9.1.2 (Appendix B).

Benchmark 9.1.2. Absence of temporal isolation effects in forward serial recall (C). This benchmark is described in Appendix B.

Benchmark 9.1.3. Isolation along non-temporal dimensions influences memory (B). The well-established *von Restorff* phenomenon occurs when a single item in a to-be-remembered list that is distinctive along any dimension is better remembered (for reviews, see Hunt, 1995; Wallace, 1965). The effect is illustrated in Figure 17 using data from Lippman (1980). The solid line represents performance in a task in which participants were required to estimate the ordinal position of each of a set of 12 items (trigrams) which had been presented at the rate of two seconds per item. The dotted line shows performance in an otherwise-identical condition in which the seventh item was surrounded by a colored rectangle at presentation: The distinctiveness induced by the highlighting rectangle leads to better memory. We consider this finding a benchmark because it is found in a wide variety of paradigms including serial recall (M. H. Smith & Stearns, 1949), probed recall (Calkins, 1894), recognition (von Restorff, 1933), free recall (Bireta, Surprenant, & Neath, 2008; Elhalal, Davelaar, & Usher, 2014; Welch & Burnett, 1924) and order reconstruction (Lippman, 1980). That said, recent theorizing on WM has not been influenced much by the von Restorff effect; therefore, we rate this benchmark as B. Survey responses support this benchmark, with ratings concentrated on A (20/52) and B (19/52).

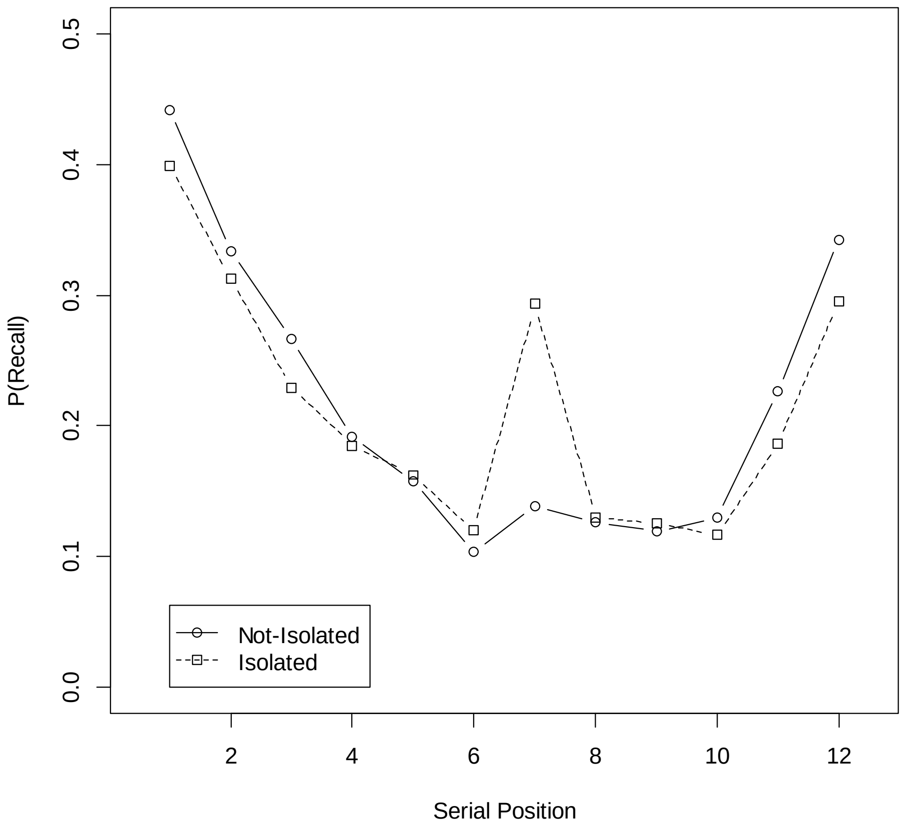

Figure 17: Probability of correctly recalling the ordinal position of CVC trigrams. In the isolated condition, the CVC in position seven was uniquely surrounded by a red rectangle (Lippman, 1980).

9.2. Grouping

At an empirical level grouping effects are closely related to distinctiveness and isolation effects, in that all involve manipulations that mark out an item or group of items. Thus, although grouping and isolation effects often receive different theoretical interpretations, we treat them together here.

Benchmark 9.2.1. Grouped lists are better recalled (A). Grouping has been most widely examined in memory for serial order. Consider presentation of a list of nine items in which, after each three items, a longer temporal gap is introduced (although a number of different manipulations may induce grouping). Our first benchmark finding is that memory performance is better, overall, when lists are grouped than when they are ungrouped (C. R. Frankish, 1989; Hartley et al., 2016; Hitch, Burgess, Towse, & Culpin, 1996; Ng & Maybery, 2002; Ryan, 1969). The improvement is primarily due to a reduction of order errors (Ryan, 1969). The basic effect is illustrated in the left-hand panel of Figure 18, which shows grouping effects for a visually presented nine-item list. Grouping benefits have also been observed for serial recall of visual (Hurlstone & Hitch, 2017) and spatial materials (Hurlstone & Hitch, 2015; Parmentier, Maybery, & Jones, 2004). Grouping effects appear to be largest when the group size is three (Wickelgren, 1967), and effects of grouping are typically larger when auditory rather than visual presentation of verbal materials is used (C. R. Frankish, 1989, 1995; Hitch et al., 1996).

Grouping effects have been highly informative for models of serial-order memory in which list items are assumed to be associated to a temporal or positional context (J. R. Anderson & Matessa, 1997; G. D. A. Brown et al., 2000; Burgess & Hitch, 1999), and therefore we consider them as a benchmark of highest priority (A), in agreement with most survey responses (29/54).

Benchmark 9.2.2. Primacy and recency effects within groups (B). Other benchmarks (e.g., 3.1) involve serial position effects: Small primacy and recency effects are typically found within groups as well as at the level of whole lists (Hitch et al., 1996; Ryan, 1969). The within-group recency effects are typically much larger when presentation is auditory rather than visual (C. R. Frankish, 1989), and this effect is illustrated in the left-hand panel of Figure 18.

Benchmark 9.2.3 Interposition errors (B). Our next benchmark finding concerns order errors that preserve within-group position. For example, consider a nine-item list organized into three groups of three. Order errors such that the fourth item is recalled in the seventh position, or vice versa, occur with relatively greater frequency in grouped as opposed to ungrouped lists. These *interposition* *errors* are seen whether presentation of verbal materials is visual (Henson, 1999) or auditory (Ryan, 1969). They have not been found, however, with visual or spatial materials (Hurlstone & Hitch, 2015, 2017).

Benchmark 9.2.4. Effects of grouping on recall latency (B). Our final grouping benchmark concerns recall latencies. Latencies preceding recall of an item are longer when that item is the first in a group, compared to items in later group positions (J. R. Anderson & Matessa, 1997; Maybery et al., 2002).

Benchmarks 9.2.2 to 9.2.4 are qualifications of the main effect of grouping (9.2.1) that add details informing theories of the (temporal) context of lists in WM; as such we rate them as B, which was also the modal response in the survey for these findings.

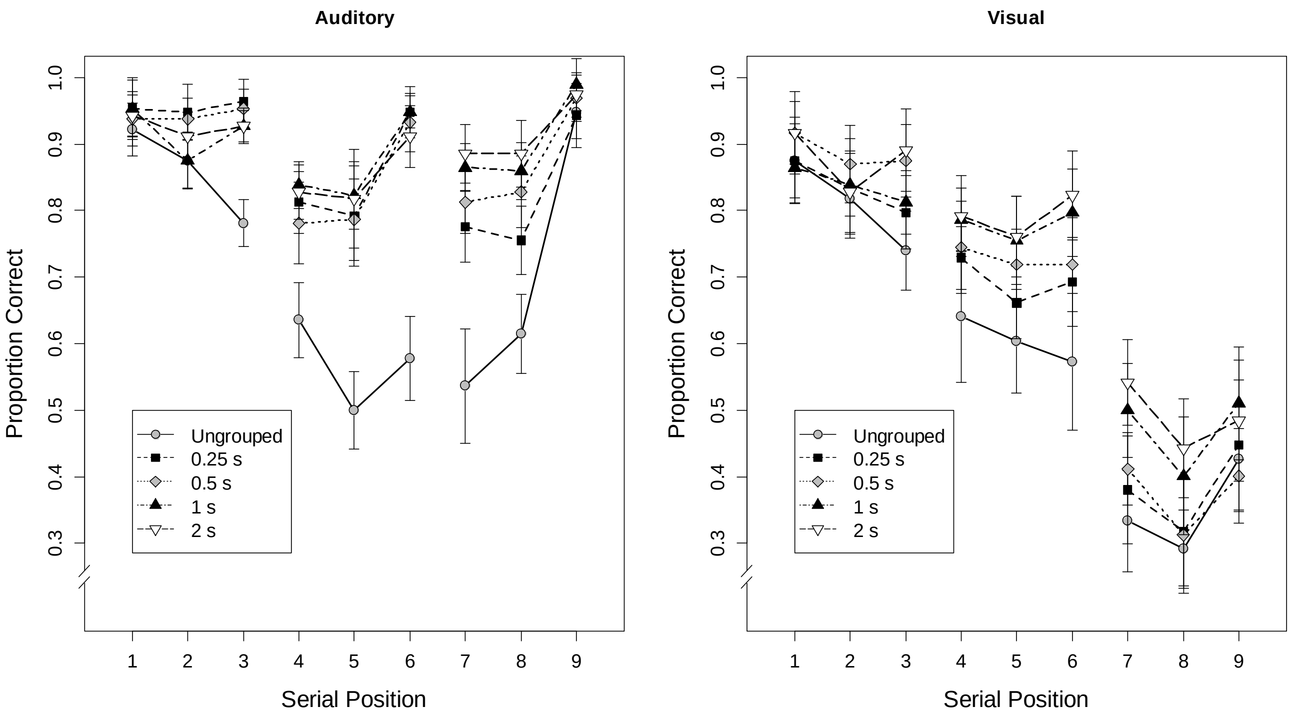

Figure 18: Serial-position curves for ungrouped lists, and lists grouped into groups of three by temporal gaps of varying length, for auditory and visual presentation modality (C. R. Frankish, 1989).

10: Prioritization of Information in WM

Individual items or subsets of information in WM can be temporarily prioritized without complete loss of not-prioritized information.

Benchmark 10.1. Effects of Retro-Cues to Individual Items in Visual WM (B).

Individual items in visual WM can be prioritized by so-called *retro-cues* during the retention interval (Griffin & Nobre, 2003; Landman, Spekreijse, & Lamme, 2003). In a typical retro-cue experiment, participants encode a simultaneously presented array of visual stimuli. The cue presented during the retention interval indicates one of the stimuli as the one most likely to be tested. When memory for the cued item is tested, response speed and accuracy are increased compared to a condition without a cue, or with a non-informative cue (Figure 19). The retro-cue is effective for delays of more than 1 s after offset of the memory array, so that its effect cannot be attributed to iconic memory, which does not last that long (Sperling, 1960). The retro-cue benefit has been observed across a broad range of methods for testing visual WM, including change detection (Landman et al., 2003), change discrimination (A. M. Murray, Nobre, Clark, Cravo, & Stokes, 2013), item recognition (Griffin & Nobre, 2003), and continuous reproduction, or delayed estimation (Pertzov, Bays, Joseph, & Husain, 2013; Souza, Rerko, Lin, & Oberauer, 2014). In most experiments the retro-cue highlights the spatial location of the cued item in the original memory array, but the retro-cue effect has also been demonstrated with cues identifying an item by its color or shape (Q. Li & Saiki, 2015; Pertzov, Bays, et al., 2013), or by verbal labels (Hollingworth & Maxcey-Richard, 2013). The beneficial effect of the retro-cue for the cued item comes at a small cost for not-cued items: When a not-cued item is tested, accuracy is slightly reduced compared to a control condition without an informative cue (Astle, Summerfield, Griffin, & Nobre, 2012; Gressmann & Janczyk, 2016). Yet, accuracy for testing not-cued items is usually much above chance, implying that not-cued items are not entirely forgotten. Moreover, items not cued by a first retro-cue can be prioritized later by a second retro-cue (Landman et al., 2003; Rerko & Oberauer, 2013).

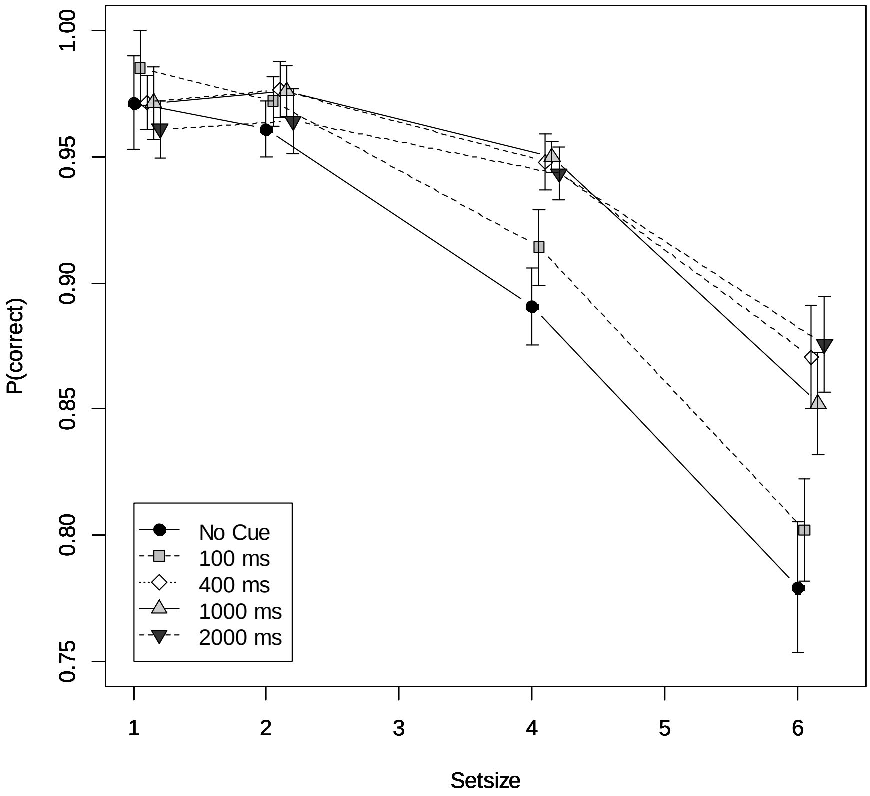

Figure 19: The beneficial effect of a retro-cue on accuracy in change detection. The effectiveness of the retro cue increases with the delay between cue and test probe (Souza, Rerko, & Oberauer, 2014)

The beneficial effect of retro-cues to single items in visual WM has been replicated numerous times, and has been found with all experimental paradigms used to study visual WM. The effect has been very informative for theories on the role of attention in WM (Gazzaley & Nobre, 2012; Sligte, Scholte, & Lamme, 2008). However, it is presently not clear whether the effects of retro-cues to single items are limited to WM for visual stimuli, as little research has been conducted with non-visual materials. Therefore, we rate this benchmark as B. Most survey responses were about evenly distributed between A (6/15) and B (5/15).

Benchmark 10.2. Item-Switch Effects (A)

Prioritization of individual items in WM has also been demonstrated with tasks involving a sequence of cognitive operations, each of which requires access to one particular item in a set held in

WM. For instance, participants could be asked to hold a small set of digits in WM, and work through a sequence of addition and subtraction tasks, each of which uses one digit from the memory set as an addend or subtrahend (Oberauer, 2003a). After access to one item in the memory set, access to the same item for the immediately following operation is faster than access to a different item (Garavan, 1998; Gehring, Bryck, Jonides, Albin, & Badre, 2003). The item-switch cost has been observed for verbal WM (Garavan, 1998; Oberauer, 2003a) and spatial WM (Hedge & Leonards, 2013; Kübler, Murphy, Kaufman, Stein, & Garavan, 2003), and for several kinds of operations on the selected item, including arithmetic (Garavan, 1998; Oberauer, 2003a), updating (Oberauer, 2003a), and local recognition (Oberauer, 2006). The switch cost increases with memory set size (Oberauer, Wendland, et al., 2003). It is reduced but not eliminated by practice (Garavan, 1998; Oberauer, 2006).

The item-switch cost is a robust effect, observed with verbal and spatial materials and several methods for testing memory. Moreover, it has been informative for theories of attention to the contents of WM (Oberauer & Hein, 2012). Therefore, we rate this benchmark as A. The most frequent rating in the survey was A (23/59), followed by B (16/59).

11: Effects of Knowledge

Knowledge from past experience has pervasive and substantial effects on performance in WM tasks. The existence of these effects continues to fuel the overarching theoretical debate as to whether WM and long-term memory are distinctly separate systems (e.g., Baddeley, 2000a) or merely different aspects of the same system. In the latter case WM is usually conceptualized as a small currently activated region of long-term memory (Cowan, 1999). However, there are many different ways in which knowledge affects WM performance, suggesting that satisfactory explanations will require richer and more detailed theoretical accounts. In what follows we give brief descriptions of the most well-established of these effects.

Benchmark 11.1 Effects of Chunking (A)

One of the principal effects of past knowledge on WM is the enhanced retention of material containing patterns encountered in previous experience. For example, immediate memory for word lists increases with their sequential redundancy measured by order of approximation to English (G. A. Miller & Selfridge, 1950). Similarly, immediate memory for letter sequences increases with the frequency of the letter bigrams in the language (Baddeley, 1964). Serial recall of letters in a complex-span task improves when the list contains known acronyms, in particular early in the list (Figure 20; Portrat, Guida, Phénix, & Lemaire, 2016). In his classic paper on our capacity to process information, G. A. Miller (1956) observed that memory span for a given type of materials can be markedly increased by becoming familiar with the patterns they contain, which he referred to as chunks. He used this to arrive at the important insight that memory span is limited in terms of number of chunks rather than number of individual items. According to Miller, span is approximately seven plus or minus two chunks (c.f. Benchmark 1.3). Thus, whereas span for binary digits is typically about the same as for decimal integers, an individual who knew how to recode from binary to decimal had a dramatically higher span of about 40 binary digits (but nevertheless about seven chunks). Similarly, a regular jogger was able to increase his digit span to the vast length of 80 items by learning to recode sequences into chunks corresponding to familiar running times for various distances (Ericsson et al., 1980). At the same time there was no improvement in his letter span, consistent with the specificity of the relationship between chunking and prior learning.

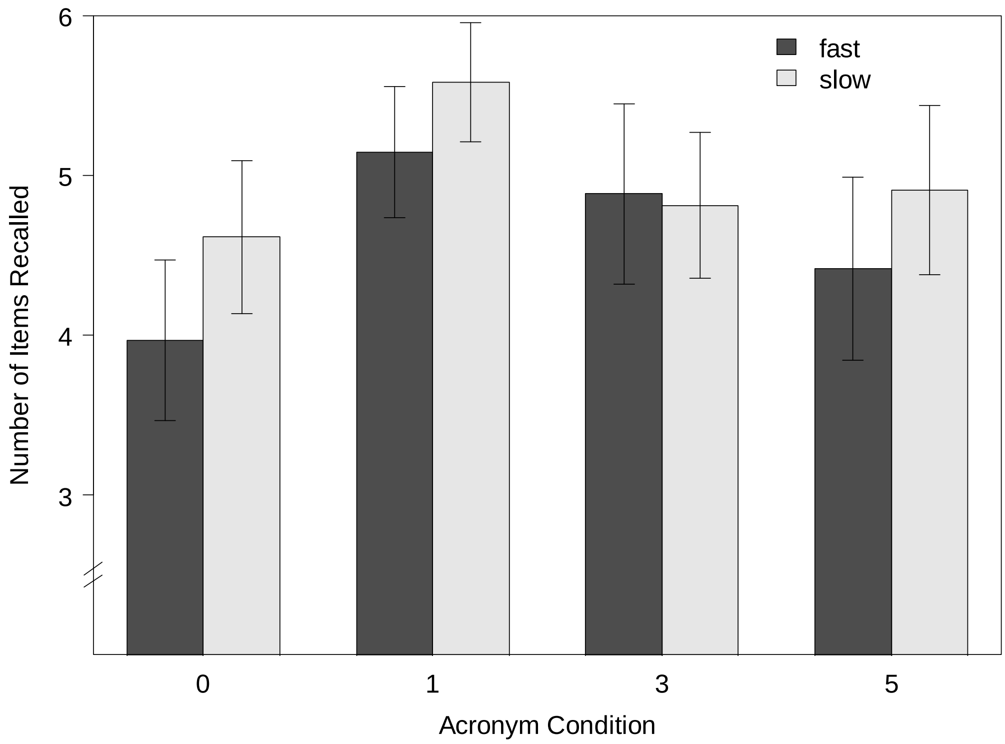

Figure 20: Performance in a complex-span task involving serial recall of letters containing no acronym (Condition 0) or an acronym starting in list position 1, 3, or 5, for fast and slow pace of distractor processing (Portrat et al., 2016).

Two further features of chunking are worth noting briefly. One is that chunking is often entirely spontaneous. For example, immediate recall of a temporally grouped sequence of letters is better when the group boundaries parse the list into known acronyms (e.g. YMCA FBI PHD TV) than when they break them up (Bower & Springston, 1970). This suggests a perceptual component to the effectiveness of chunking. Secondly, chunking is by no means confined to verbal materials. For example, immediate visual memory for colored objects increases when spatially grouped pairs of colors co-occur frequently across trials (Brady, Konkle, & Alvarez, 2009). In another visual memory task in which pieces on a chessboard are recalled immediately following a brief exposure, chess experts are able to recall substantially more pieces than novices and achieve this by encoding larger chunks (Chase & Simon, 1973b). This advantage depends on the distribution of pieces coming from a real game of chess. When the pieces are randomly distributed the difference between experts and novices is greatly reduced, consistent with the dependence of chunking on specific knowledge. Subsequent research has confirmed these findings and developed improved methods for identifying the chunks used in chess (Gong, Ericsson, & Moxley, 2015).

The beneficial effect of chunking is robust and general, and has theoretical leverage in at least two regards. First, it reflects the influence of long-term knowledge on performance in tests of WM. Second, it provides the basis for contemporary estimates of WM capacity in terms of the number of chunks that can be remembered. Reappraisal of evidence from a variety of sources suggests that capacity is significantly lower than Miller's (1956) original estimate of seven, being limited to just three or perhaps four chunks (Broadbent, 1975; Cowan, 2001); we consider evidence pertaining to the chunk capacity limit above (Benchmark 1.3). The central role of the chunking effect for theories of WM justifies its status as a high-priority (A) benchmark.

Benchmark 11.2 Sentence superiority effect (C)

This benchmark is described in Appendix B.

Benchmark 11.3 Effects of Lexicality, Word Frequency, and Phonotactic Frequency (B)

A somewhat different effect of knowledge concerns the extent to which the immediate recall of individual items benefits from prior learning of these items. This benefit is most evident in the lexicality effect, whereby memory span for known words is one or two items higher than span for pronounceable nonwords (Hulme, Maughan, & Brown, 1991). Subtler effects are observed as a function of word frequency, a typical finding being that span for words with a high frequency of occurrence is about half an item higher than span for low frequency words (Hulme et al., 1997). Importantly for theoretical accounts, the effect of word frequency cannot be attributed to potential confounding variables such as articulation rate and age of acquisition (Roodenrys, Hulme, Alban, Ellis, & Brown, 1994) and persists under concurrent articulation (Gregg, Freedman, & Smith, 1989), though there is probably some contribution from articulatory fluency (Woodward, Macken, & Jones, 2008).

Knowledge at the sublexical level can also affect verbal short-term memory, as in the effect of phonotactic frequency. This is the finding that nonwords constructed from high-frequency pairs of phonemes are better recalled than those containing low-frequency phoneme pairs (Gathercole, Frankish, Pickering, & Peaker, 1999; Majerus & Van der Linden, 2003; Thorn, Gathercole, & Frankish, 2005). A further sublexical effect is the finding that errors in the immediate recall of nonwords reflect linguistic consonant-rime syllable structure (Treiman & Danis, 1988). In general, short-term memory for nonwords appears to be influenced by a combination of sublexical and lexical knowledge, the latter being reflected in effects of lexical neighborhood size (i.e. the number of words that differ from a nonword by altering one of its phonemes). Nonwords that are more word-like on this measure are better recalled (Thorn & Frankish, 2005). This finding has been very influential on theorizing about the relation between WM and LTM, and the majority of survey respondents rated it as A (17/26). However, because this benchmark is necessarily limited to verbal material, we gave it a B rating.

Benchmark 11.4. Regularization (C)

This benchmark is described in Appendix B.

Benchmark 11.5. Hebb Repetition Effect (A)

Hebb (1961) studied an immediate serial recall task in which the same list was used on multiple trials, without informing participants, and found that recall of the repeated list improved as a function of number of repetitions (Figure 21). Hebb took this as evidence that there is long-term learning of information even when it is held only briefly in short-term memory. However, the repetition effect can equally be regarded as a further instance of the effect of prior learning on immediate recall. The repetition effect has a visuo-spatial analogue that has similar characteristics (Couture & Tremblay, 2006; Horton, Hay, & Smyth, 2008; Turcotte, Gagnon, & Poirier, 2005), suggesting that the underlying mechanism is very general. There is, nevertheless, evidence that the Hebb repetition effect may serve a specific role in the verbal domain, namely that of learning novel phonological sequences and thereby adding new words to the lexicon (Szmalec, Duyck, Vandierendonck, Mata, & Page, 2009).

An important characteristic of the verbal Hebb repetition effect is its dependence on the rhythm and timing of the repeated sequence. Thus, learning is reduced when the temporal grouping pattern of a sequence changes across repetitions (Bower & Winzenz, 1969; Hitch, Flude, & Burgess, 2009). There is also evidence that the learning effect builds from the beginning of a repeated sequence (Bower & Winzenz, 1969; Hitch, Fastame, & Flude, 2005).

The Hebb effect is a highly general phenomenon of memory for serial order, and it has been an explanatory target of several computational models (Burgess & Hitch, 1999, 2006; Page & Norris, 2009). For these reasons, and in agreement with the majority of survey responses (30/55), we regard it as a benchmark of high priority (A).

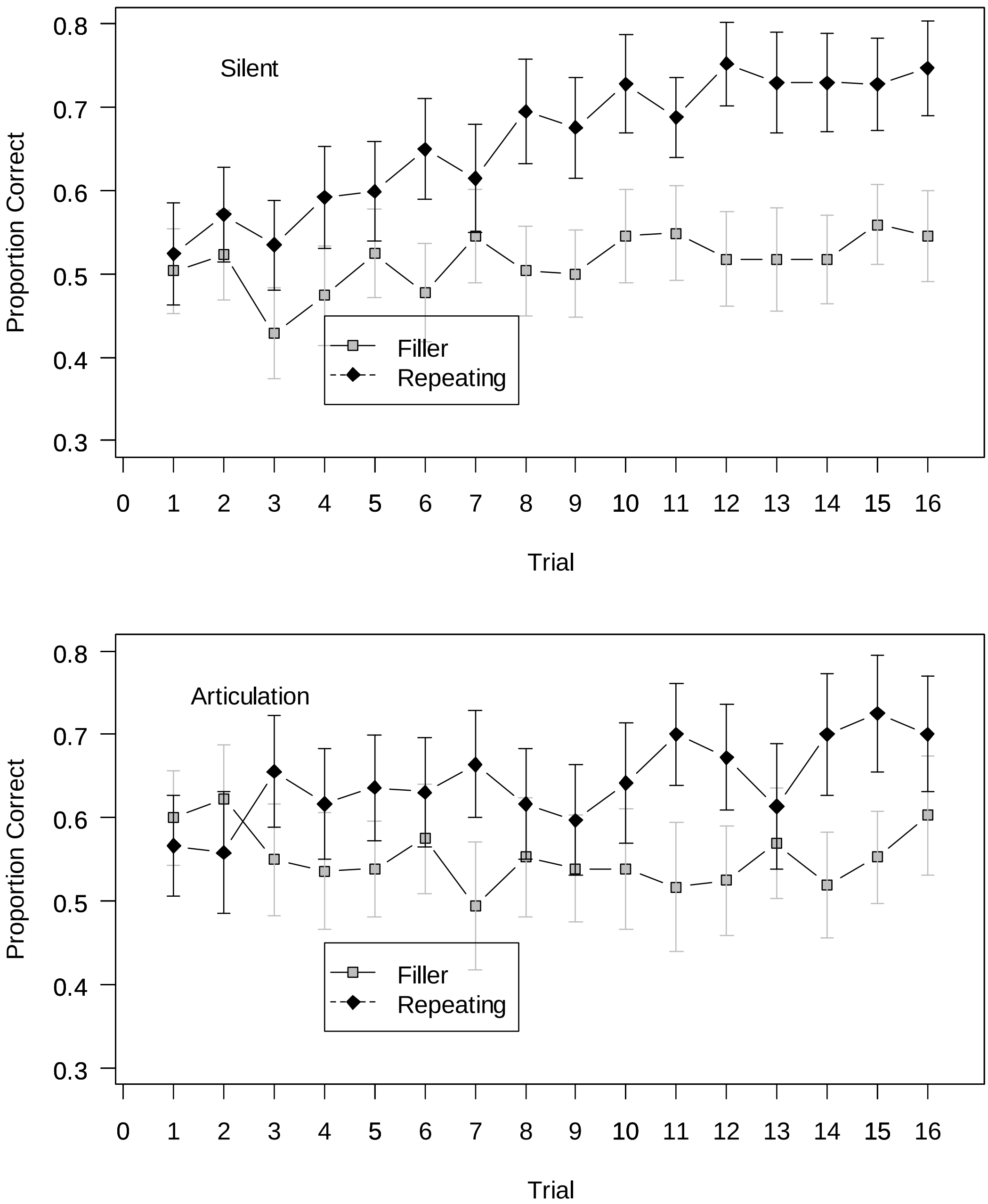

Figure 21: The Hebb repetition effect in standard conditions (Silent) and with concurrent articulation: Serial recall of letter lists improves over repetitions of the same list across trials, whereas performance on filler lists, which are new on each trial, does not improve (Page, Cumming, Norris, Hitch, & McNeill, 2006, Exp. 1).

12: Individual Differences

Benchmark 12.1. Positive Manifold (A).

With respect to research on individual differences in WM capacity, a fundamental benchmark is that performance on WM tasks correlates positively for all kinds of tasks and materials (Engle, Tuholski, Laughlin, & Conway, 1999; Kane et al., 2004; Kyllonen & Christal, 1990; Oberauer, Süß, Wilhelm, & Wittman, 2003; Unsworth, Fukuda, Awh, & Vogel, 2014). For example, scores on complex span tasks with verbal stimuli are positively correlated with scores on complex span tasks with visual/spatial stimuli (Kane et al., 2004; Shah & Miyake, 1996) and scores on complex span tasks are positively correlated with scores on other types of WM tasks, such as simple span tasks (Engle et al., 1999), change detection tasks (Unsworth et al., 2014), *n*-back tasks (Chuderski, 2014; Kane, Conway, Miura, & Colflesh, 2007), and updating tasks (Engle et al., 1999; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000). As this finding is fundamental to establish the notion of general working-memory capacity as an individual-differences construct, we regard it as a benchmark of highest priority (A). The survey data support this assessment, with A as the modal response (12/27).

Benchmark 12.2. Higher Correlations Within Domains (B).

Among this pattern of positive correlations, a second benchmark finding is that correlations tend to be higher within than across domains, and correlations tend to be higher within the same types of task than across different task types. For example, correlations within the verbal domain and within the spatial domain are higher than correlations across the verbal/spatial domains (Bayliss, Jarrold, Gunn, & Baddeley, 2003; Kane et al., 2004; Shah & Miyake, 1996). Because benchmark 12.2 is a qualification of benchmark 12.1, we rated it as B, in agreement with the modal survey response (10/24).

Benchmark 12.3. Higher Correlations among Complex than among Simple Spans (B).

Cross-domain correlations are higher for complex span tasks than for simple span tasks in adults (Kane et al., 2004) and in children (Alloway, Gathercole, & Pickering, 2006; Bayliss et al., 2003). More specifically, correlations between verbal and spatial complex span tasks are higher than correlations between verbal and spatial simple span tasks. Yet, consistent with Benchmark 12.1, all correlations are positive. This pattern of correlations is considered a benchmark but was rated to have only intermediate priority (B) because it is specific to one pair of WM paradigms (i.e., simple and complex span tasks). In addition, whereas this particular empirical pattern has been repeatedly observed, other findings call into question whether separate factors can be established for simple and complex span (for a review see Unsworth & Engle, 2007b). The modal survey response was also a B (23/55).

Benchmark 12.4. Separation of Primary and Secondary Memory (B).

When estimates from several WM tasks are subjected to factor analysis, a factorial separation of indicators of “primary memory” and “secondary memory” has been observed (Unsworth & Engle, 2007a; Unsworth, Spillers, & Brewer, 2010). For example, Unsworth et al. (2010) had subjects perform an immediate serial recall task, and they extracted measures of primary memory (PM) and secondary memory (SM) based on the method developed by Tulving and Colotla (1970). They found that both PM and SM correlated with WM capacity, and PM and SM each accounted for unique variance in WM capacity. We propose this result as a *conditional* benchmark because the measurement of PM and SM is based on a theoretical assumption about the underlying structure of memory, and on additional assumptions underlying the methods of measuring PM and SM separately. In agreement with the modal survey response (17/55) we rated it as B.

WM task performance is correlated with several measures of “component processes” of WM, that is, processes that have been proposed to contribute to performance in WM tasks. We identified correlations with indicators of two such component processes as benchmarks (12.5 and 12.6).

Benchmark 12.5. Correlation Between Verbal WM and Measures of Articulation and Retrieval Speed (B).

Measures of WM capacity correlate with articulation speed (Cowan et al., 1994; Nicolson, 1981), and retrieval speed in children (Cowan, 1992). This benchmark is robust and has theoretical leverage because it is predicted by theories that assign articulatory rehearsal and retrieval the role of a component process in WM. At the same time its relevance is limited to the verbal domain, therefore we assign it rating B, in agreement with the most frequent survey response (21/50).

Benchmark 12.6. Correlation Between WM and Attention Indicators (A).

WM task performance is positively correlated with measures of attention that place minimal demands on memory. By "minimal demands on memory" we mean tasks that require maintenance of task instructions, task goals, and response rules, but not multiple stimuli, as in WM tasks. For example, the Stroop task requires maintenance of instructions, a goal, and perhaps response mappings but there is not a memory load per se, because none of the presented words needs to be remembered. Correlations between WM task performance and such attention measures are considered a benchmark because of their theoretical leverage – they are predicted by theories assuming a relation between WM and controlled attention (Kane & Engle, 2002; Unsworth & Engle, 2007a). In addition, the finding generalizes over several indicators of controlled attention.

The evidence for correlations of WMC with three indicators of attention is sufficiently strong and replicable to warrant benchmark status: (1) WM task performance is correlated negatively with the size of the Stroop effect. This correlation is consistently found if the majority of trials is congruent and only a minority is incongruent, but not when congruent and incongruent trials occur equally often (Chuderski, Taraday, Nęcka, & Smoleń, 2012; Hutchison, 2011; Kane & Engle, 2003; Meier & Kane, 2013; Morey et al., 2012). (2) WMC is correlated with accuracy in the anti-saccade task (Chuderski, 2014, 2015; Kane, Bleckley, Conway, & Engle, 2001; Meier, Smeekens, Silvia, Kwapil, & Kane, 2018; Redick et al., 2016; Shipstead, Lindsey, Marshall, & Engle, 2014; Unsworth et al., 2014; Unsworth & Spillers, 2010). (3) WMC is negatively correlated with the prevalence of mind wandering during longer periods of working on a cognitive task, assessed with thought probes interspersed at random times during task performance, or through post-task questionnaires (McVay & Kane, 2009, 2012); a meta-analysis estimated a correlation of $r = .12$ (Randall, Oswald, & Beier, 2014).

For other indicators of controlled attention (i.e., the flanker effect and the Simon effect), some studies have found a correlation with WMC (e.g., Heitz & Engle, 2007) but others did not (e.g., Wilhelm, Hildebrandt, & Oberauer, 2013); therefore we regard them to be insufficiently robust for being a benchmark.

Taken together, this benchmark is well established; it generalizes across several indicators of controlled attention and of WM capacity, and it has substantial theoretical leverage. Therefore, we assign it rating A.

Benchmark 12.7. Correlation of WM with Fluid Intelligence (A).

Finally, the seventh benchmark finding on individual differences is that WM task performance is strongly correlated with measures of general fluid intelligence. Two meta-analyses of latent variable studies investigating the relationship between WM capacity and fluid intelligence found that the two constructs are correlated somewhere between $r = .71$ (Kane, Hambrick, & Conway, 2005) and $r = .85$ (Oberauer, Schulze, Wilhelm, & Süß, 2005). These results have recently been corroborated by a large sample study ($N = 2200$) demonstrating a correlation between WM and fluid intelligence of $r = .77$ (Gignac, 2014). This finding is considered a high-priority (A) benchmark because, as noted, it is supported by two meta-analyses and a recent large sample study, and because it links research on individual differences in WM capacity to the broader field of intelligence. At the same time, we acknowledge that a theory or model of WM could legitimately focus first on explaining how WM works, before turning to the relation of WM to other constructs.

13: Neuroscience

Neuroscience offers architectural constraints and mechanistic insight that provide important considerations for models of WM. First, revealing the neural substrates of WM phenomena enable inferences based on the observed regional dependencies of those phenomena. If distinct neural regions or processes are associated with different phenomena, this provides evidence that those phenomena are dissociable suggesting that such phenomena should be considered distinct aspects of a model of WM. Second, the nature of the neural code underlying information representation and processing indicate the mechanisms by which the brain instantiates cognitive functions commonly ascribed to WM. For example, sustained firing of neurons during delay intervals (Fuster & Alexander, 1971; Kubota &

Niki, 1971) is widely considered the neural basis of short-term retention, but other codes may yet be just as important (e.g. Stokes, 2015). Such data provide knowledge of the coding scheme used by the brain that may be essential for reproducing the benchmark findings of WM. Given the wealth of neuroscience literature, we will consider only the most essential and replicable findings here. Furthermore, for the sake of synergy with the rest of the benchmarks, we will consider mostly human findings, using animal models largely as a backdrop for analogous human findings.

Benchmark 13.1. Dissociable Neural Substrates of Different Content Domains (A)

One of the key contributions of neuroscience is the demonstration that dissociable neural networks are associated with the retention of different types of information. Goldman-Rakic (1987) marshalled an impressive collection of non-human primate data from single-unit recordings, lesion-induced behavioral impairments, and neuroanatomical projections to put forward the view that distinct circuits subserve the retention of spatial and identity-based information, respectively. Courtney and colleagues first using PET (Courtney, Ungerleider, Keil, & Haxby, 1996) and then fMRI (Courtney, Petit, Maisog, Ungerleider, & Haxby, 1998) provided human evidence consistent with these findings. These results revealed that dorsal areas of cortex including the superior frontal sulcus and parietal cortex support spatial WM, while ventral areas of cortex including the inferior frontal gyrus and temporal cortex support object WM. Furthermore, a wealth of human neuroimaging evidence implicates left inferior frontal areas (i.e. Broca's area) and peri-sylvian areas in the short-term retention of verbal information (Awh et al., 1996; Chein & Fiez, 2001; Cohen et al., 1997; Fiez et al., 1996; Paulesu, Frith, & Frackowiak, 1993; Postle, Berger, & D'Esposito, 1999; E. E. Smith, Jonides, & Koeppe, 1996). These broad networks have been further substantiated through meta-analysis of the numerous neuroimaging studies conducted on human WM (Nee et al., 2013; Owen, McMillan, Laird, & Bullmore, 2005; Rottschy et al., 2012; Wager & Smith, 2003). Similar conclusions have been reached on the basis of lesion (D'Esposito & Postle, 1999; Muller & Knight, 2006) and transcranial magnetic stimulation data (Mottaghy, Gangitano, Sparing, Krause, & Pascual-Leone, 2002).

There have been suggestions that the prefrontal parts of these networks are particularly important when storage is accompanied by additional processing demands (D'Esposito & Postle, 1999;

Wager & Smith, 2003). However, that dissociable networks are involved in representing different forms of content for WM has been evidenced in a wide-variety of tasks including those involving simple retention (e.g. item-recognition), as well as those involving more complex demands (Rottschy et al., 2012). Collectively, these data suggest that distinct representational bases exist for broad classes of information (i.e. verbal, spatial, object). This benchmark has broad empirical support, and it speaks to the much debated question whether there are domain-specific working-memory mechanisms, and how they should be conceptualized (Baddeley, 1986; Depoorter & Vandierendonck, 2009; Vergauwe et al., 2010). Therefore, we rated this benchmark as high priority (A), although survey responses were about evenly distributed between A, B, and C.

Benchmark 13.2. Preserved WM in Amnesia (A)

Whereas the above data indicate that WM is sub-divided by information domain, other data indicate that WM is distinct from other forms of memory. The most often cited finding in this regard is that damage to the medial temporal lobe (MTL) produces a profound impairment in the ability to form new long-term memories, while relatively sparing many aspects of WM (Baddeley & Warrington, 1970; Cave & Squire, 1992; Scoville & Milner, 1957). For example, digit span, a measure of verbal WM capacity, is intact in patients with MTL damage. Although a growing literature has demonstrated that MTL damage does impair some forms of WM especially for novel or relational information (Finke et al., 2008; Hannula, Tranel, & Cohen, 2006; Olson, Page, Moore, Chatterjee, & Verfaellie, 2006; Pertzov, Miller, et al., 2013; Ranganath & Blumenfeld, 2005), the impact on long-term memory, or on short-term memory tests exceeding the capacity of WM, is unquestionably more severe (Jeneson & Squire, 2012).

The relative preservation of short-term maintenance after MTL damage is well replicated, and has been foundational for the concepts of short-term and WM, and we therefore rate it as a high-priority (A) benchmark. This rating is also supported by the survey, with A as the modal rating (12/24).

Benchmark 13. 3 Measures of Neural Activity Track Amount of Information in WM (A)

Another popular distinction between long-term and WM is the nature of the neural code for storage. For long-term memory, lasting synaptic changes (e.g. long-term potentiation) are thought to provide a mechanism for retention. WM, on the other hand, is presumed to be related to more transient phenomena (Goldman-Rakic, 1995). As alluded to above, active neural firing is the most widely accepted view of how information is retained in WM (Fuster & Alexander, 1971; Kubota & Niki, 1971). Whereas direct neural recordings in humans are rare, the BOLD signal measured with fMRI and electrical signal measured with EEG are related to underlying neural activity (Logothetis, 2003). Both measures yield signals that reflect the load on WM during the retention interval (Cowan et al., 2011; Manoach et al., 1997; Veltman, Rombouts, & Dolan, 2003).

A particularly tight connection between BOLD and EEG signals and an estimate of the number of items retained has been established for visual WM (Todd & Marois, 2004, 2005; Vogel & Machizawa, 2004; Vogel, McCollough, & Machizawa, 2005; Xu & Chun, 2006). In the examined paradigms, the number of visual objects to-be-retained in visual WM (i.e., the memory set size) is varied, and the number of items retained is estimated from a person's accuracy at different set sizes. Typical estimates for young adults are 3 to 4 items (see Benchmarks 1.3). Both BOLD and EEG signals in posterior cortical areas reach a plateau at set size 3 to 4 when averaged across participants. The change in BOLD activity evoked by the storage of a single item and BOLD activity evoked by storage of an individual's maximum capacity has been shown to correlate with working memory capacity; however, a caveat for this result is that it has only been observed in a single study (Todd & Marois, 2005). A much larger body of evidence has linked contralateral delay activity (CDA) in posterior EEG electrodes with individual differences in WM capacity. A meta-analysis of 12 samples (286 subjects) across 11 studies revealed a robust correlation between the increase in CDA amplitude with increasing numbers of items stored and behavioral measures of WM capacity (Luria, Balaban, Awh, & Vogel, 2016).

Likewise, multiple studies have shown that CDA activity provides a sensitive index of "irrelevant storage" during tasks in which observers attempt to store target items that are presented amongst irrelevant distractors; a meta-analysis of 9 samples (200 subjects) across 7 studies revealed a strong correlation between behavioral estimates of filtering efficiency and a CDA measure of irrelevant storage (Luria et al., 2016). Finally, in line with the known links between WM capacity and intelligence, a latent variable analysis showed that CDA amplitude is a robust predictor of fluid intelligence and attentional control (Unsworth et al., 2014). Thus, a large body of evidence suggests that CDA activity taps into core aspects of WM ability as well as other constructs that have been linked with WM capacity via analyses of individual variations in cognitive ability.

These findings also motivate basic questions about which aspects of memory load are tracked by these neural measures of WM storage, because increases in the number of items stored in WM are often confounded with increases in the number of physical elements on the screen and the total amount of information that is contained within the memory array. There is some evidence separating these aspects: When multiple stimuli in a display are perceived as a single object based on Gestalt grouping cues, the amplitude of both BOLD and EEG signals is determined by the number of perceived objects, not the number of physical stimuli or the total amount of information contained within each perceived object (Balaban & Luria, 2015; Xu & Chun, 2006).

This benchmark has so far been best established for the visual domain; whether a similarly tight connection between neural signals and the estimated number of items holds in verbal and purely spatial WM has been less carefully studied and is presently unclear. Nevertheless, we rated benchmark 13.3 as high priority (A) because it is replicable, it generalizes across different methods (i.e., EEG and BOLD signals), and it is of high theoretical importance. The ramping activity up to a person's estimated item limit for visual WM provides important constraints for the mechanisms of capacity limits. For instance, it could reflect a mechanism that is weakly deployed at low loads, and strongly deployed up to a limit at high loads. This empirical pattern challenges models that assert a full distribution of all available mnemonic resources regardless of the number of memoranda (e.g., Van den Berg et al., 2012; Wilken & Ma, 2004; Zhang & Luck, 2008). The A rating is supported by modal response in the survey (20/51 for A).

Benchmark 13. 4 Short-Term Retention Without Measurable Neurally Active Representations (C)

This benchmark is described in Appendix B.

Discussion

In an ideal scientific world, the accumulation of empirical evidence on a subject gradually leads us towards a better theoretical understanding of that subject. In the long run, we hope to arrive at a single unified theory that explains all extant findings. Looking back on half a century of research on WM leaves us with the impression that such a scenario will not come to pass unaided. Empirical knowledge is accumulating at an impressive pace. Our theoretical understanding of WM, however – though arguably also making progress – is lagging more and more behind: The rate at which new empirical phenomena are established outpaces the rate at which we provide explanations for them within a unified theory. As a consequence, we observe a proliferation of theories which, rather than competing for the best explanation of all empirical findings about WM, live side by side in their respective explanatory niches: Each theory defines its own set of supporting findings on which it thrives. Efforts towards building a unified theory are discouraged by the fact that any such theory, if formulated precisely enough to be testable, immediately clashes with dozens, if not hundreds of findings.

The present work is motivated by our conviction that we can hope to make progress toward a unified theory of WM, if two conditions are met. First, we need to acknowledge that, for the foreseeable future, the expectation that such a theory explains *all* empirical phenomena in the field is unrealistic. Second, as a consequence, we need to work towards a rational way of prioritizing phenomena as targets for explanation. This means that we need criteria for judging how important it is for a theory to explain a given phenomenon. These criteria can be used to define a set of benchmark findings that every theory that intends to provide a comprehensive account of WM should strive to explain. With this article we aim to initiate a discussion about the criteria for benchmark findings, and about the question which findings, in light of our current knowledge, should be regarded as benchmarks.

In this article we proposed criteria for benchmarks, and presented a set of benchmarks rated by priority. We took several steps to facilitate consensus in the field about these proposals. First, we ensured that the Benchmarks Team consists of a diverse set of researchers with heterogeneous theoretical views on WM. Second, we made an effort to formulate each benchmark in a theory-neutral way. Third, we conducted an informal survey among experts of WM to ensure that we did not inadvertently overlook important findings, or overrate the importance of some findings.

Notwithstanding these efforts, we cannot be fully confident that our selection and our ratings of benchmark findings is free from bias. The best way to eliminate any remaining bias is through an open discussion among all experts in the field, and subsequent revision of the set of benchmarks. To that end we established an online forum for discussion of the benchmarks, and we invite all scholars of WM to join the Benchmarks Team for preparing a revised set of benchmarks in four to five years[7].

We believe that a set of benchmarks can be useful for a field in several ways. First, on a purely descriptive level benchmarks provide a snapshot of the state of empirical knowledge in the field, concentrated on findings that are robust and reasonably general. As such, it can help students and researchers to navigate the wealth of empirical results and give them information on how well supported and general each finding is. Benchmarks also reveals gaps in our knowledge: Even a cursory glance at the overview tables in Appendix C and D shows that a large number of benchmarks is so far established only in one content domain (primarily the verbal domain) and only in young adults, and many benchmarks have been studied only with a limited set of experimental paradigms. Researchers could use benchmarks as a guide to systematically extend the generality of findings they consider to be of theoretical importance.

A second use of benchmarks is to serve as a sanity check for theoretical efforts. Any new theory in a field should endeavor not to contradict benchmark findings. Obviously, we want much more from a good theory than an account of existing, well established findings: Theories should offer a compelling mechanistic explanation that help us understand the phenomena, and they should imply new predictions. Therefore, it is important that the rapid accumulation of empirical findings does not

[7] The online forum is here: https://wmbenchmarks.wordpress.com/

unduly tie down theoretical innovations: A new theoretical proposal should not be rejected just because there exists an empirical result that contradicts it. If theoretical progress is not to be stifled by empirical constraints, theorists must be allowed to ignore some findings, at least for a while. An agreed-upon set of benchmarks therefore has a liberating effect on theory building: It limits the number of empirical findings that a theorist should aim to not contradict, and provides guidance on their priority. We suggest that a theory that handles benchmark results, but is contradicted by other more “niche” results, must be accorded greater credit than a theory that fails to handle benchmarks but accounts for some other results that are arguably of lesser importance.

A third, arguably most ambitious use of benchmarks is as basis for the development of increasingly comprehensive theories, that is, theories that can explain a larger proportion of empirical findings relevant for a topic such as WM. Benchmarks facilitate this development in two ways. First, they provide well-defined explanatory targets for theorists who aim to develop a new unified theory or improve an existing one. Second, they provide a common empirical ground for evaluating theories. Evaluating theories against an entire set of diverse findings, such as the present set of benchmarks, raises a number of challenges: How do we measure the goodness of a theory or computational model in explaining a set of benchmark findings? There has been much progress in measuring the goodness of fit of formal models to individual data sets (for an overview see Lewandowsky & Farrell, 2011; Shiffrin, Lee, Kim, & Wagenmakers, 2008), but these methods do not easily generalize to the problem of evaluating theories against an entire set of findings from different paradigms, with different dependent variables, each of which is represented by multiple data sets.

A related problem concerns what it means for a theory to explain a benchmark. Many theories and models of WM are highly flexible, such that they are compatible with many findings but do not predict these findings – they are equally compatible with the absence of the effect in question, or even with an effect in the opposite direction. This flexibility is to some extent inherent in verbal theories because they are often formulated vaguely, leaving much room for interpretation. Computational models avoid this vagueness, but they do not entirely escape the problem of flexibility: Most computational models have free parameters that can enable them to be compatible with an effect, its absence, and its opposite. This flexibility is substantially reduced when we aspire for a model to account for several benchmark findings with a common set of parameter values. Our proposed set of benchmarks should facilitate efforts towards more comprehensive computational models. We hope that our proposal for benchmarks will, among other things, instigate a discussion on how to evaluate and compare the empirical adequacy of models across a broad range of findings from different experimental paradigms.

A good starting point for such a discussion could be the proposal of Wills and Pothos (2012) for how to assess the adequacy of computational models across several experiments. They propose as a criterion the number, or the proportion, of *ordinal*, *irreversible*, and *penetrable* successes of a model in explaining the findings. An ordinal success is defined as the accurate prediction of the ordinal pattern of dependent variables across experimental conditions, that is, getting the direction of the experimental effects right. A model's success is irreversible if the modelers commit to holding parameter values constant across multiple applications of a model to different experiments and different phenomena, so that the model's explanatory success cannot be undone by a later change in parameter values (e.g., when fitting the model to a new data set). A model is penetrable to the degree that it is easy to apply, and explained in psychological terms, so that in addition to predicting the data, it also advances our understanding of the phenomena.

Successful theories and models of WM are likely to differ in their scope, either because the theory implies that certain benchmarks are not relevant (e.g., when a theory ascribes some benchmarks to episodic long-term memory rather than to WM), or because the theorist decides to "start small" and aim for a detailed explanation of a coherent subset of WM benchmarks (e.g., only findings on serial recall). To accommodate scope differences, Wills and Pothos (2012) propose to evaluate models not only by the absolute number of successes, but also by the proportion of successes within their scope. One prerequisite for that criterion is that a theory or model includes a clear definition of its scope. In addition, we argue that the scope of a model should also be well justified, rather than "gerrymandering" benchmarks out of a model's scope simply because the model cannot account for them. For instance, limiting a model's scope to serial recall is convincing to the extent that benchmark findings from serial-recall tests differ in important ways from those obtained with other paradigms

(Bhatarah et al., 2009), and limiting a model's scope to verbal materials is convincing to the extent that benchmarks differ between WM for verbal and non-verbal contents (Hurlstone et al., 2014).

In closing, we wish to emphasize that the set of benchmarks we proposed here is not intended as a definitive summary of our empirical knowledge about WM. Rather, we see it as a first proposal of how to organize and prioritize the wealth of data that we have accumulated so far; it will prove useful to the degree that researchers use it to guide and evaluate theoretical efforts; it is subject to revision, and it will obviously have to be updated in light of new empirical discoveries. We also do not wish to constrain theory development to address only those findings that we identified as benchmarks. A good theory should not only explain as much as possible of what we already know but also make new predictions. Tests of these predictions generate new findings that, once they are firmly established, become future benchmarks.

Author Note

The preparation of this manuscript was facilitated by two grants from the Swiss National Science Foundation that funded two workshops (grant numbers IZ 32Z0 150896 and IZ 32Z0 160296). We gratefully acknowledge the contribution of Jeff Rouder during the early phase of this project.

Appendix A – Survey Results

The overall battery of survey questions comprised 110 candidate benchmarks identified during the 2014 meeting of the Benchmarks Team. Each participant received a random subset of 71 of those items, subject to the following constraints: (1) Subordinate findings were always presented together with their corresponding main finding. For example, the superordinate potential benchmark “Monotonic decrease of accuracy with increasing set size (list length)” would be presented together with one randomly chosen item out of 3 further subordinate candidates (e.g., “Set-size effect is found in change detection (Luck & Vogel 1997), change localization (van den Berg et al., 2008)”). There were 23 such clusters of items. (2) Candidates that received particularly variable ratings by the Benchmark Team were presented to all participants. (3) Respondents could add further proposals in free-text form. (4) A single competence item was included for all participants that queried “who introduced the magical number seven, ‘plus or minus two’, into the literature”? (All respondents who completed the item answered it correctly).

For each candidate item, participants chose one of 4 response options, corresponding to the three levels of benchmark (A, B, or C), plus the “not a benchmark” option.

There were 156 participants who contributed to the survey, of whom 51 were complete. Because the incompletes also contained useful data, we report the analysis on the full set of participants. Owing to the random sampling of items and the large number of incompletes, the total number of responses differed considerably across items, ranging from 6 to 81, with a mean of 36.88 (median 50).

The table below reports the data for the 74 items that received 20 or more responses, in descending order of their endorsement (in percentages) as benchmark of type A. The second column gives the number of the benchmark in the text and the reference table; survey items not included in the final list of benchmarks are marked "N"; when a survey item was subsumed under a benchmark as a special case or qualification, we set the benchmark number in parentheses. Frequencies of responses are given as numbers out of N responses to a given item, and as percentages of that N.

	BM number	Candidate benchmark	A	B	C	Not BM	N	%A	%B	%C	%Not BM
1	11.3	Lexicality, word frequency, and bigram/phonotactic frequency have effects on recall accuracy.	17	4	4	1	26	0.65	0.15	0.15	0.04
2	1.1	Monotonic decrease of accuracy with increasing set size (list length).	50	17	9	5	81	0.62	0.21	0.11	0.06
3	5.1 + 5.2	Dual-task studies show some domain-general and some domain-specific mutual impairment of two concurrently maintained memory sets, or of maintenance and a concurrent processing task	37	10	8	5	60	0.62	0.17	0.13	0.08
4	2.1	Gradual forgetting of small sets of items over increasing retention interval filled with interpolated activity in the same domain	19	6	4	5	34	0.56	0.18	0.12	0.15
5	(1.1)	Set-size effect is found in change detection and change localization	16	4	7	2	29	0.55	0.14	0.24	0.07
6	11.5	Immediate serial recall improves if the same list is used on multiple trials (Hebb, 1961)	30	13	7	5	55	0.55	0.24	0.13	0.09
7	9.2.1	Grouping enhances recall relative to ungrouped lists	29	15	7	3	54	0.54	0.28	0.13	0.06
8	1.2	Monotonic increase of mean RT with set size	39	24	9	5	77	0.51	0.31	0.12	0.06
9	11.2	Sentence superiority: lists of words forming a sentence are recalled better than lists of random words, or jumbled sentences	13	4	5	4	26	0.5	0.15	0.19	0.15

	BM number	Candidate benchmark	A	B	C	Not BM	N	%A	%B	%C	%Not BM
10	13.2	Performance in some short-term memory tasks is preserved even when brain damage has caused profound amnesia in other forms of memory	12	7	3	2	24	0.5	0.29	0.12	0.08
11	1.3	The number of items recalled in serial recall of verbal lists under articulatory suppression is about four	33	21	10	6	70	0.47	0.3	0.14	0.09
12	7	Lists of words with more syllables are recalled worse than lists of words with fewer syllables	24	14	8	7	53	0.45	0.26	0.15	0.13
13	12.1	Performance on WM tasks (and in memory tasks in general) correlates positively for all kinds of tasks and materials	12	9	4	2	27	0.44	0.33	0.15	0.07
14	4.3	Intrusions of currently (or permanently) irrelevant memory contents occur in various paradigms.	26	17	10	7	60	0.43	0.28	0.17	0.12
15	(3.1)	With output order controlled, there tend to be equal amounts of primacy and recency and the overall curve tends to be symmetric	24	17	8	8	57	0.42	0.3	0.14	0.14
16	6.1	To-be-ignored speech or tones during visual presentation of verbal lists for serial recall (or during the retention interval) impair recall	10	6	5	3	24	0.42	0.25	0.21	0.12
17	N	Memory sets of more complex items are recalled worse than sets of simpler items	22	16	10	5	53	0.42	0.3	0.19	0.09

	BM number	Candidate benchmark	A	B	C	Not BM	N	%A	%B	%C	%Not BM
18	13.3	Parietal BOLD signal and contralateral delay activity (CDA) track memory set size up to about 3 to 4 simple items and then reach a plateau	20	14	7	10	51	0.39	0.27	0.14	0.2
19	10.2	Item-switch cost in tasks requiring multiple successive retrievals of individual items: Latency of access to a new item in a memory set takes longer than repeated access to the same item	23	16	10	10	59	0.39	0.27	0.17	0.17
20	9.1.3	Non-temporal sources of saliency (e.g., an item in a unique color) improve recall of the salient item	20	19	7	6	52	0.38	0.37	0.13	0.12
21	1.1	Set-size effect is found for proportion correct in serial recall, complex span, WM updating, immediate free recall, and probed recall	9	10	4	2	25	0.36	0.4	0.16	0.08
22	(13.3)	The plateau [of the CDA] is reached at smaller set sizes in individuals with worse memory performance.	18	12	8	13	51	0.35	0.24	0.16	0.25
23	5.2.4	Cognitive-load effect of distractor processing in the retention interval: Performance on serial-recall tasks and single-item recall and recognition tasks declines with increasing ratio of time needed for distractor task to time available	20	19	13	7	59	0.34	0.32	0.22	0.12
24	4.1.1	Transposition gradients obey the locality constraint (over a wide range of scales): transpositions with close	10	9	7	4	30	0.33	0.3	0.23	0.13

	BM number	Candidate benchmark	A	B	C	Not BM	N	%A	%B	%C	%Not BM
		neighbors are more likely than with more distant neighbors									
25	13.4	There are both activation-based and non-activation-based neural signatures of WM.	16	14	10	11	51	0.31	0.27	0.2	0.22
26	N	Change detection: Hit rate decreases monotonically with set size, false-alarm rate increases monotonically with set size	8	8	5	5	26	0.31	0.31	0.19	0.19
27	9.2.2	There are primacy and recency effects within groups	16	21	13	4	54	0.3	0.39	0.24	0.07
28	9.1.1	Temporal isolation effects are found in most list-recall paradigms: isolated items are recalled better	8	9	7	4	28	0.29	0.32	0.25	0.14
29	2.3	No forgetting of sequentially presented verbal lists over retention interval filled with repetitive activity	18	21	14	10	63	0.29	0.33	0.22	0.16
30	(4.3)	Serial recall: some errors are protrusions (i.e., intrusions of items from preceding list); they tend to come from the corresponding serial position of the recalled sequence of the preceding trial	17	21	18	4	60	0.28	0.35	0.3	0.07
31	(8.1.1)	Phonological similarity dominates over visual similarity whenever the material is nameable.	15	21	8	11	55	0.27	0.38	0.15	0.2
32	4.5	Distributions of errors in continuous recall are distinctly non-normal, e.g. they are more peaked and have fatter tails	14	14	13	11	52	0.27	0.27	0.25	0.21

	BM number	Candidate benchmark	A	B	C	Not BM	N	%A	%B	%C	%Not BM
		than a normal (von-Mises) distribution									
33	(2.1)	Partial forgetting over some seconds even with no interpolated activity if visual stimuli are presented briefly in arrays or in a rapid sequence	17	28	13	6	64	0.27	0.44	0.2	0.09
34	12.4	In factor analyses of memory tests there is a factorial separation of "primary memory" and "secondary memory"	13	17	10	10	50	0.26	0.34	0.2	0.2
35	N	Attentional capture (of visual search, of eye movements) by contents of WM	13	14	17	7	51	0.25	0.27	0.33	0.14
36	12.3	Correlations between tests with materials from different content domains (verbal vs. visual-spatial) are larger in complex span than simple span	13	23	7	8	51	0.25	0.45	0.14	0.16
37	N	Change detection for changes within a category of stimuli is harder than between categories (Awh, Barton, & Vogel, 2007)	13	18	13	8	52	0.25	0.35	0.25	0.15
38	3.5.2	Semantic clustering during output: When words on a list can be grouped into semantic categories, people tend to recall words from the same category together	13	28	8	4	53	0.25	0.53	0.15	0.08
39	4.4	Ranschburg effect: in forward serial recall people struggle to recall an item twice even though it was repeated in the list	6	9	5	5	25	0.24	0.36	0.2	0.2

	BM number	Candidate benchmark	A	B	C	Not BM	N	%A	%B	%C	%Not BM
40	13.1	Dissociable sensory and association networks are recruited by broad classes (domains) of WM content: Dorsal frontal-dorsal parietal for spatial; ventral frontal-ventral temporal for object; ventral frontal-dorsal temporal/ventral parietal for verbal materials	6	8	8	5	27	0.22	0.3	0.3	0.19
41	9.2.3	Tendency to confuse items in same within-group positions (“interpositions”) (Ryan, 1969).	12	22	14	6	54	0.22	0.41	0.26	0.11
42	(1.1)	Deviation of response from target feature in continuous recall: Distribution gets wider with increasing set size, including an increasing proportion of responses very far from the target...	17	20	30	11	78	0.22	0.26	0.38	0.14
43	(12.2)	A factorial differentiation between verbal and visual-spatial WM tasks, and between simple and complex span tasks, has been found in children	11	17	12	11	51	0.22	0.33	0.24	0.22
44	13.4	Short-term retention is possible in the absence of measurable neurally active representation.	11	14	9	17	51	0.22	0.27	0.18	0.33
45	11.4	Serial recall of lists tends to reproduce well-learned transition probabilities between items	12	17	19	9	57	0.21	0.3	0.33	0.16
46	N	Regularity in the memory set, or links to knowledge, enable the representation of abstract summaries ("gist"), while more detailed information is lost	12	25	9	11	57	0.21	0.44	0.16	0.19

	BM number	Candidate benchmark	A	B	C	Not BM	N	%A	%B	%C	%Not BM
47	1.3	People can remember about three individual words and/or learned word-pairs under articulatory suppression in serial recall under non-serial scoring	14	27	19	9	69	0.2	0.39	0.28	0.13
48	8.1.2	Phonologically dissimilar items on mixed lists are recalled as well as or better than on purely dissimilar lists	11	22	12	10	55	0.2	0.4	0.22	0.18
49	2.2	Forgetting over filled retention interval is much diminished (Keppel & Underwood, 1962) but perhaps not entirely eliminated (Baddeley & Scott, 1971) when proactive interference is minimized	6	14	6	4	30	0.2	0.47	0.2	0.13
50	9.2.4	Effect of grouping on recall latency: longer recall time for first item in a group	10	19	17	8	54	0.19	0.35	0.31	0.15
51	N	In factor analyses of data form the continuous-recall paradigm for visual working memory, there is a factorial separation of "probability of recall" and "precision parameters (measured through the mixture model of Zhang & Luck, 2008)	9	17	7	17	50	0.18	0.34	0.14	0.34
52	N	In factor analyses of working memory tests, there is more factorial differentiation at the higher end of the ability distribution	9	8	18	16	51	0.18	0.16	0.35	0.31
53	8.2	Accuracy in change detection decreases with smaller magnitudes of change	9	16	22	5	52	0.17	0.31	0.42	0.1

	BM number	Candidate benchmark	A	B	C	Not BM	N	%A	%B	%C	%Not BM
54	N	A recency effect is also observed in continuous-distractor free recall, although on a lower level than in immediate free recall	9	19	14	11	53	0.17	0.36	0.26	0.21
55	(9.2.1)	Group sizes of 3 result in best performance	9	23	17	5	54	0.17	0.43	0.31	0.09
56	12.2	Correlations between tests of working memory tend to be higher within than between domains (verbal vs. visual-spatial)	4	10	6	4	24	0.17	0.42	0.25	0.17
57	(8.1.1)	Sometimes phonological similarity is beneficial for item memory (boundary conditions to be established)	9	19	16	11	55	0.16	0.35	0.29	0.2
58	12.5	There is a correlation between (verbal) memory span and both articulation speed and retrieval speed in children, older adults, and normal young adults that helps explain both individual differences and developmental differences in memory span	8	21	15	6	50	0.16	0.42	0.3	0.12
59	N	Mean of the features of a group or ensemble of objects in a visual array biases the estimation of the features of individual elements towards that mean	9	21	17	10	57	0.16	0.37	0.3	0.18
60	N	Trial-to-trial confidence in continuous recall is predictive of memory performance (Rademaker et al., 2012)	8	11	13	20	52	0.15	0.21	0.25	0.38

BM	number	Candidate benchmark	A	B	C	Not BM	N	%A	%B	%C	%Not BM
61	(11.5)	Hebb effect is reduced when temporal grouping changes across repetitions	7	23	16	9	55	0.13	0.42	0.29	0.16
62	9.1.2	Temporal isolation effects are absent in forward serial recall	3	7	6	8	24	0.12	0.29	0.25	0.33
63	4.1.2	Ratio of fill-in errors to in-fill errors (given a list ABC, recalling ACB is a fill-in error, and ACD is an in-fill error): Fill-in errors outweigh in-fill errors, and the ratio tends to be 2:1	6	16	21	12	55	0.11	0.29	0.38	0.22
64	6.2	Changing-state irrelevant sound (sequence of different tones) is more disruptive for serial recall than steady-state sound (repetition of the same tone)	3	7	11	7	28	0.11	0.25	0.39	0.25
65	5.1.3	Spatial memory is more susceptible to disruption by an added verbal memory set than vice versa	6	26	21	7	60	0.1	0.43	0.35	0.12
66	N	"Probability of recall" [from the Zhang & Luck mixture model] forms a general factor but "precision" does not.	5	11	13	21	50	0.1	0.22	0.26	0.42
67	N	The CDA is (largely) insensitive to stimulus complexity, with the exception of polygons...	5	13	13	20	51	0.1	0.25	0.25	0.39
68	(9.2.1)	Grouping effects are amplified with auditory presentation.	5	15	24	10	54	0.09	0.28	0.44	0.19
69	N	Relative distribution of transpositions, intrusions, omissions, repetitions,	5	19	15	16	55	0.09	0.35	0.27	0.29

	BM number	Candidate benchmark	A	B	C	Not BM	N	%A	%B	%C	%Not BM
		protrusions: tend to depend on stimulus material									
70	N	People maintain item-by-item memory of small arrays of objects, and statistical properties of larger arrays	5	21	19	12	57	0.09	0.37	0.33	0.21
71	N	Retention of verbatim information for last clause in a sentence, but only meaning of previous clauses, ...	5	13	19	20	57	0.09	0.23	0.33	0.35
72	8.1.4	Adults and older children show primarily phonological similarity effects on serial-order memory for pictures of nameable objects, whereas children younger than 6 years show only visual similarity effects	4	18	14	14	50	0.08	0.36	0.28	0.28
73	N	Effects of grouping are observed not only when group boundaries are marked by longer pauses, but also when group boundaries are marked by pitch or by voice (male vs. female)	3	17	15	19	54	0.06	0.31	0.28	0.35
74	(1.3)	In serial recall, number of items recalled in correct position increases to about three, then decreases again with increasing list length	1	7	7	6	21	0.05	0.33	0.33	0.29

Appendix B: Benchmarks with C Ratings

Benchmark 3.3.2: Particularly Fast Access to the Last List Item (C). There is some indication that the most recently presented item in a study list is identified much faster than other items. For example, by systematically varying the time at which participants were asked to respond, Wickelgren, Corbett, and Dosher (1980) found that participants were sooner able to match the most recent study item to the test item than study items presented earlier in the list (Figure A1; for a more comprehensive review see McElree, 2006). This finding qualifies as benchmark because of its replicability, and because it plays an important role in the case for a focus of attention in working memory (McElree, 2006; Oberauer & Hein, 2012; Vergauwe et al., 2016). Because it has limited generality, and qualifies benchmark 3.3.1, we assign it a low priority (C), although survey participants tended more towards a B rating (7/15).

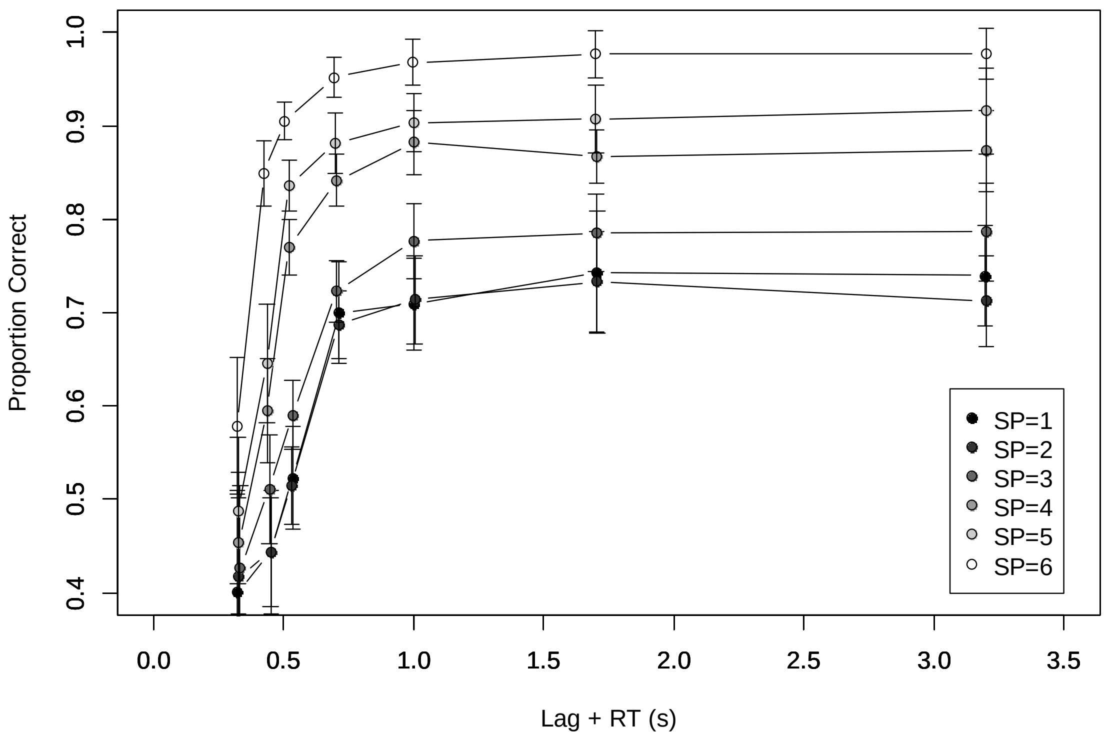

Figure A1: Speed-accuracy trade-off curves for item recognition: Accuracy of matching probes in different serial positions (SP) is plotted as a function of decision time until the response deadline (Öztekin & McElree, 2010)

Benchmark 3.3.3. Serial-Position Effects on Recall Latencies (C). Recall latency is measured as the time between successive keypresses in recalling digits or letters (e.g., Farrell & Lewandowsky, 2004), inter-item pauses in spoken recall (e.g., Murdock & Okada, 1970), or times between successive onsets or offsets in typed recall (Thomas, Milner, & Haberlandt, 2003). Latency patterns show regularities for particular tasks, but do vary across tasks (see Figure A2). In the case of serial recall, an extended pause is left before outputting the first item in response to the recall cue, and latencies then follow an inverse U-shaped function, with responses slowing and then speeding across successive serial positions (Maybery et al., 2002). Probed recall shows a similar inverse-U shape when latencies are plotted by input position (Sanders & Willemsen, 1978). These effects are replicable and show some generality across paradigms, but they have played only a minor role in theorizing about serial recall so far; therefore, we propose them as a low-priority benchmark (C).

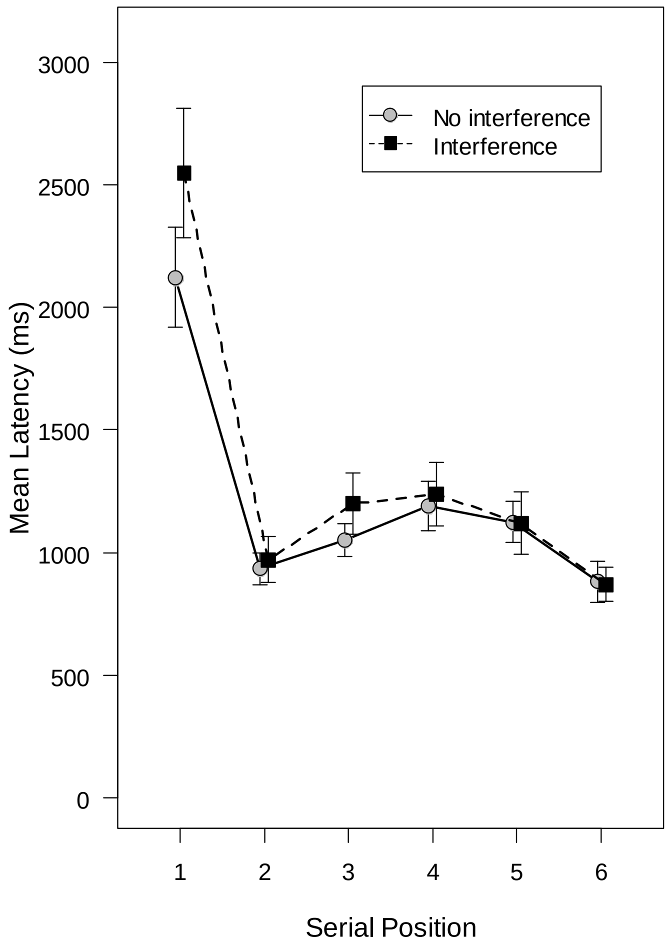

Figure A2: Serial-position effect on recall latencies for immediate forward serial recall (No interference) and for forward serial recall after reading aloud four distractor digits (Interference) (Farrell & Lewandowsky, 2004, Exp. 2)

Benchmark 3.5.2: Semantic Clustering in Free Recall (C). In addition to the tendency to recall items in forward order (Benchmark 3.4.3), participants tend to cluster semantically related items together at output in free-recall tasks (Bousfield, 1953; Jenkins & Russell, 1952). In addition, people subjectively organize lists of words during encoding, and this organization is reflected in their patterns of retrieval (Mandler, 1967; Tulving, 1962). Semantic relatedness effects have also been observed in the retrieval of lists of ostensibly unrelated items (Howard & Kahana, 2002): Semantically more similar pairs of words are more likely to be recalled in immediate succession. Semantic clustering effects are informative for theories of free recall (e.g., Healey & Kahana, 2014), but as they are limited to free recall of words, we ranked this benchmark as C, although the most frequent survey response was B (28/53).

Benchmark 4.1.2. Fill-in effect in serial recall (C). Transposition errors in serial recall (including complex span) exhibit a systematic pattern of sequential dependency: If an item *i* is recalled a position too early (e.g., when given sequence ABC, starting recall with B instead of A), recall of item *i*-1 (e.g., B<u>A</u>; a fill-in error) is more likely at the next output position than item *i*+1 (e.g., B<u>C</u>; an infill error). Available data show that fill-in errors outweigh infill errors by a ratio of approximately 2:1—a result dubbed the *fill-in effect* (Farrell, Hurlstone, & Lewandowsky, 2013; Guérard & Tremblay, 2008; Henson, 1996; Surprenant, Kelley, Farley, & Neath, 2005). This finding has been important in informing theories of how serial order is represented. At the same time, it is specific to forward serial recall, and has been firmly established only for verbal materials (for a potential boundary condition see Osth & Dennis, 2015). Therefore, we rate it as C, in agreement with the modal survey response (21/55).

Benchmark 4.2. Serial position effects on error-types in serial recall (C).

Errors in serial recall sometimes involve the loss of item information. These *item errors* can be divided into *intrusions* (reporting items not part of the study sequence), *omissions* (failure to report any item in a position), and *repetitions* (incorrect report of an item already produced). The frequency of item and transposition errors varies according to serial position: item errors increase with serial position, whereas transpositions increase initially, but decrease thereafter (Avons & Mason, 1999; Guérard & Tremblay, 2008; Henson et al., 1996). This finding places constraints on models of memory for serial order, but it is specific to forward serial recall; hence we rate it as C.

Benchmark 4.4. Ranschburg effect in serial recall (C).

In serial recall, people tend to fail to report an item twice when it was repeated in a sequence—the *Ranschburg effect*. This effect occurs when serial recall is compared under two conditions: In the repetition condition, study sequences contain two occurrences of the same item separated by several intervening items, whilst in the control condition sequences always contain unique items. The typical finding is that recall of the second occurrence of a repeated item is impaired, relative to items in corresponding positions in the control condition (Crowder, 1968; Henson, 1998a; Jahnke, 1969). We regard this finding a benchmark because it is replicable and forms a key piece of evidence for the assumption of response suppression in serial recall (Henson, 1998a). At the same time, its generality is limited to (verbal) serial recall; so we assigned it rating C, although survey respondents were leaning more towards B (9/25).

Benchmark 5.1.3. Asymmetric Effects Between Verbal and Spatial Sets (C). This benchmark qualifies benchmark 5.1.1: Visual-spatial memories are more susceptible to disruption by an added verbal memory set than vice versa. Morey and colleagues (Morey et al., 2013) paired various visual and verbal memory sets and measured decreases in capacity estimates from a single-set baseline as the number of items in the simultaneously-held cross-domain set increased. Whereas visual memory capacity reliably shrank as verbal memory load increased, verbal memory capacity was far less impaired by the concurrent visual memory load. This pattern has been replicated for serial verbal and spatial memory tasks (Morey & Mall, 2012; Morey & Miron, 2016). We rated this emerging finding as a C benchmark because its consistency and generalizability are still under consideration. This rating was further justified by the survey respondents, who endorsed ratings of B (26/60) and C (21/60) at similar rates.

Benchmark 5.2.3. Processing of material from same or different category as the memory materials (C). It is typically found that, when the memory items and processing items come from the same domain, the disruption is attenuated when the memory items and processing items come from different categories within that domain (Conlin & Gathercole, 2006; Conlin, Gathercole, & Adams, 2005; Turner & Engle, 1989). In complex span tasks, memory for words is better when numbers are processed concurrently than when words are processed concurrently and, in the same way, memory for numbers is better when words are processed concurrently than when numbers are processed concurrently (Conlin et al., 2005; Turner & Engle, 1989). Survey respondents rated this benchmark equally often B and C (6/15 of the respondents for each rating), but because the finding is quite specialized and has only been studied in a limited set of experimental paradigms, we rated this benchmark as C.

Benchmark 6.3 Auditory deviant effect (C)

The irrelevant sound effect (Benchmarks 6.1 and 6.2) has been contrasted with the finding that a deviant auditory distractor during visual presentation of verbal lists impairs memory (Hughes et al., 2007; Lange, 2005; Sörqvist, 2010). Participants read a list of items for recall whilst ignoring an irrelevant sound sequence. Serial recall is impaired when, in a few trials, one token in the sequence is unexpectedly spoken in a different voice, or is out of rhythm (e.g., Hughes et al., 2007; Sörqvist, 2010). In contrast to the changing-state effect, the disruptive effect of a deviant does not only occur in serial recall, but also in a missing-item task (Hughes et al., 2007). We assigned this benchmark a C rating because the auditory deviant effect constitutes a relatively novel finding in the WM literature that is of high theoretical leverage but for which robustness and generality still need to be ascertained.

Benchmark 8.1.4. Development of phonological similarity effect (C). Phonological similarity effects are subject to a developmental trend. Whereas adults and older children show phonological similarity effects on serial-order memory for pictures of nameable objects, this effect only emerges around age 7, and is not observed in younger children (Hayes & Schulze, 1977; Hitch, Woodin, & Baker, 1989; S. Palmer, 2000). Rather, younger children's recall is primarily dominated by visual confusions (Hayes & Schulze, 1977; Hitch, Woodin, et al., 1989). This finding is a benchmark because it is well replicated, and it is theoretically important because it has been taken as reflecting the development of the use of verbal rehearsal, although a recent analysis (Jarrold & Citroen, 2013) calls this interpretation into question. We assigned this benchmark a lower priority (C) because it qualifies the more general benchmark 8.1, and because theories of WM need to explain this finding only if they aim to explain the development of WM. In agreement with this assessment, survey responses were about evenly distributed between B (18/50), C (14/50) and "not a benchmark" (14/50).

Benchmark 8.1.5. Effect of visual similarity on serial recall (C). Visual similarity also has detrimental effect on serial-order memory of visual stimuli. A visual similarity effect has been found for words varying in orthographic similarity (Logie, Della Sala, Wynn, & Baddeley, 2000; Saito, Logie, Morita, & Law, 2008), visual patterns (Avons & Mason, 1999), and faces (Smyth et al., 2005). Jalbert, Saint-Aubin, and Tremblay (2008) varied the similarity of colored squares in a serial reconstruction task, and found that similarity in color hindered memory for both order and location. The nature of visual stimuli does not easily permit the separate examination of item and order memory as in phonological memory (Benchmark 8.1.1). Nonetheless, Saito et al. (2008) found that the visual similarity of Kanji characters primarily affected ordering errors, with no significant effect on item errors. We nominate the visual-similarity effect as a benchmark because it generalizes the theoretically highly important phonological-similarity effect. At the same time, we assign it lower priority (C) because there are only a few studies demonstrating the effect, and it is limited to serial-order paradigms. We have only sparse survey data on this benchmark, with B (5/12) the most frequent response.

Benchmark 8.2. Effects of Item-Probe Similarity in Recognition and Change Detection (C)

In the change-detection test of visual WM, an observer reports whether or not a change occurred either in a single probed item or in a whole array (see Figure 1H). Until recently, it was common to use highly distinguishable items as stimuli, such as colors deliberately chosen to be far separated in color space. More recent studies have instead varied the magnitude of the change in order to obtain a richer characterization of behavior (Bays et al., 2009; Devkar, Wright, & Ma, 2015; Keshvari, Van den Berg, & Ma, 2012; Keshvari, Van den Berg, & Ma, 2013; Lara & Wallis, 2012; Van den Berg et al., 2012). At a qualitative level, these studies have universally found that accuracy decreases smoothly with decreasing magnitude of change (increasing similarity), not only in change detection but also in change localization (Van den Berg et al., 2012), and across species (Devkar et al., 2015; Heyselaar, Johnston, & Pare, 2011; Lara & Wallis, 2012). Beyond this qualitative finding, the theoretical importance of this benchmark lies in the quantitative shapes of the psychometric curves (see Figure A3): The behavioral richness obtained by varying both set size and change magnitude can be effectively exploited for comparing models of WM, such as slot models and noise-based models (Devkar et al., 2015; Keshvari et al., 2013). We regard this finding as a category C benchmark: So far, evidence for it is limited to one paradigm and one content domain, but models aiming at explaining visual change detection must get it right.

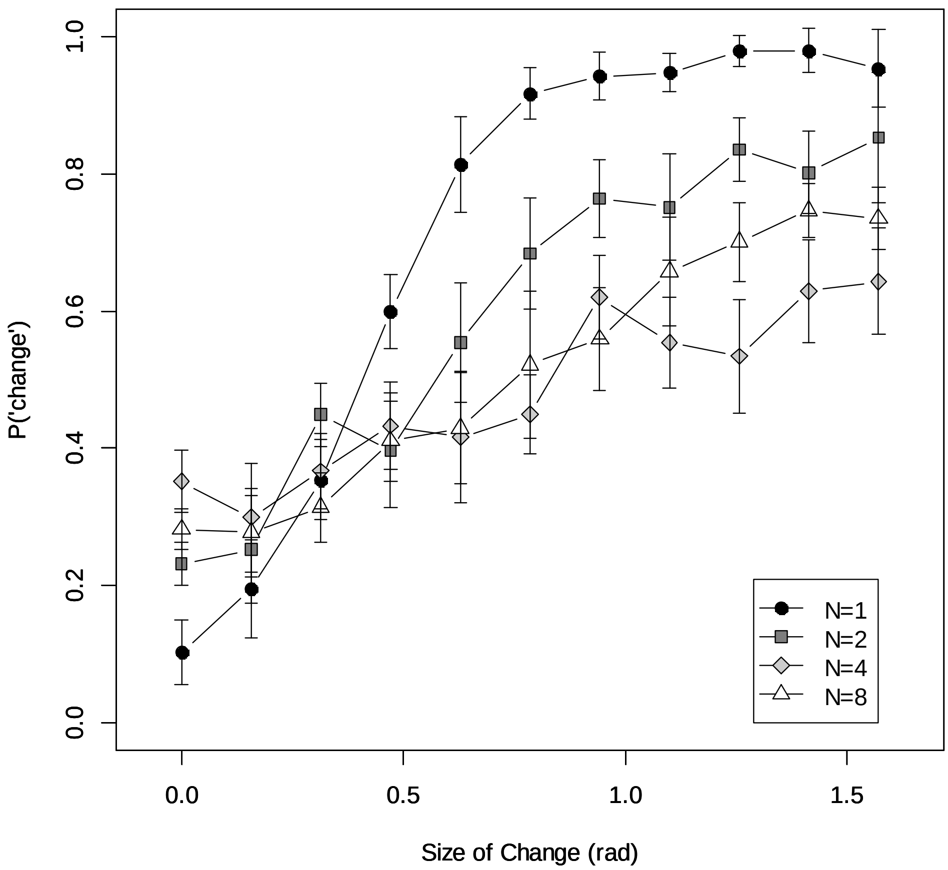

Figure A3: Probability of detecting a change as a function of memory set size (N) and the size of the change in a change-detection task (Van den Berg et al., 2012).

Benchmark 9.1.2. Absence of temporal isolation effects in forward serial recall (C). Early experiments on temporal distinctiveness (Benchmark 9.1.1) examined memory for serial order. Studies using predictable presentation schedules found an advantage for more isolated items (Neath & Crowder, 1990, 1996). Several subsequent studies found that the apparent effect of temporal isolation disappears when random presentation schedules are used (Lewandowsky, Brown, Wright, & Nimmo, 2006; Nimmo & Lewandowsky, 2005, 2006). As this finding qualifies benchmark 9.1.1, we rate it as C. In support, most survey respondents rated this finding as B (7/24), C (6/24) or "not a benchmark" (8/24).

Benchmark 11.2 Sentence superiority effect (C)

One phenomenon reflecting the effect of knowledge is the observation that immediate memory for a series of words is greatly enhanced when they form a sentence. For example, Brener (1940) estimated that span was more than ten words when they were presented in sentences as compared with less than six when they were ordered randomly. Baddeley, Hitch, and Allen (2009) showed that the sentence superiority effect is observed with strictly controlled materials and persists when participants are required to perform concurrent articulation or even a more demanding concurrent task. Sentence recall clearly benefits from knowledge of syntactic and semantic information. This effect may be different from the chunking benefit (11.1) in that it does not rest on participant's familiarity with the specific sentences presented, but with more general linguistic knowledge. The sentence-superiority effect is a benchmark because it informs theories on how general long-term knowledge assists memory for order; at the same time it is limited to verbal materials, and to memory for serial order, hence we rate it as C, although the most frequent rating in the survey was A (13/26).

Benchmark 11.4. Regularization (C)

Another observation associated with prior learning is the occurrence of what can be termed *regularization errors* in immediate recall. In general terms these are cases where the content of errors is biased in the direction of increasing the familiarity of what is recalled. Bartlett (1932) first drew attention to distortions of this type in long-term episodic memory, and they are also found in short-term memory, though this is much less well documented. For example, errors in the immediate recall of nonwords often involve the production of real words (Jefferies, Frankish, & Lambon Ralph, 2006), and extensive experience of recalling sequences of nonwords obeying an artificial grammar results in errors biased towards respecting the learned grammar (Botvinick & Bylsma, 2005). Regularization effects have been informative for computational models of serial recall (Botvinick & Plaut, 2006), but their empirical foundation is still relatively thin, so that we rate it as C, although the most frequent survey response was B (25/57).

Benchmark 13. 4 Short-Term Retention Without Measurable Neurally Active Representations (C)

Correlates of active neural firing provide the most direct means to measure WM-related phenomena in the brain. However, there is mounting evidence that information can be retained in the short-term in the absence of detectable neural activity related to the retained information. Strong evidence for this comes from a pair of studies that used machine learning algorithms to track items in WM while attention was endogenously shifted between the items (LaRocque, Lewis-Peacock, Drysdale, Oberauer, & Postle, 2013; Lewis-Peacock, Drysdale, Oberauer, & Postle, 2011). In both BOLD and EEG signals, attended items were well tracked by these algorithms, but unattended items could not be identified. Critically, if attention was switched to an initially unattended item, the newly attended item could suddenly be detected. These data demonstrate that unattended items are not lost, they are simply not detectable through correlates of neural activity. Data from single-unit recordings provide evidence that such phenomena may not be due to limits of non-invasive human recording techniques. For example, neural activity related to WM can disappear during the start of a retention interval, only to ramp up near the time that that information is needed for a decision (Barak, Tsodyks, & Romo, 2010; E. K. Miller, Erickson, & Desimone, 1996; Romo, Brody, Hernandez, & Lemus, 1999). The observed sustained activity related to WM can be an artifact of averaging over trials, each of which show only intermittent bursts of stimulus-related neural activity (Lundqvist et al., 2016). Taken together, this research suggests that information can be retained in the short-term without measurable sustained neural firing (Stokes, 2015). These findings are theoretically informative because they support a growing body of computational models that suggest that short-term retention is at least partly mediated by short-lived synaptic plasticity (Barak & Tsodyks, 2007; Lundqvist, Herman, & Lansner, 2011; Mongillo, Barak, & Tsodyks, 2008; Sugase-Miyamoto, Liu, Wiener, Optican, & Richmond, 2008).

The circumstances under which short-term retention is supported by different neural codes (e.g. sustained neural firing, activity-silent mechanisms) remain to be elucidated. Considering the various forms of neural codes will be important to provide neural plausibility to any model of WM and its interaction with attention and long-term memory, and to explain findings that cannot be readily accommodated by a single representational coding scheme. Nevertheless, given that the lion’s share of neural data have examined correlates of active neural firing, additional research into activity-silent mechanisms is needed to understand their properties and generalizability. Therefore, we consider activity-silent mechanisms a lower priority benchmark (C). This rating also receives some support from the survey, where the most frequent ratings were B (14/51) and "not a benchmark" (17/51).

Appendix C: Cross-Reference Table

Benchmark (Rating)	Verbal	Visual	Spatial
1.1. Set-size effect on accuracy (A)	FR: Grenfell-Essam and Ward (2012) SR: Crannell and Parrish (1957) **CS**: Unsworth and Engle (2006a) PR: Murdock (1968b) **MU**: Oberauer and Kliegl (2001) **NB**: Jonides et al. (1997) IRec: R. E. Morin et al. (1967)	CD: Luck and Vogel (1997) CR: Wilken and Ma (2004)	SR: Jones et al. (1995); Woods et al. (2016) **CS**: Shah and Miyake (1996) FR: Cortis et al. (2015) **MU**: Oberauer and Kliegl (2006)
1.2. Set-size effect on RT (A)	IRec: Donkin and Nosofsky (2012b); Monsell (1978); Sternberg (1966, 1969) SR: Maybery et al. (2002) **CS**: Towse et al. (2008) FR: Rohrer (1996) **NB**: (Oberauer, 2006); Verhaeghen et al. (2004); Jaeggi, Schmid, Buschkuehl, and Perrig (2008) **MU**: Kessler and Meiran (2008); Oberauer, Wendland, et al. (2003)	CD: Donkin et al. (2013); Gilchrist and Cowan (2014); Pearson, Raškevičius, Bays, Pertzov, and Husain (2014); Souza, Rerko, and Oberauer (2014) **MU**: Kessler and Meiran (2008)	**MU**: Oberauer and Kliegl (2006); **NB**: Jaeggi et al. (2008)
1.3. Number of chunks remembered (B)	SR: Broadbent (1975); SRI: Chen and Cowan (2005, 2009); Cowan et al. (2004); IRec: Cowan et al. (2012); PR: Sperling (1960)	FR: Chase and Simon (1973a); Gobet and Clarkson (2004); Gobet and Simon (1998); CD: Luck and Vogel (1997) CR: (Adam, Vogel, & Awh, 2017)	PR: Cleeremans and McClelland (1991); also, visual references involve the spatial placement of visually distinct items.
2.1. (A) : Effect of filled (verbal, spatial) or unfilled (visual) retention interval	**BP**: J. Brown (1958); Peterson and Peterson (1959) FR: Glanzer, Gianutsos, and Dubin (1969)	CD: Mercer and Duffy (2015); Phillips (1974) CR: Pertzov, Bays, et al. (2013) IRec: (Kerr, Avons, & Ward, 1999); Ricker and Cowan (2010); Sakai and Inui (2002)	**BP**: Kopelman and Stanhope (1997) **BP**: Meudell (1977)
2.2. Effect of RI reduced in absence of PI (B)	**BP**: Baddeley and Scott (1971); Keppel and Underwood (1962)	CD: Shipstead and Engle (2013) CR: Souza and Oberauer (2015) IRec: Mercer (2014)	FR: Meudell (1977)
2.3. No effect of RI when filled	SR: Lewandowsky et al. (2004); Lewandowsky, Geiger, et al. (2008); Phaf and Wolters		

with constant distractor (B)	(1993); Vallar and Baddeley (1982) **CS**: Lewandowsky et al. (2010)		
2.4. Presentation duration effects (B)	SR: Matthews and Henderson (1970); Tan and Ward (2008) FR: Glanzer and Cunitz (1966); Roberts (1972) IRec: Ratcliff and Murdock (1976)	SR: Smyth et al. (2005) CD: Bays et al. (2011); Vogel et al. (2006)	
3.1. Primacy and recency effect on accuracy (A)	SR: Drewnowski and Murdock (1980); Madigan (1971) SR-B: Guérard, Saint-Aubin, Burns, and Chamberland (2012); S. C. Li and Lewandowsky (1993); Madigan (1971) SRI: Drewnowski and Murdock (1980) FR: Murdock (1962) PR: Murdock (1968a) ROO: Ward et al. (2010) **CS**: Unsworth and Engle (2006b) IRec: Monsell (1978); Oberauer (2003b) RRec: Oberauer (2003b)	ROO: Avons (1998); Guérard and Tremblay (2008); Ward, Avons, and Melling (2005) IRec: Ward et al. (2005)	SR: Jones et al. (1995); Smyth and Scholey (1996) SR-B: Farrand and Jones (1996) FR: Cortis et al. (2015)
3.2. Modality and its interaction with recency (B)	SR: Conrad and Hull (1964); Watkins et al. (1974); (Beaman, 2002). SR-B: Madigan (1971) FR:Watkins et al. (1974) SRI: Watkins et al. (1974) PR: Murdock and vom Saal (1967); Murdock (1967)		ROO: (Tremblay et al., 2006)
3.3.1. Serial position effects on recognition latencies (B)	IRec: Corballis (1967); Donkin and Nosofsky (2012b); Forrin and Morin (1969); McElree and Dosher (1989) RRec: Oberauer (2003b)	RRec: Nosofsky and Donkin (2016)	
3.3.2. Fast access to last item (C)	IRec: McElree and Dosher (1989); Öztekin and McElree (2010); Wickelgren et al. (1980)		
3.3.3. Serial position effects on recall latencies (C)	SR: Maybery et al. (2002) PR: Sanders and Willemsen (1978)	ROO: Hurlstone and Hitch (2017)	ROO: Hurlstone and Hitch (2015)
3.4.1. Effects of Output Order on Accuracy (B)	SR: Cowan et al. (2002); (Tan & Ward, 2007) PR: Oberauer (2003b) Tulving and Arbuckle (1966) IRec: Oberauer (2003b) FR: Dalezman (1976)		
3.4.2. Effects of Output Order on	PR: Oberauer (2003b) RRec: Lange et al. (2011); Oberauer (2003b)		

Retrieval Latency (B)	FR: Murdock and Okada (1970)		
3.4.3. Effects of output contiguity (B)	PR: Nairne et al. (2007); RRec: Lange et al. (2011) FR: Kahana (1996) ROO: Lewandowsky et al. (2009)		
3.5.1 Self-chosen start of recall (B)	SR: Grenfell-Essam and Ward (2012); Ward et al. (2010) FR: Grenfell-Essam and Ward (2012); Ward et al. (2010) OR: Ward et al. (2010) ROO: Ward et al. (2010)		FR: Cortis et al. (2015)
3.5.2 Semantic clustering of recall (C)	FR: Bousfield (1953); Golomb, Peelle, Addis, Kahana, and Wingfield (2008) Jenkins and Russell (1952) SR: Golomb et al. (2008)		
4.1. Confusions of target item with other items in memory set (A)	SR: Henson et al. (1996) PR: Fuchs (1969) RRec: Oberauer (2005)	CD: Donkin et al. (2015); Wilken and Ma (2004) CR: Bays et al. (2009) ROO: Smyth et al. (2005)	SR & ROO: Guérard and Tremblay (2008)
4.1.1. Locality constraint on transposition errors (A)	SR: Henson et al. (1996) ROO: Surprenant et al. (2005) PR: Fuchs (1969) **CS**: Oberauer et al. (2012) **NB**: Szmalec et al. (2011)	ROO: Smyth et al. (2005) CD, PR: (Rerko et al., 2014) **NB**: McCabe and Hartman (2008)	ROO: Hurlstone and Hitch (2015)
4.1.2. Fill-in effect in serial recall (C)	SR, **CS**: Farrell et al. (2013) ROO: Surprenant et al. (2005)	SR: Miozzo, Petrov, Fischer-Baum, and Peressotti (2016)	SR & ROO: Guérard and Tremblay (2008)
4.2. Serial position effects on error types in serial recall (C)	SR & ROO: Guérard and Tremblay (2008) **CS**: Oberauer et al. (2012)	ROO: Avons and Mason (1999)	SR & ROO: Guérard and Tremblay (2008)
4.3. Intrusions from previous memory sets (B)	SR: Drewnowski and Murdock (1980); Fischer-Baum and McCloskey (2015) **BP**: Quinlan et al. (2015) IRec: Atkinson et al. (1974); Berman et al. (2009); Jonides et al. (1998)	RRec: Hartshorne (2008)	
4.4. Ranschburg effect in serial recall (C)	SR: Henson (1998a); Jahnke (1969)		
4.5. Error distributions on continuous response scales (B)		CR: Zhang and Luck (2008); Van den Berg et al. (2012); (Adam et al., 2017); Bays (2016)	

5.1.1. Multiple-set effects between domains (A)	IRec: Fougnie and Marois (2006) PR Morey et al. (2013) SR: Allen et al. (2006); Cowan and Morey (2007); Logie et al. (1990); Morey and Cowan (2005); Sanders and Schroots (1969) ROO: Morey and Mall (2012); Morey and Miron (2016)	CD: Cowan and Morey (2007); Cowan et al. (2014); Morey and Cowan (2004, 2005); Morey et al. (2013); van Lamsweerde, Beck, and Elliott (2015) IRec: Allen et al. (2006)	IRec: Depoorter and Vandierendonck (2009) RRec: Depoorter and Vandierendonck (2009); Logie et al. (1990); Vandierendonck (2015). ROO: Morey and Mall (2012), Morey and Miron (2016); Vandierendonck (2015).
5.1.2 Multiple-set effects within domains (B)	SR: Cowan and Morey (2007); Logie et al. (1990); Sanders and Schroots (1969)	CD: Cowan and Morey (2007); Delvenne and Bruyer (2004); Fougnie and Marois (2006); Fougnie and Marois (2009) SR: Smyth and Pendleton (1990)	IRec: Depoorter and Vandierendonck (2009) RRec: Depoorter and Vandierendonck (2009)
5.1.3 Asymmetric effects between verbal and spatial sets (C)	PR: Morey et al. (2013) ROO: Morey and Mall (2012); Morey and Miron (2016)	see column "verbal"	see column "verbal"
5.2.1 Processing impairs memory in same domain (A)	Concurrent articulation: SR: Baddeley et al. (1984) **CS**: Camos et al. (2009) **BP**: Meiser and Klauer (1999) ROO, FR: Camos, Lagner, and Loaiza (2016) Rrec: Lange et al. (2011) Other forms of processing: **CS**: Chein et al. (2011); Shah and Miyake (1996) SR: Hale et al. (1996) **BP**: Jarrold et al. (2010); Logie et al. (1990) ROO, FR: Camos et al. (2016)	**BP**: Klauer and Zhao (2004); Della Sala, Gray, Baddeley, Allamano, and Wilson (1999); Logie and Marchetti (1991); Tresch et al. (1993)	**CS**: Chein et al. (2011); Shah and Miyake (1996) SR: Hale et al. (1996) **BP**: Della Sala et al. (1999); Klauer and Zhao (2004)
5.2.2. Processing impairs memory across domains (A) (references are listed by the domain of the memory task)	**CS**: Chein et al. (2011); Jarrold et al. (2011); Vergauwe et al. (2010); Vergauwe et al. (2012) **BP**: Jarrold et al. (2011) CD: Makovski (2012)	**CS**: Vergauwe et al. (2009) **BP** & IRec: Vergauwe et al. (2009)	**CS**: Vergauwe et al. (2009) **BP** & IRec: Vergauwe et al. (2009)
5.2.3. Processing effect within domain less serious when	**CS**: Conlin et al. (2005); Conlin and Gathercole (2006); Turner and Engle (1989)		

different categories (C)			
5.2.4 Effect of cognitive load (A)	**CS**: Barrouillet et al. (2004); Barrouillet et al. (2007); Barrouillet et al. (2011); Camos et al. (2009); Hudjetz and Oberauer (2007); Liefooghe et al. (2008); Plancher and Barrouillet (2013); Vergauwe et al. (2010) **BP** & SR: Liefooghe et al. (2008) **BP** & LRec: Vergauwe et al. (2015)	**CS**: Vergauwe et al. (2009) **BP** & IRec: Vergauwe et al. (2009)	**CS**: Vergauwe et al. (2009); Vergauwe et al. (2010) **BP** & IRec: Vergauwe et al. (2009) **BP** & RRec: Vergauwe et al. (2015)
5.2.5 Effect of secondary task on items and bindings (B)		CD & **BP**: Allen et al. (2006); Allen et al. (2009); Morey and Bieler (2013); Vergauwe et al. (2014)	
6.1. Irrelevant-sound effect (B)	SR: Colle and Welsh (1976); Miles et al. (1991); Salamé and Baddeley (1982) FR: Beaman and Jones (1998); (Salamé & Baddeley, 1990); ROO: Tremblay, Parmentier, Hodgetts, Hughes, and Jones (2012); Tremblay, Macken, and Jones (2000) IRec: LeCompte (1994); Bell, Röer, and Buchner (2013) PR: Beaman and Jones (1997); LeCompte (1994)		ROO: Jones et al. (1995); Tremblay et al. (2012)
6.2. Changing-state modulation of IS effect (B)	SR: Jones et al. (1992); Meiser and Klauer (1999) ROO: Hughes, Hurlstone, Marsh, Vachon, and Jones (2013); Schlittmeier, Weisz, and Bertrand (2011) SRec: Gisselgård, Uddén, Ingvar, and Peterson (2007)		ROO: Tremblay et al. (2001)
6.3. Auditory deviant effect (C)	SR: Hughes et al. (2007); Hughes et al. (2013); Lange (2005)		
7. Syllable-based word-length effect (B)	SR: Baddeley et al. (1975); Mackworth (1963) SR-SPAN: Baddeley, Chincotta, Stafford, and Turk (2002); LaPointe and Engle (1990) SR-B: Cowan et al. (1992); Guérard et al. (2012) FR: Bhatarah et al. (2009); Watkins (1972) PR: Avons et al. (1994)	N.A.	N.A.

	ROO: Nairne et al. (1997); Tehan and Tolan (2005) **CS**: LaPointe and Engle (1990); Tehan, Hendry, and Kocinski (2001) Serial recognition: Baddeley et al. (2002)		
8.1.1 Phonological similarity effect (A)	SR: Conrad and Hull (1964); Farrell and Lewandowsky (2003); Wickelgren (1965) FR: Sperling and Speelman (1970) PR: Sperling and Speelman (1970) SRec: Nimmo and Roodenrys (2005) **MU**: Oberauer and Kliegl (2006) **CS**, often not observed: Chow et al. (2016); Macnamara et al. (2011)	N.A.	N.A.
8.1.2. Mixed list effect of phonological or visual similarity (B)	SR: Henson et al. (1996); Baddeley (1968); Farrell and Lewandowsky (2003)	SR: Logie, Saito, Morita, Varma, and Norris (2016)	
8.1.3. Phonological similarity interacts with articulatory suppression (B)	SR: Larsen and Baddeley (2003); D. J. Murray (1968)	N.A.	N.A.
8.1.4. Development of phonological and visual similarity effects (C)	SR: Hayes and Schulze (1977); Hitch, Woodin, et al. (1989)		
8.1.5. Effects of visual similarity on serial recall (C)	SR: Logie et al. (2000); Saito et al. (2008)	ROO: Avons and Mason (1999)	
8.2. Effect of size of change on recognition and change detection (C)		IRec: Kahana, Zhou, Geller, and Sekuler (2007) CD: Van den Berg et al. (2012)	
9.1.1 Temporal isolation effects in most list-recall paradigms: isolated items are recalled better (B).	FR: G. D. A. Brown, Morin, and Lewandowsky (2006) Running memory span: Geiger and Lewandowsky (2008) IRec: C. Morin et al. (2010) ROO: Lewandowsky et al. (2006)		

9.1.2. No temporal isolation effects in forward serial recall and serial recognition (C)	SR: Nimmo and Lewandowsky (2005); Nimmo and Lewandowsky (2006) SRec: Farrell and McLaughlin (2007)	SR: Peteranderl and Oberauer (2017)	SR: Parmentier, King, et al. (2006)
9.1.3 Non-temporal isolation effects (B)	SR: Smith & Stearns (1949) (anticipation method) IRec: von Restorff (1933)/ Hunt (1995) PR: Calkins (1894) FR: Bireta et al. (2008); Welch and Burnett (1924) ROO: Lippman (1980); Lippman and Lippman (1978)	Experiments with verbal material usually involve manipulations of visual distinctiveness	SR: Guérard, Hughes, & Tremblay (2008)
9.2.1. Grouping enhances recall relative to ungrouped lists (A)	SR: Wickelgren (1964) PR: Farrell, Wise, and Lelievre (2011) FR: Sometimes for recency items; negative effect for early-list items: Gianutsos (1972); Tzeng and Hung (1973)	ROO: Hurlstone and Hitch (2017)	SR: Parmentier et al. (2004)
9.2.2. Primacy and recency within groups (B)	SR: Ryan (1969); C. R. Frankish (1995) PR: G. D. A. Brown, Vousden, and McCormack (2009)	ROO: Hurlstone and Hitch (2017)	SR: Parmentier, Andrés, Elford, and Jones (2006)
9.2.3 Tendency to confuse items in same within-group positions (interpositions) (B)	SR: Farrell and Lelièvre (2009); Henson (1999); Wickelgren (1964)		
9.2.4. Effect on recall latency: longer recall time for first item in group (B)	SR, temporally-induced grouping: (Maybery et al., 2002) SR, spatially-induced grouping: J. R. Anderson and Matessa (1997) PR: Farrell, Wise, and Lelievre (2011)	ROO: Hurlstone and Hitch (2017)	SR (temporally-induced grouping: Parmentier et al. (2004) SR (spatially and voice-induced grouping): Parmentier and Maybery (2008):
10.1. Retro-cue effects: item cues (B)		CD: Landman et al. (2003) RRec: Griffin and Nobre (2003); Makovski, Sussman, and Jiang (2008) IRec: Lepsien, Griffin, Devlin, and Nobre (2005) CR: Pertzov, Bays, et al. (2013); Souza, Rerko, Lin, et al. (2014)	RRec: Q. Li and Saiki (2015)

10.2. Item-switch costs (B)	**MU**: Garavan (1998); Oberauer (2003a) **NB**: Oberauer (2006)		**MU**: Hedge and Leonards (2013); Hedge, Oberauer, and Leonards (2015); Kübler et al. (2003)
11.1. Chunking benefit (A)	OR: Baddeley (1964); Bower and Springston (1970) SRI: G. A. Miller and Selfridge (1950) SPAN: G. A. Miller (1956) Ericsson et al. (1980) FR: Chen and Cowan (2005) SR: Chen and Cowan (2005, 2009); Mathy and Feldman (2012) **CS**: (Portrat et al., 2016)	PR: Brady et al. (2009) CD: Gao, Gao, Tang, Shui, and Shen (2015)	ROL: (Chase & Simon, 1973b) Gong et al. (2015)
11.2. Sentence-superiority effect (C)	SPAN: Brener (1940) SR: Baddeley et al. (2009)	N.A.	N.A.
11.3. Lexicality, frequency, and phonotactic effects (B)	SPAN: Gregg et al. (1989); Hulme et al. (1991); Hulme et al. (1997); Roodenrys et al. (1994) SR: Gathercole et al. (1999); Gregg et al. (1989); Hulme et al. (1997); Majerus and Van der Linden (2003); Thorn and Frankish (2005); Thorn et al. (2005); Treiman and Danis (1988); Woodward et al. (2008) **CS**: Conlin and Gathercole (2006)	CD, familiarity: Buttle and Raymond (2003); Jackson and Raymond (2008)	
11.4. Regularization (C)	SR: Botvinick and Bylsma (2005); Jefferies et al. (2006)		
11.5. Hebb effect (A)	SR: Bower and Winzenz (1969); Hebb (1961); Hitch et al. (2005); Hitch et al. (2009) ROO: Szmalec et al. (2009) **CS**: Oberauer, Jones, and Lewandowsky (2015)	ROO: Horton et al. (2008); Page et al. (2006)	ROO: Couture and Tremblay (2006); Turcotte et al. (2005)
12.1. Positive intercorrelation of WM tasks (A)	**CS**, SR: Engle et al. (1999) **CS**, SR: Kane et al. (2004) **MU**, ROO, SR: Kyllonen and Christal (1990) **CS**, **MU**, SR: Oberauer, Süß, et al. (2003) **CS**: Unsworth et al. (2014)	CD, CR: Chow and Conway (2015); Unsworth et al. (2014)	**CS**, SR: (Engle et al., 1999) **CS**, SR: Kane et al. (2004) **CS**, **MU**, SR: Oberauer, Süß, et al. (2003) **CS**: Unsworth et al. (2014)
12.2. Correlations higher within than between domains (B)	**CS**, SR: Bayliss et al. (2003) **CS**, SR: Kane et al. (2004) **CS**: Shah and Miyake (1996)		**CS**, SR: Bayliss et al. (2003) **CS**, SR: Kane et al. (2004) **CS**: Shah and Miyake (1996)

12.3. Correlations between domains larger for complex span (B)	**CS**, SR, SR-B: Alloway et al. (2006) **CS**, SR: Bayliss et al. (2003) **CS**, SR: Kane et al. (2004)		**CS**, SR: Alloway et al. (2006) **CS**, SR: Bayliss et al. (2003) **CS**, SR: Kane et al. (2004)
12.4. Separate factors for primary and secondary memory (B)	**CS**, FR: Unsworth and Engle (2007a); Unsworth et al. (2010)		**CS**: Unsworth et al. (2010)
12.5. Correlation of WM with articulation and retrieval speed (B)	SR: Cowan (1992) SR: Cowan et al. (1994) SR: Nicolson (1981)		
12.6. Correlation of WM with measures of attention (A)	**CS**, Stroop: Kane and Engle (2003) **CS**, anti-saccade: Chuderski (2014); Kane et al. (2001); Kane et al. (2016) **NB**, anti-saccade: Chuderski (2015) **CS**, mind-wandering: McVay and Kane (2009, 2012)	CD, Stroop & anti-saccade: Chuderski et al. (2012) CD, anti-saccade: Chuderski (2015) CD, mind-wandering: Unsworth and Robison (2016)	**CS**, anti-saccade: Chuderski (2014); McVay and Kane (2012) **NB**, anti-saccade: Chuderski (2015)
12.7. Correlation of WM with fluid intelligence (A)	ROO, SR, SR-B: Gignac (2014) **CS**, **MU**, ROO, SR-B: Kane et al. (2005) **CS**, **MU**, ROO, SR-B: Oberauer et al. (2005)	CD, CR: Chow and Conway (2015); Unsworth et al. (2014)	**CS**, **MU**, ROO, SR-B: Kane et al. (2005) **CS**, **MU**, ROO, SR-B: Oberauer et al. (2005)
13.1. Separate neural networks for different content domains (A)	IRec: Awh et al. (1996); Gruber and von Cramon (2003); Paulesu et al. (1993); E. E. Smith et al. (1996) SR: Chein and Fiez (2001) FR: Fiez et al. (1996) **NB**: Cohen et al. (1997) Lesion meta-analysis: D'Esposito and Postle (1999) Voxel-based meta-analysis: Nee et al. (2013); Owen et al. (2005); Rottschy et al. (2012)	IRec: Courtney et al. (1998); Courtney et al. (1996); Mohr, Goebel, and Linden (2006); Mottaghy et al. (2002); Sala and Courtney (2007); Sala, Rama, and Courtney (2003) Voxel-based meta-analysis: Nee et al. (2013); Owen et al. (2005); Rottschy et al. (2012)	IRec: Courtney et al. (1998); Courtney et al. (1996); Gruber and von Cramon (2003); Mohr et al. (2006); Mottaghy et al. (2002); Rama et al. (2004); Sala and Courtney (2007); Sala et al. (2003); E. E. Smith et al. (1996) Lesion meta-analysis: D'Esposito and Postle (1999) Voxel-based meta-analysis: Nee et al. (2013); Owen et al. (2005); Rottschy et al. (2012)
13.2. Preserved short-term memory in amnesia (A)	SR: Baddeley and Warrington (1970) FR: Baddeley and Warrington (1970); Carlesimo, Marfia, Loasses, and Caltagirone (1996)	IRec: Baddeley, Allen, and Vargha-Khadem (2010)	CR: Warrington and Baddeley (1974) ROO: Cave and Squire (1992);

	BP: Baddeley and Warrington (1970)		Jeneson, Mauldin, and Squire (2010)
13.3. Parietal BOLD and CDA track set size up to about 3 or 4 items (A)	IRec: (Cowan et al., 2011); Manoach et al. (1997); Veltman et al. (2003) **NB**: Cohen et al. (1997); Veltman et al. (2003)	CD: Todd and Marois (2004, 2005); Vogel and Machizawa (2004); Xu and Chun (2006)	
13.4. Activation-based and non-activation-based neural signatures (C)	PC: LaRocque et al. (2013); Lewis-Peacock et al. (2011)	PC: Wolff, Ding, Myers, and Stokes (2015); LaRocque et al. (2013); Lewis-Peacock et al. (2011)	CR: Sprague, Ester, and Serences (2016)

Acronyms (paradigms involving processing in addition to maintenance are printed in bold):

SR – serial recall (SR-B for backward serial recall, SPAN for span measure from serial recall)

SRI – serial recall, item-memory scoring (ignoring order)

FR – free recall

PR – probed recall

OR – ordered recall (recall the items in any order, placing them in the correct ordinal list position)

ROO – reconstruction of order (assign a given set of items to their correct ordinal list position)

ROL - reconstruction of (spatial) location (assign a given set of items to their correct spatial location)

CD – change detection/change discrimination

CR – continuous reproduction (a.k.a. delayed estimation)

IRec – Item recognition

RRec – relational recognition (e.g., recognizing an item in a spatial location, or a conjunction of two words or two visual features).

SRec – serial-order recognition (i.e., deciding whether the order of a test list matches that of a memory list)

PC – Probe comparison (comparing a probe to an item on a given feature dimension)

CS – complex span

BP – Brown-Peterson

NB – N-back

MU – Working memory updating

Appendix D: Reference Table for Benchmarks in Children and Old Adults

Benchmark (Rating)	Children (age range in years)	Old Adults
1.1. Set-size effect on accuracy (A)	IR, **CS**, CD (9, 11): Cowan et al. (2005) CD (5, 7, 10): Riggs, McTaggart, Simpson, and Freeman (2006) **CS**: Camos and Barrouillet (2011) NB (7-13): Pelegrina et al. (2015)	ROO: Burke et al. (2015) CD: Duarte et al. (2013) **MU**: Oberauer and Kliegl (2001) NB: Verhaeghen and Basak (2005)
1.2. Set-size effect on RT (A)	IRec (9, 13): Hoving, Morin, and Konick (1970); Keating and Bobbitt (1978) IRec (5, 9, 11): Spitzer (1976)	IRec: Ferraro and Balota (1999); Lange and Verhaeghen (2009) **MU**: Oberauer, Wendland, et al. (2003) NB: Jaeggi et al. (2008)
1.3. Number of chunks remembered (B)		
2.1.: Effect of filled retention interval on verbal and spatial memory (A)	**BP** (8-17): Towse, Hitch, and Hutton (2002)	**BP**: Floden et al. (2000); Puckett and Lawson (1989)
2.2. Effect of RI reduced in absence of PI (B)	**BP** (9-13): Kail (2002)	**BP**: Puckett and Lawson (1989)
2.3. No effect of RI when filled with constant distractor (B)		
2.4. Presentation duration effects (B)	SR, presentation duration effect with 10 but not 7 years: D. J. Murray and Roberts (1968)	CD: Sander, Werkle-Bergner, and Lindenberger (2011a) FR, SR: Golomb et al. (2008)
3.1. Primacy and recency effect on accuracy (A)	IRec (0.5): Cornell and Bergstrom (1983) IRec (5, 9, 11): Spitzer (1976) SR (7, 9, 11): McCormack, Brown, Vousden, and Henson (2000); (5, 8): Pickering, Gathercole, and Peaker (1998) ROO (5-12): Koppenol-Gonzalez, Bouwmeester, and Vermunt (2014); (Morey, Mareva, Lelonkiewicz, & Chevalier, 2017) FR (7, 9, 14): Recency independent of age, primacy only for age 9, 14: Cole, Frankel, and Sharp (1971); Jarrold et al. (2015)	SR, SR-B: Elliott et al. (2011); Maylor, Vousden, and Brown (1999) ROO: Korsnes and Magnusson (1996) FR: Golomb et al. (2008)
3.2. Modality and its interaction with recency (B)	SR (7-10): D. J. Murray and Roberts (1968) FR (9+): Dempster and Rowher (1983)	SR: Manning and Geenhut-Wertz (1990); (Maylor et al., 1999)

3.3.1. Serial position effects on recognition latencies (B)	IRec (5, 9, 11): Spitzer (1976)	IRec: Lange and Verhaeghen (2009)
3.3.2. Fast access to last item (C)		IRec: Öztekin, Güngör, and Badre (2012)
3.3.3. Serial position effects on recall latencies (C)	SR (4), no serial position effect: (Cowan, 1992)	ROO: Korsnes and Magnusson (1996)
3.4.1. Effects of Output Order on Accuracy (B)		
3.4.2. Effects of Output Order on Retrieval Latency (B)		
3.4.3. Effects of output contiguity (B)	FR (5-8): Jarrold et al. (2015)	FR, reduced contiguity effect: Golomb et al. (2008)
3.5.1 Self-chosen start of recall (B)		
3.5.2 Semantic clustering of recall (C)	FR (9-12): Cole et al. (1971) FR; clustering tendency increasing with age (4,6,10): Wingard (1980)	FR, reduced but still present in old age: Denney (1974); Taconnat et al. (2009); Wingard (1980)
4.1. Confusions of target item with other items in memory set (A)	SR (7, 9, 11): McCormack et al. (2000)	SR: Maylor et al. (1999) RRec: Oberauer (2005) CR: Peich, Husain, and Bays (2013)
4.1.1. Locality constraint on transposition errors (A)	SR (7, 9, 11): McCormack et al. (2000); (5, 8): Pickering et al. (1998)	SR: Maylor et al. (1999) NB: McCabe and Hartman (2008)
4.1.2. Fill-in effect in serial recall (C)		
4.2. Serial position effects on error types in serial recall (C)	SR (7, 9, 11): McCormack et al. (2000)	SR: Maylor et al. (1999)
4.3. Intrusions from previous memory sets (B)	IRec (8-10): Loosli, Rahm, Unterrainer, Weiller, and Kaller (2014)	IRec: Loosli et al. (2014) **CS**: Zeintl and Kliegel (2010)
4.4. Ranschburg effect in serial recall (C)	SR (7, 9, 11): McCormack et al. (2000)	SR: Maylor and Henson (2000)
4.5. Error distributions on continuous response scales (B)	CR (7-12): Sarigiannidis, Crickmore, and Astle (2016)	CR: Peich et al. (2013)
5.1.1. Multiple-set effects between domains (A)		
5.1.2 Multiple-set effects within domains (B)		
5.1.3 Asymmetric effects between verbal and spatial sets (C)		
5.2.1 Processing impairs memory in same domain (A)	SR (5, 8): Rattat (2010) **CS** (8, 10): Hale, Bronik, and Fry (1997) **BP** (6, 8): S. Miller, McCulloch, and Jarrold (2015); Tam, Jarrold, Baddeley, and Sabatos-DeVito (2010)	**CS**: Myerson, Hale, Rhee, and Jenkins (1999) **BP**: (Puckett & Stockburger, 1988)
5.2.2. Processing impairs memory across domains (A)	SR (7-9): M. Anderson, Bucks, Bayliss, and Della Sala (2011); Bayliss et al. (2003)	SR: M. Anderson et al. (2011) **BP**: Puckett and Stockburger (1988)

5.2.3. Processing effect within domain less serious when different categories (C)	**CS** (9-10): Conlin and Gathercole (2006); Conlin et al. (2005)	**CS**: K. Z. H. Li (1999)
5.2.4 Effect of cognitive load (A)	**CS**, effect of cognitive load at age 7 and older but not at age 5: Barrouillet, Gavens, Vergauwe, Gaillard, and Camos (2009); Camos and Barrouillet (2011); Portrat, Camos, and Barrouillet (2009)	**CS**: Baumans, Adam, and Seron (2012)
5.2.5 Effect of secondary task on items and bindings (B)		
6.1. Irrelevant-sound effect (B)	SR (7+): Elliott (2002); Elliott et al. (2016); Klatte, Lachmann, Schlittmeier, and Hellbruck (2010)	SR: Rouleau and Belleville (1996)
6.2. Changing-state modulation of IS effect (B)	SR (7+): Elliott (2002); Elliott et al. (2016)	SR: Röer, Bell, Marsh, and Buchner (2015)
6.3. Auditory deviant effect (C)		SR: Röer et al. (2015)
7. Syllable-based word-length effect (B)	SR, spoken words (4+): (Cowan et al., 1994); Hitch, Halliday, Dodd, and Littler (1989); Hulme, Thomson, Muir, and Lawrence (1984); Hulme and Tordoff (1989) PR, WLE at 7 but not 4 years: Henry (2007)	SR: Bireta, Fine, and VanWormer (2013); Collette, Van der Linden, Bechet, and Salmond (1999)
8.1.1 Phonological similarity effect in recall (A)	SR, no effect with 3 years, effect potentially with 4-5 years, certainly with 7 years and older: Conrad (1971); Henry (1991); Hulme and Tordoff (1989) SRec (5-9): Jarrold, Cocksey, and Dockerill (2008) **BP** (5-8): Tam et al. (2010)	SR: Bireta et al. (2013); Collette et al. (1999)
8.1.2. Mixed list effect of phonological or visual similarity (B)		
8.1.3. Phonological similarity interacts with articulatory suppression (B)	SR (7, 10): Halliday, Hitch, Lennon, and Pettipher (1990); Hasselhorn and Grube (2003)	SR: Bireta et al. (2013)
8.1.4. Development of phonological and visual similarity effects (C)	N.A.	N.A.
8.1.5. Effects of visual similarity on serial recall (C)	SR, visual-similarity effect in 5 but not 10 year olds: (Hitch, Halliday, Schaafstal, & Schraagen, 1988)	
8.2. Effect of size of change on recognition and change detection (C)		
9.1.1 Temporal isolation effects in most list-recall		

paradigms: isolated items are recalled better (B).		
9.1.2. No temporal isolation effects in forward serial recall and serial recognition (C)		
9.1.3 Non-temporal isolation effects (B)		FR: Bireta et al. (2008); Vitali et al. (2006)
9.2.1. Grouping enhances recall relative to ungrouped lists (A)	SR, Grouping benefit at age 8 and older but not younger: Harris and Burke (1972); Towse, Hitch, and Skeates (1999)	
9.2.2. Primacy and recency within groups (B)	SR: Harris and Burke (1972)	
9.2.3 Tendency to confuse items in same within-group positions (interpositions) (B)		
9.2.4. Effect on recall latency: longer recall time for first item in group (B)		
10.1. Retro-cue effects: item cues (B)	CD (7, 10): Astle, Nobre, and Scerif (2012); Shimi and Scerif (2017)	CR: Souza (2016) CD, no retro-cue effect in old adults: Duarte et al. (2013); Newsome et al. (2015)
10.2. Item-switch costs (B)	**MU** (7-11): Magimairaj and Montgomery (2012)	**MU**: Oberauer, Wendland, et al. (2003); Verhaeghen and Basak (2005)
11.1. Chunking benefit (A)	SR (8, 10): Cowan et al. (2010) SR (10, 12, 14): Burtis (1982)	SR: Naveh-Benjamin, Cowan, Kilb, and Chen (2007)
11.2. Sentence-superiority effect (C)		
11.3. Lexicality, frequency, and phonotactic effects (B)	PR, SRec: Jarrold et al. (2008)	
11.4. Regularization (C)		
11.5. Hebb effect (A)	SR (5-8): Majerus et al. (2009); Mosse and Jarrold (2008)	SR, Hebb effect for verbal but not spatial materials: Turcotte et al. (2005)
12.1. Positive intercorrelation of WM tasks (A)	**CS**, SR, SR-B, SPAN (4-15): Alloway et al. (2006); Bayliss et al. (2003); Gathercole, Pickering, Ambridge, and Wearing (2004)	**CS**, SR: Hale et al. (2011); Park et al. (2002)
12.2. Correlations higher within than between domains (B)	**CS**, SR, SR-B, SPAN (4-11): Alloway et al. (2006); Gathercole et al. (2004)	**CS**, SR: Hale et al. (2011); Park et al. (2002)
12.3. Correlations between domains larger for complex span (B)	**CS**, SR (4-11): Alloway et al. (2006)	
12.4. Separate factors for primary and secondary memory (B)		

12.5. Correlation of WM with articulation and retrieval speed (B)	SR (7-11): Cowan et al. (1998); (1989)	SR: Correlation with articulation speed Collette et al. (1999)
12.6. Correlation of WM with measures of attention (B)		**CS**, Stroop: Aschenbrenner and Balota (2015)
12.7. Correlation of WM with fluid intelligence (A)	SR, SR-B, **CS**, ROO (4): Gustafsson and Wolff (2015) SR, SR-B, **CS** (8, 12): Engel de Abreu, Conway, and Gathercole (2010); Shahabi, Abad, and Colom (2014)	**CS**: Hertzog, Dixon, Hultsch, and MacDonald (2003)
13.1. Separate neural networks for different content domains (A)		
13.2. Preserved short-term memory in amnesia (A)		
13.3. Parietal BOLD and CDA track set size up to about 3 or 4 items (A)	IRec: Klingberg, Forssberg, and Westerberg (2002)	CD, CDA: Jost, Bryck, Vogel, and Mayr (2010); Sander, Werkle-Bergner, and Lindenberger (2011b)
13.4. Activation-based and non-activation-based neural signatures (C)		

Acronyms (paradigms involving processing in addition to maintenance are printed in bold):SR – serial recall (SR-B for backward serial recall, SPAN for span measure from serial recall)

SRI – serial recall, item-memory scoring (ignoring order)

FR – free recall

PR – probed recall

OR – ordered recall (recall the items in any order, placing them in the correct ordinal list position)

ROO – reconstruction of order (assign a given set of items to their correct ordinal list position)

ROL - reconstruction of (spatial) location (assign a given set of items to their correct spatial location)

CD – change detection/change discrimination

CR – continuous reproduction (a.k.a. delayed estimation)

IRec – Item recognition

RRec – relational recognition (e.g., recognizing an item in a spatial location, or a conjunction of two words or two visual features).

SRec – serial-order recognition (i.e., deciding whether the order of a test list matches that of a memory list)

PC – Probe comparison (comparing a probe to an item on a given feature dimension)

CS – complex span

BP – Brown-Peterson

NB – N-back

MU – Working memory updating

References

Aaronson, D. (1968). Temporal course of perception in an immediate recall task. *Journal of Experimental Psychology, 76*(1), 129-140.

Adam, K. C. S., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The contribution of attentional lapses to individual differences in visual working memory capacity. *Journal of Cognitive Neuroscience, 27*, 1601-1616. doi:10.1162/jocn_a_00811

Adam, K. C. S., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology, 97*, 79-97. doi:10.1016/j.cogpsych.2017.07.001

Allen, R. J., Baddeley, A. D., & Hitch, G. J. (2006). Is the binding of visual features in working memory resource-demanding? *Journal of Experimental Psychology: General, 135*, 298-313.

Allen, R. J., Hitch, G. J., & Baddeley, A. D. (2009). Cross-modal binding and working memory. *Visual Cognition, 17*, 83-102. doi:http://doi.org/10.1080/13506280802281386

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuo-spatial short-term and working memory in children: Are they separable? *Child Development, 77*, 1698-1716.

Anderson, J. R., & Matessa, M. (1997). A production system model of serial memory. *Psychological Review, 104*, 728-748.

Anderson, M., Bucks, M. S., Bayliss, D. M., & Della Sala, S. (2011). Effect of age on dual-task performance in children and adults. *Memory & Cognition, 39*, 1241-1252.

Aschenbrenner, A. J., & Balota, D. A. (2015). Interactive effects of working memory and trial history on Stroop interference in cognitively healthy aging. *Psychology and Aging, 30*, 1-8. doi:10.1037/pag0000012

Astle, D. E., Nobre, A. C., & Scerif, G. (2012). Attentional control constrains visual short-term memory: insights from developmental and individual differences. *Quarterly Journal of Experimental Psychology, 65*, 277-294.

Astle, D. E., Summerfield, J., Griffin, I. C., & Nobre, A. (2012). Orienting attention to locations in mental representations. *Attention, Perception, & Psychophysics, 74*, 146-162.

Atkinson, R. C., Herrmann, D. J., & Wescourt, K. T. (1974). Search processes in recognition memory. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 101-146). Potomac: Erlbaum.

Avons, S. E. (1998). Serial report and item recognition of novel visual patterns. *British Journal of Psychology, 89*, 285-308.

Avons, S. E., & Mason, A. (1999). Effects of visual similarity on serial report and item recognition. *Quarterly Journal of Experimental Psychology, 52A*, 217-240.

Avons, S. E., Wright, K. L., & Pammer, K. (1994). The word-length effect in probed and serial recall. *Quarterly Journal of Experimental Psychology, 74A*, 1047-1054.

Awh, E., Jonides, J., Smith, E. E., Schumacher, E. H., Koeppe, R. A., & Katz, S. (1996). Dissociation of storage and rehearsal in verbal working memory: Evidence from positron emission tomography. *Psychological Science, 7*(1), 25-31. doi:DOI 10.1111/j.1467-9280.1996.tb00662.x

Baddeley, A. D. (1964). Immediate memory and the “perception” of letter sequences. *Quarterly Journal of Experimental Psychology, 16*(364-367).

Baddeley, A. D. (1968). How does acoustic similarity influence short-term memory? *Quarterly Journal of Experimental Psychology, 20*, 249-263.

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.

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

Baddeley, A. D. (2000b). The phonological loop and the irrelevant speech effect: Some comments on Neath (2000). *Psychonomic Bulletin & Review, 7*, 544-549.

Baddeley, A. D. (2012). Working memory: theories, models, controversies. *Annual Review of Psychology, 63*, 1-23.

Baddeley, A. D., Allen, R., & Vargha-Khadem, F. (2010). Is the hippocampus necessary for visual and verbal binding in working memory? *Neuropsychologia, 48*, 1089-1095.

Baddeley, A. D., Chincotta, D., Stafford, L., & Turk, D. (2002). Is the word length effect in STM entirely attributable to output delay? Evidence from serial recognition. *Quarterly Journal of Experimental Psychology, 55*, 353-369.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *Recent Advances in Learning and Motivation* (Vol. VIII, pp. 47-90). New York: Academic Press.

Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory & Language, 61*, 438-456.

Baddeley, A. D., Lewis, V. J., & Vallar, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology, 36*, 233-252.

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple component model. In A. Miyake & P. Shah (Eds.), *Models of working memory* (pp. 28-61). Cambridge: Cambridge University Press.

Baddeley, A. D., & Scott, D. (1971). Short term forgetting in the absence of proactive interference. *Quarterly Journal of Experimental Psychology, 23*, 275-283.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short term memory. *Journal of Verbal Learning and Verbal Behavior, 14*, 575-589.

Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long- and short-term memory. *Journal of Verbal Learning and Verbal Behavior, 9*(2), 176-189. doi:Doi 10.1016/S0022-5371(70)80048-2

Bakeman, R., & McArthur, D. (1996). Picturing repeated measures: Comments on Loftus, Morrison, and others. *Behavioral Research Methods, Instruments, & Computers, 28*, 584-589.

Balaban, H., & Luria, R. (2015). The number of objects determines visual working memory capacity allocation for complex items. *NeuroImage, 119*, 54-62. doi:10.1016/j.neuroimage.2015.06.051

Barak, O., & Tsodyks, M. (2007). Persistent activity in neural networks with dynamic synapses. *PLoS Computational Biology, 3*(2), e35. doi:10.1371/journal.pcbi.0030035

Barak, O., Tsodyks, M., & Romo, R. (2010). Neuronal population coding of parametric working memory. *The Journal of neuroscience : the official journal of the Society for Neuroscience, 30*(28), 9424-9430. doi:10.1523/JNEUROSCI.1875-10.2010

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General, 133*, 83-100. doi:http://doi.org/10.1037/0096-3445.133.1.83

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 33*, 570-585. doi:http://doi.org/10.1037/0278-7393.33.3.570

Barrouillet, P., & Camos, V. (2015). Working memory: Loss and reconstruction. *Psychology Press (Hove, UK)*.

Barrouillet, P., Gavens, N., Vergauwe, E., Gaillard, V., & Camos, V. (2009). Working memory span development: A time-based resource-sharing model account. *Developmental psychology, 45*, 477-490.

Barrouillet, P., & Lecas, J. F. (1999). Mental models in conditional reasoning and working memory. *Thinking & Reasoning, 5*, 289-302.

Barrouillet, P., Portrat, S., & Camos, V. (2011). On the law relating processing to storage in working memory. *Psychological Review, 118*, 175-192. doi:http://doi.org/10.1037/a0022324

Bartlett, F. C. (1932). Remembering: A Study in Experimental and Social Psychology. *Cambridge: Cambridge University Press*.

Baumans, C., Adam, S., & Seron, X. (2012). Effect of cognitive load on working memory forgetting in aging. *Experimental Psychology, 59*, 311-321. doi:10.1027/1618-3169/a000158

Bayliss, D. M., Jarrold, C., Gunn, D. M., & Baddeley, A. D. (2003). The complexities of complex span: Explaining individual differences in working memory in children and adults. *Journal of Experimental Psychology: General, 132*, 71-92.

Bays, P. M. (2016). Evaluating and excluding swap errors in analogue tests of working memory. *Scientific Reports, 6*. Retrieved from doi:10.1038/srep19203

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9*, 1-11.
Bays, P. M., Gorgoraptis, N., Wee, N., Marshall, L., & Husain, M. (2011). Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision, 11*, 1-15.
Beaman, C. P. (2002). Inverting the modality effect in serial recall. *Quarterly Journal of Experimental Psychology, 55A*, 371-389.
Beaman, C. P., & Jones, D. M. (1997). The role of serial order in the irrelevant speech effect: Tests of the changing-state hypothesis. *Journal of Experimental Psychology: Learning, Memory & Cognition, 23*, 459-471.
Beaman, C. P., & Jones, D. M. (1998). Irrelevant sound disrupts order information in free as in serial recall. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 51A*(615-636).
Bell, R., Mund, I., & Buchner, A. (2011). Disruption of short-term memory by distractor speech: Does content matter? *The Quarterly Journal of Experimental Psychology, 64*, 146-168.
Bell, R., Röer, J. P., & Buchner, A. (2013). Irrelevant speech disrupts item-context binding. *Experimental Psychology, 60*, 376-384.
Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In search of decay in verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 317-333.
Bhatarah, P., Ward, G., Smith, J., & Hayes, L. (2009). Examining the relationship between free recall and immediate serial recall: Similar patterns of rehearsal and similar effects of word length, presentation rate, and articulatory suppression. *Memory & Cognition, 37*, 689-713.
Bhatarah, P., Ward, G., & Tan, L. (2008). Examining the relationship between free recall and immediate serial recall: The serial nature of recall and the effect of test expectancy. *Memory & Cognition, 36*, 20-34.
Bireta, T. J., Fine, H. C., & VanWormer, L. A. (2013). Does working memory change with age? The interactions of concurrent articulation with the effects of word length and acoustic confusion. *Aging, Neuropsychology, and Cognition, 20*, 125-147. doi:10.1080/13825585.2012.657607
Bireta, T. J., Surprenant, A. M., & Neath, I. (2008). Age-related differences in the von Restorff isolation effect. *Quarterly Journal of Experimental Psychology, 61*, 345-352. doi:10.1080/17470210701626608
Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology, 6*, 173-189.
Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*, 624-652.
Botvinick, M. M., & Bylsma, L. M. (2005). Regularization in short-term memory for serial order. *Journal of Experimental Psychology: Learning, Memory & Cognition, 31*, 351-358.
Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review, 113*, 201-233.
Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology, 49*, 229-240.
Bower, G. H., & Springston, F. (1970). Pauses as recoding points in letter series. *Journal of Experimental Psychology, 83*, 421-430.
Bower, G. H., & Winzenz, D. (1969). Group structure, coding, and memory for digit series. *Journal of Experimental Psychology Monograph, 80*, 1-17.
Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General, 138*, 487-502.
Brener, R. (1940). An experimental investigation of memory span. *Journal of Experimental Psychology, 26*, 467-482.
Broadbent, D. E. (1975). The magic number seven after fifteen years. In A. Kennedy & A. Wilkes (Eds.), *Studies in long term memory* (pp. 3-18). Oxford, England: John Wiley & Sons.

Brown, G. D. A. (2015). Temporal distinctiveness models of memory. In J. D. Wright (Ed.). *International Encyclopedia of the Social & Behavioral Sciences, 24*, 199-205.

Brown, G. D. A., Morin, C., & Lewandowsky, S. (2006). Evidence for time-based models of free recall. *Psychonomic Bulletin & Review, 13*(4), 717-723.

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review, 107*, 127-181.

Brown, G. D. A., Vousden, J. I., & McCormack, T. (2009). Memory retrieval as temporal discrimination. *Journal of Memory and Language, 60*(1), 194-208. doi:10.1016/j.jml.2008.09.003

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology, 10*, 12-21.

Brown, J. W., Reynolds, J. R., & Braver, T. S. (2007). A computational model of fractionated conflict-control mechanisms in task-switching. *Cognitive Psychology, 55*, 37-85.

Brown, L. A., & Brockmole, J. B. (2010). The role of attention in binding visual features in working memory: Evidence from cognitive ageing. *Quarterly Journal of Experimental Psychology, 63*, 2067-2079. doi:10.1080/17470211003721675

Bunting, M. F., Cowan, N., & Saults, J. S. (2006). How does running memory span work? *Quarterly Journal of Experimental Psychology, 59*, 1691-1700.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review, 106*, 551-581.

Burgess, N., & Hitch, G. J. (2006). A revised model of short-term memory and long-term learning of verbal sequences. *Journal of Memory and Language, 55*, 627-652.

Burke, M. R., Poyser, C., & Schiessl, I. (2015). Age-related deficits in visuospatial memory are due to changes in preparatory set and eye-hand coordination. *Journals of Gerontology: Psychological Sciences, 70*, 682-690. doi:10.1093/geronb/gbu027

Burtis, P. J. (1982). Capacity increase and chunking in the development of short-term memory. *Journal of Experimental Child Psychology, 34*, 387-413.

Buttle, H., & Raymond, J. E. (2003). High familiarity enhances visual change detection for face stimuli. *Perception & Psychophysics, 65*, 1296-1306.

Calkins, M. W. (1894). Association. *Psychological Review, 1*, 476-483.

Camos, V., & Barrouillet, P. (2011). Developmental change in working memory strategies: From passive maintenance to active refreshing. *Developmental psychology, 47*, 898-904.

Camos, V., Lagner, P., & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language, 61*, 457-469. doi:http://doi.org/10.1016/j.jml.2009.06.002

Camos, V., Lagner, P., & Loaiza, V. M. (2016). Maintenance of item and order information in verbal working memory. *Memory*. doi:10.1080/09658211.2016.1237654

Campbell, T., Beaman, C. P., & Berry, D. C. (2002). Auditory memory and the irrelevant sound effect: Further evidence for changing-state disruption. *Memory, 10*, 199-214.

Campoy, G., & Baddeley, A. D. (2008). Phonological and semantic strategies in immediate serial recall. *Memory, 16*, 329-340.

Cardozo, B. L., & Leopold, F. F. (1963). Human code transmission. *Ergonomics, 6*, 133-141.

Carlesimo, G. A., Marfia, G. A., Loasses, A., & Caltagirone, C. (1996). Recency effect in anterograde amnesia: Evidence for distinct memory stores underlying enhanced retrieval of terminal items in immediate and delayed recall paradigms. *Neuropsychologia, 34*, 177-184.

Cave, C. B., & Squire, L. R. (1992). Intact verbal and nonverbal short-term memory following damage to the human hippocampus. *Hippocampus, 2*(2), 151-163. doi:10.1002/hipo.450020207

Chase, W. G., & Simon, H. A. (1973a). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215-281). New York: Academic Press.

Chase, W. G., & Simon, H. A. (1973b). Perception in Chess. *Cognitive Psychology, 4*, 55-81.

Chein, J. M., & Fiez, J. A. (2001). Dissociation of verbal working memory system components using a delayed serial recall task. *Cerebral Cortex, 11*(11), 1003-1014.

Chein, J. M., Moore, A. B., & Conway, A. R. A. (2011). Domain-general mechanisms of complex working memory span. *NeuroImage, 54*, 550-559. doi:http://doi.org/10.1016/j.neuroimage.2010.07.067

Chen, Z., & Cowan, N. (2005). Chunk limits and length limits in immediate recall: a reconciliation. *Journal of Experimental Psychology: Learning, Memory & Cognition, 31*, 1235-1249.

Chen, Z., & Cowan, N. (2009). Core verbal working-memory capacity: The limit in words retained without covert articulation. *Quarterly Journal of Experimental Psychology, 62*, 1420-1429.

Cho, S., Holyoak, K. J., & Cannon, T. D. (2007). Analogical reasoning in working memory: Resources shared among relational integration, interference resolution, and maintenance. *Memory & Cognition, 35*, 1445-1455.

Chow, M., & Conway, A. R. A. (2015). The scope and control of attention: Sources of variance in working memory capacity. *Memory & Cognition, 43*, 325-339. doi:10.3758/s13421-014-0496-9

Chow, M., Macnamara, B. N., & Conway, A. R. A. (2016). Phonological similarity in working memory span tasks. *Memory & Cognition, 44*, 937-949. doi:10.3758/s13421-016-0609-8

Christie, D. F. M., & Phillips, W. A. (1979). Simple drawing and pattern completion techniques for studying visualization and long-term visual knowledge. *Memory & Cognition, 7*, 360-367.

Chuderski, A. (2014). The relational integration task explains fluid reasoning above and beyond other working memory tasks. *Memory & Cognition, 42*, 448-463. doi:10.3758/s13421-013-0366-x

Chuderski, A. (2015). The broad factor of working memory is virtually isomorphic to fluid intelligence tested under time pressure. *Personality and Individual Differences, 85*, 98-104. doi:10.1016/j.paid.2015.04.046

Chuderski, A., Taraday, M., Nęcka, E., & Smoleń, T. (2012). Storage capacity explains fluid intelligence but executive control does not. *Intelligence, 40*, 278-295.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General, 120*, 235-253.

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature, 386*(6625), 604-608. doi:10.1038/386604a0

Cole, M., Frankel, F., & Sharp, D. (1971). Development of free recall learning in children. *Developmental psychology, 4*(109-123).

Colle, H. A., & Welsh, A. (1976). Acoustic masking in primary memory. *Journal of Verbal Learning and Verbal Behavior, 15*, 17-32.

Collette, F., Van der Linden, M., Bechet, S., & Salmond, E. (1999). Phonological loop and central executive functioning in Alzheimer's disease. *Neuropsychologia, 37*, 905-918.

Conlin, J. A., & Gathercole, S. E. (2006). Lexicality and interference in working memory in children and in adults. *Journal of Memory and Language, 55*, 363-380. doi:http://doi.org/10.1016/j.jml.2006.05.003

Conlin, J. A., Gathercole, S. E., & Adams, J. W. (2005). Stimulus similarity decrements in children's working memory span. *Quarterly Journal of Experimental Psychology, 58A*, 1434-1446.

Conrad, R. (1971). The chronology of the development of covert speech in children. *Developmental psychology, 5*, 398-405.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Pychology, 55*, 429-432.

Conrad, R., & Hull, A. J. (1968). Input modality and the serial position curve in short-term memory. *Psychonomic Science, 10*, 135-136. doi:10.3758/BF03331446

Copeland, D. E., & Radvansky, G. A. (2001). Phonological similarity in working memory. *Memory & Cognition, 29*, 774-776.

Corballis, M. C. (1967). Serial order in recognition and recall. *Journal of Experimental Psychology, 74*, 99-105.

Cornell, E. H., & Bergstrom, L. I. (1983). Serial-position effects in infants' recognition memory. *Memory & Cognition, 11*, 494-499.

Cortis, C., Dent, K., Kennett, S., & Ward, G. (2015). First things first: Similar list length and output order effects for verbal and nonverbal stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 1179-1214. doi:10.1037/xlm0000086

Courtney, S. M., Petit, L., Maisog, J. M., Ungerleider, L. G., & Haxby, J. V. (1998). An area specialized for spatial working memory in human frontal cortex. *Science, 279*(5355), 1347-1351.

Courtney, S. M., Ungerleider, L. G., Keil, K., & Haxby, J. V. (1996). Object and spatial visual working memory activate separate neural systems in human cortex. *Cerebral Cortex, 6*(1), 39-49.

Couture, M., & Tremblay, S. (2006). Exploring the characteristics of the visuo-spatial Hebb repetition effect. *Memory & Cognition, 34*, 1720-1729

Cowan, N. (1992). Verbal memory span and the timing of spoken recall. *Journal of Memory and Language, 31*, 668-684.

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

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

Cowan, N. (in press). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-016-1191-6

Cowan, N., Chen, Z., & Rouder, J. N. (2004). Constant capacity in an immediate serial-recall task: A logical sequel to Miller (1956). *Psychological Science, 15*, 634-640.

Cowan, N., Day, L., Saults, S., Keller, T. A., Johnson, T., & Flores, L. (1992). The role of verbal output time in the effects of word length on immediate memory. *Journal of Memory and Language, 31*, 1-7.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology, 51*, 42-100.

Cowan, N., Hismjatullina, A., AuBuchon, A. M., Saults, J. S., Horton, N., Leadbitter, K., & Towse, J. N. (2010). With Development, list recall includes more chunks, not just larger ones. *Developmental psychology, 46*, 1119-1131. doi:10.1037/a0020618

Cowan, N., Keller, T., Hulme, C., Roodenrys, S., McDougall, S., & Rack, J. (1994). Verbal memory span in children: Speech timing clues to the mechanisms underlying age and word length effects. *Journal of Memory and Language, 33*, 234-250.

Cowan, N., Li, D., Moffitt, A., Becker, T. M., Martin, E. A., Saults, J. S., & Christ, S. E. (2011). A neural region of abstract working memory. *Journal of Cognitive Neuroscience, 23*, 2852-2863.

Cowan, N., & Morey, C. C. (2007). How can dual-task working memory retention limits be investigated? *Psychological Science, 18*, 686-688. doi:http://doi.org/10.1111/j.1467-9280.2007.01960.x

Cowan, N., Rouder, J. N., Blume, C. L., & Saults, J. S. (2012). Models of verbal working memory capacity: What does it take to make them work? *Psychological Review, 119*, 480-499 doi:10.1037/a0027791

Cowan, N., Saults, J. S., & Blume, C. L. (2014). Central and peripheral components of working memory storage. *Journal of Experimental Psychology: General, 143*, 1806-1836. doi:10.1037/a0036814

Cowan, N., Saults, J. S., Elliott, E. M., & Moreno, M. V. (2002). Deconfounding serial recall. *Journal of Memory and Language, 46*, 153-177.

Cowan, N., Wood, N. L., Wood, P. K., Keller, T. A., Nugent, L. D., & Keller, C. V. (1998). Two separate verbal processing rates contributing to short-term memory span. *Journal of Experimental Psychology: General, 127*, 141-160.

Crannell, C. W., & Parrish, J. M. (1957). A comparison of immediate memory span for digits, letters, and words. *Journal of Psychology, 44*, 319-327.

Crowder, R. G. (1968). Intraserial repetition effects in immediate memory. *Journal of Verbal Learning & Verbal Behavior, 7*, 446-451.

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics, 5*, 365-373.
D'Esposito, M., & Postle, B. R. (1999). The dependence of span and delayed-response performance on prefrontal cortex. *Neuropsychologia, 37*(11), 1303-1315.
Dalezman, J. J. (1976). Effects of output order on immediate, delayed, and final recall performance. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 597–608
Della Sala, S., Gray, C., Baddeley, A. D., Allamano, N., & Wilson, L. (1999). Pattern span: a tool for unwelding visuo-spatial memory. *Neuropsychologia, 37*, 1189-1199.
Delvenne, J.-B., & Bruyer, R. (2004). Does visual short-term memory store bound features? *Visual Cognition, 11*, 1-27.
Dempster, F. N., & Rowher, W. D., Jr. (1983). Age differences and modality effects in immediate and delayed free recall. *Child Development, 54*, 30-41.
Denney, N. W. (1974). Clustering in middle and old age. *Developmental psychology, 10*, 471-4775.
Depoorter, A., & Vandierendonck, A. (2009). Evidence for modality-independent order coding in working memory. *Quarterly Journal of Experimental Psychology, 62*, 531-549.
Devkar, D., Wright, A. A., & Ma, W. J. (2015). The same type of visual working memory limitations in humans and monkeys. *J Vision, in press*.
Donkin, C., & Nosofsky, R. M. (2012a). A power-law model of psychological memory strength in short- and long-term recognition. *Psychological Science, 23*, 625-634.
Donkin, C., & Nosofsky, R. M. (2012b). The structure of short-term memory scanning: an investigation using response time distribution models. *Psychonomic Bulletin & Review, 19*, 363-394.
Donkin, C., Nosofsky, R. M., Gold, J. M., & Shiffrin, R. M. (2013). Discrete-slot models of visual working-memory response times. *Psychological Review, 120*, 873-902.
Donkin, C., Tran, S. C., & Pelley, M. L. (2015). Location-based errors in change detection: A challenge for the slots model of visual working memory. *Memory & Cognition, 43*, 421-431. doi:10.3758/s13421-014-0487-x
Dornbush, R. (1968a). Input variables in bisensory memory. *Perception & Psychophysics, 4*, 41-44.
Dornbush, R. (1968b). Shadowing in bisensory memory. *Quarterly Journal of Experimental Psychology, 20*, 225-231.
Dornbush, R. (1969). Stimulus information and stimulus interference in bisensory short-term memory. *Perception & Psychophysics, 5*, 303-304.
Drewnowski, A., & Murdock, B. B. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Learning, Memory & Cognition, 6*, 319-332.
Duarte, A., Hearons, P., Jiang, Y., Delvin, M. C., Newsome, R. N., & Verhaeghen, P. (2013). Retrospective attention enhances visual working memory in the young but not the old: An ERP study. *Psychophysiology, 50*, 465-476.
Duncan, M., & Murdock, B. (2000). Recognition and recall with precuing and postcuing. *Journal of Memory and Language, 42*, 301-313.
Elhalal, A., Davelaar, E. J., & Usher, M. (2014). The role of the frontal cortex in memory: an investigation of the Von Restorff effect. *frontiers in Human Neuroscience, 8*. Retrieved from doi:10.3389/fnhum.2014.00410
Ellermeier, W., & Zimmer, K. (2014). The psychoacoustics of the irrelevant sound effect. *Acoustical Science and Technology, 35*, 10-16.
Elliott, E. M. (2002). The irrelevant speech effect and children: Theoretical implications of developmental change. *Memory & Cognition, 30*, 478-487.
Elliott, E. M., Cherry, K. E., Brown, J. S., Smitherman, E. A., Jazwinski, S. M., Yu, Q., & Volaufova, J. (2011). Working memory in the oldest-old: evidence from output serial position curves. *Memory & Cognition, 39*, 1423-1434. doi:10.3758/s13421-011-0119-7
Elliott, E. M., Hughes, R. W., Briganti, A., Joseph, T. N., Marsh, J. E., & Macken, B. (2016). Distraction in verbal short-term memory: Insights from developmental differences. *Journal of Memory and Language, 88*, 39-50. doi:10.1016/j.jml.2015.12.008

Emrich, S. M., & Ferber, S. (2012). Competition increases binding errors in visual working memory. *Journal of Vision, 12*, 1-16.
Endress, A. D., & Potter, M. C. (2014). Large capacity temporary visual memory. *Journal of Experimental Psychology: General, 143*, 548-565. doi:10.1037/a0033934
Engel de Abreu, P. M. J., Conway, A. R. A., & Gathercole, S. E. (2010). Working memory and fluid intelligence in young children. *Intelligence, 38*, 552-561.
Engle, R. W., Tuholski, S. W., Laughlin, J., & Conway, A. R. A. (1999). Working memory, short-term memory and general fluid intelligence: A latent variable model approach. *Journal of Experimental Psychology: General, 128*, 309-331.
Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science, 208*, 1181-1182.
Fallon, A. B., Groves, K., & Tehan, G. (1999). Phonological similarity and trace degradation in the serial recall task: When CAT helps RAT, but not MAN. *International Journal of Psychology (Special Issue: Short-Term/Working Memory), 34*, 301-307.
Farrand, P., & Jones, D. (1996). Direction of report in spatial and verbal serial short-term memory. *Quarterly Journal of Experimental Psychology, 49A*, 140-158.
Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review, 119*(2), 223-271.
Farrell, S., Hurlstone, M. J., & Lewandowsky, S. (2013). Sequential dependencies in recall of sequences: Filling in the blanks. *Memory & Cognition, 41*, 938-952. doi:10.3758/s13421-013-0310-0
Farrell, S., & Lelièvre, A. (2009). End anchoring in short-term order memory. *Journal of Memory and Language, 60*, 209-227.
Farrell, S., & Lewandowsky, S. (2003). Dissimilar items benefit from phonological similarity in serial recall. *Journal of Experimental Psychology: Learning, Memory & Cognition, 29*, 838-849.
Farrell, S., & Lewandowsky, S. (2004). Modelling transposition latencies: Constraints for theories of serial order memory. *Journal of Memory and Language, 51*, 115-135.
Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science, 19*, 329-335.
Farrell, S., & McLaughlin, K. (2007). Short-term recognition memory for serial order and timing. *Memory & Cognition, 35*(7), 1724-1734.
Farrell, S., Wise, V., & Lelievre, A. (2011). Relations between timing, position, and grouping in short-term memory. *Memory & Cognition, 39*(4), 573-587. doi:10.3758/s13421-010-0053-0
Farrell, S., Wise, V., & Lelièvre, A. (2011). Relations between timing, position, and grouping in short-term memory. *Memory & Cognition, 39*, 573-587.
Fell, J. C., & Laugherty, K. R. (1969). Short-term memory: Mode of presentation for alphanumeric information. *Human Factors, 11*, 401-406.
Ferraro, F. R., & Balota, D. A. (1999). Memory scanning performance in healthy young adults, healthy older adults, and individuals with dementia of the Alzheimer type. *Aging, Neuropsychology, and Cognition, 6*, 260-272.
Fiez, J. A., Raife, E. A., Balota, D. A., Schwarz, J. P., Raichle, M. E., & Petersen, S. E. (1996). A positron emission tomography study of the short-term maintenance of verbal information. *Journal of Neuroscience, 16*(2), 808-822.
Finke, C., Braun, M., Ostendorf, F., Lehmann, T. N., Hoffmann, K. T., Kopp, U., & Ploner, C. J. (2008). The human hippocampal formation mediates short-term memory of colour-location associations. *Neuropsychologia, 46*(2), 614-623. doi:10.1016/j.neuropsychologia.2007.10.004
Fischer-Baum, S., & McCloskey, M. (2015). Representation of item position in immediate serial recall: evidence from intrusion errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 1426-1446. doi:10.1037/xlm0000102
Floden, D., Stuss, D. T., & Craik, F. I. M. (2000). Age differences in performance on two versions of the Brown-Peterson task. *Aging, Neuropsychology, and Cognition, 7*, 245-259.
Forrin, B., & Morin, R. E. (1969). Recognition times for items in short- and long-term memory. *Acta Psychologica, 30*, 126-141.

Fougnie, D., & Marois, R. (2006). Distinct capacity limits for attention and working memory: Evidence from attentive tracking and visual working memory paradigms. *Psychological Science, 17*, 526-534.
Fougnie, D., & Marois, R. (2009). Dual-task interference in visual working memory: A limitation in storage capacity but not in encoding or retrieval. *Attention, Perception, & Psychophysics, 71*, 1831-1841.
Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nat Commun, 3*, 1229.
Frankish, C. (2008). Precategorical acoustic storage and the perception of speech. *Journal of Memory and Language, 58*, 815-836. doi:10.1016/j.jml.2007.06.003
Frankish, C. R. (1989). Perceptual organization and precategorical acoustical storage. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15*(3), 469-479.
Frankish, C. R. (1995). Intonation and auditory grouping in immediate serial recall. *Applied Cognitive Psychology, 9*, 5-22.
Fuchs, A. H. (1969). Recall for order and content of serial word lists in short-term memory. *Journal of Experimental Psychology, 82*, 14-21.
Fuster, J. M., & Alexander, G. E. (1971). Neuron activity related to short-term memory. *Science, 173*(3997), 652-654.
Gao, Z., Gao, Q., Tang, N., Shui, R., & Shen, M. (2015). Organization principles in visual working memory: Evidence from sequential stimulus display. *Cognition, 146*, 277-288. doi:10.1016/j.cognition.2015.10.005
Garavan, H. (1998). Serial attention within working memory. *Memory & Cognition, 26*, 263-276.
Gardiner, J. M., Craik, F. I. M., & Birtwistle, J. (1972). Retrieval cues and release from proactive inhibition. *Journal of Verbal Learning and Verbal Behavior, 11*, 778-783.
Gardiner, J. M., & Gregg, V. H. (1979). When auditory memory is not overwritten. *Journal of Verbal Learning and Verbal Behavior, 18*, 705-719.
Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 84-95.
Gathercole, S. E., Pickering, S. J., Ambridge, S. J., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental psychology, 40*, 177-190.
Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends in Cognitive Sciences, 16*, 129-135. doi:10.1016/j.tics.2011.11.014
Gehring, W. J., Bryck, R. L., Jonides, J., Albin, R. L., & Badre, D. (2003). The mind's eye, looking inward? In search of executive control in internal attention shifting. *Psychophysiology, 40*, 572-585.
Geiger, S. M., & Lewandowsky, S. (2008). Temporal isolation does not facilitate forward serial recall - or does it? *Memory & Cognition, 36*, 957-967.
Gerver, D. (1969). Effects of grammaticalness, presentation rate, and message length on auditory short-term memory. *Quarterly Journal of Experimental Psychology, 21*, 203-208.
Gescheider, G. A. (1997). *Psychophysics: the fundamentals*: Psychology Press.
Gianutsos, R. (1972). Free-recall of grouped words. *Journal of Experimental Psychology, 95*(2), 419-+. doi:10.1037/h0033639
Gignac, G. E. (2014). Fluid intelligence shares closer to 60% of its variance with working memory capacity and is a better indicator of general intelligence. *Intelligence, 47*, 122-133.
Gilchrist, A. L., & Cowan, N. (2014). A two-stage search of visual working memory: investigating speed in the change-detection paradigm. *Attention, Perception, & Psychophysics, 76*, 2031-2050.
Gisselgård, J., Uddén, J., Ingvar, M., & Peterson, K. M. (2007). Disruption of order information by irrelevant items: A serial recognition paradigm. *Acta Psychologica, 124*, 356-369.
Glahn, D. C., Kim, J., Cohen, M. S., Poutanen, V.-P., Therman, S., Bava, S., . . . Cannon, T. D. (2002). Maintenance and manipulation in spatial working memory: Dissociations in the prefrontal cortex. *NeuroImage, 17*, 201-213. doi:10.1006/nimg.2002.1161

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning & Verbal Behavior, 5*, 351-360.

Glanzer, M., Gianutsos, R., & Dubin, S. (1969). The removal of items from short-term storage. *Journal of Verbal Learning and Verbal Behavior, 8*, 435-447.

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

Glenberg, A. M. (1984). A retrieval account of the long-term modality effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 16-31.

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: Evidence for the magical number four -- or is it two? *Memory, 12*, 732-747.

Gobet, F., & Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology, 31*, 1-40.

Gobet, F., & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory, 6*, 225-255.

Goldman-Rakic, P. S. (1987). Circuitry of the primate prefrontal cortex and the regulation of behavior by representational memory. In F. Plum & V. B. Mountcastle (Eds.), *Handbook of physiology* (Vol. 5, pp. 373-417). Bethseda, MD: American Physiological Society.

Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron, 14*(3), 477-485.

Golomb, J. D., Peelle, J. E., Addis, K. M., Kahana, M. J., & Wingfield, A. (2008). Effects of adult aging on utilization of temporal and semantic associations during free and serial recall. *Memory & Cognition, 36*, 947-956. doi:10.3758/MC.36.5.947

Gong, Y., Ericsson, K. A., & Moxley, J. H. (2015). Recall of briefly presented chess positions and its relation to chess skill. *PLoS ONE, 10*. Retrieved from doi:10.1371/journal.pone.0118756

Gorgoraptis, N., Catalao, R. F. G., Bays, P. M., & Husain, M. (2011). Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience, 31*, 8502-8511.

Gregg, V. H., Freedman, C. M., & Smith, D. K. (1989). Word frequency, articulatory suppression and memory span. *British Journal of Psychology, 80*, 363-374.

Grenfell-Essam, R., & Ward, G. (2012). Examining the relationship between free recall and immediate serial recall: the role of list length, strategy use, and test expectancy. *Journal of Memory and Language, 67*, 106-148.

Gressmann, M., & Janczyk, M. (2016). The (un)clear effects of invalid retro-cues. *frontiers in Psychology, 7*. Retrieved from doi:10.3389/fpsyg.2016.00244

Griffin, I. C., & Nobre, A. C. (2003). Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience, 15*, 1176-1194.

Groeger, J. A., Banks, A. P., & Simpson, P. J. (2008). Serial memory for sound-specified locations: Effects of spatial uncertainty and motor suppression. *The Quarterly Journal of Experimental Psychology, 61:2*, 248-262. doi:10.1080/17470210601138746

Gruber, O., & von Cramon, D. Y. (2003). The functional neuroanatomy of human working memory revisited. Evidence from 3-T fMRI studies using classical domain-specific interference tasks. *NeuroImage, 19*(3), 797-809.

Guerard, K., Hughes, R. W., & Tremblay, S. (2008). An isolation effect in serial memory for spatial information. *Quarterly Journal of Experimental Psychology, 61*(5), 752-762. doi:10.1080/17470210701402331

Guérard, K., Neath, I., Surprenant, A., & Tremblay, S. (2010). Distinctiveness in serial memory for spatial information. *Memory & Cognition, 38*(1), 83-91.

Guérard, K., Saint-Aubin, J., Burns, S. C., & Chamberland, C. (2012). Revisiting backward recall and benchmark memory effects: A reply to Bireta et al. (2010). *Memory & Cognition, 40*, 388-407.

Guérard, K., & Tremblay, S. (2008). Revisiting evidence for modularity and functional equivalence across verbal and spatial domains in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 556-569.

Guilford, J. P., & Dallenbach, K. M. (1925). The determination of memory span by the method of constant stimuli. *American Journal of Psychology, 36*, 621-628.

Gustafsson, J.-E., & Wolff, U. (2015). Measuring fluid intelligence at age four. *Intelligence, 50*, 175-185. doi:10.1016/j.intell.2015.04.008

Hale, S., Bronik, M. D., & Fry, A. E. (1997). Verbal and spatial working memory in school-age children: Developmental differences in susceptibility to interference. *Developmental psychology, 33*, 364-371.

Hale, S., Myerson, J., Rhee, S. H., Weiss, C. S., & Abrams, R. A. (1996). Selective interference with the maintenance of location information in working memory. *Neuropsychology, 20*, 228-240.

Hale, S., Rose, N. S., Myerson, J., Strube, M. J., Sommers, M., Tye-Murray, N., & Spehar, B. (2011). The structure of working memory abilities across the adult life span. *Psychology and Aging, 26*, 92-110.

Halliday, M. S., Hitch, G. J., Lennon, B., & Pettipher, C. (1990). Verbal short-term memory in children: The role of the articulatory loop. *European Journal of Cognitive Psychology, 2*, 23-38.

Handley, S., Capon, A., Copp, C., & Harper, C. (2002). Conditional reasoning and the Tower of Hanoi: The role of spatial and verbal working memory. *British Journal of Psychology, 93*, 501-518.

Hannula, D. E., Tranel, D., & Cohen, N. J. (2006). The long and the short of it: relational memory impairments in amnesia, even at short lags. *Journal of Neuroscience, 26*(32), 8352-8359. doi:10.1523/JNEUROSCI.5222-05.2006

Harris, G. J., & Burke, D. (1972). The effecs of grouping on short-term serial recall of digits by children: Developmental trends. *Child Development, 43*, 710-716.

Hartley, T., Hurlstone, M. J., & Hitch, G. J. (2016). Effects of rhythm on memory for spoken sequences: A model and tests of its stimulus-driven mechanism. *Cognitive Psychology, 87*, 135-178. doi:10.1016/j.cogpsych.2016.05.001

Hartshorne, J. K. (2008). Visual working memory capacity and proactive interference. *PLoS ONE, 3*, e2716.

Harvey, A. J., & Beaman, C. P. (2007). Input and output modality effects in immediate serial recall. *Memory, 15*, 693-700.

Hasselhorn, M., & Grube, D. (2003). The phonological similarity effect on memory span in children: Does it depend on age, speech rate, and articulatory suppression? . *International Journal of Behavioral Development, 7*, 145-152.

Hayes, D. S., & Schulze, S. A. (1977). Visual encoding in preschoolers' serial retention. *Child Development*, 1066-1070.

Healey, M. K., & Kahana, M. J. (2014). Is memory search governed by universal principles or idiosyncratic strategies? *Journal of Experimental Psychology: General, 143*, 575-596. doi:10.1037/a0033715

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. In J. F. Delafresnaye (Ed.), *Brain mechanisms and learning* (pp. 37-46). Oxford: Blackwell.

Hecht, S. A. (2002). Counting on working memory in simple arithmetic when counting is used for problem solving. *Memory & Cognition, 30*, 447-455.

Hedge, C., & Leonards, U. (2013). Using eye movements to explore switch costs in working memory. *Journal of Vision, 13*, 1-19. doi:10.1167/13.4.18

Hedge, C., Oberauer, K., & Leonards, U. (2015). Selection in spatial working memory is independent of perceptual selective attention, but they interact in a shared spatial priority map. *Attention, Perception & Psychophysics, 77*, 26653-22668. doi:10.3758/s13414-015-0976-4

Heitz, R., & Engle, R. W. (2007). Focusing the spotlight: Individual differences in visual attention control. *Journal of Experimental Psychology: General, 136*, 217-240.

Henry, L. A. (1991). The effects of word length and phonemic similarity in young children's short-term memory. *Quarterly Journal of Experimental Psychology, 43*, 35-52.

Henry, L. A. (2007). The effects of word length and phonemic similarity in young children's short-term memory. *Quarterly Journal of Experimental Psychology, 43A*, 35-52.

Henson, R. N. A. (1996). *Short-term memory for serial order.* (Dissertation), University of Cambridge. Retrieved from http://www.mrc-cbu.cam.ac.uk/~rh01/thesis.html

Henson, R. N. A. (1998a). Item repetition in short-term memory: Ranschburg repeated. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1162-1181.

Henson, R. N. A. (1998b). Short-term memory for serial order: The Start-End Model. *Cognitive Psychology, 36*, 73-137.
Henson, R. N. A. (1999). Positional information in short-term memory: Relative or absolute? *Memory & Cognition, 27*, 915-927.
Henson, R. N. A., Norris, D. G., Page, M. P. A., & Baddeley, A. D. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. *Quarterly Journal of Experimental Psychology, 49A*, 80-115.
Hertzog, C., Dixon, R. A., Hultsch, D. F., & MacDonald, S. W. S. (2003). Latent change models of adult cognition: Are changes in processing speed and working memory associated with changes in episodic memory? *Psychology and Aging, 18*, 755-769.
Heyselaar, E., Johnston, K., & Pare, M. (2011). A change detection approach to study visual working memory of the macaque monkey. *J Vision, 11*(3), 11:11-10.
Hitch, G. J. (1974). Short-term memory for spatial and temporal information. *Quarterly Journal of Experimental Psychology, 26*, 503-513.
Hitch, G. J., Burgess, N., Towse, J. N., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *Quarterly Journal of Experimental Psychology, 49A*, 116-139.
Hitch, G. J., Fastame, M. C., & Flude, B. (2005). How is the serial order of a verbal sequence coded? Some comparisons between models. *Memory, 13*, 247-258.
Hitch, G. J., Flude, B., & Burgess, N. (2009). Slave to the rhythm: Experimental tests of a model for verbal short-term memory and long-term sequence learning. *Journal of Memory and Language, 61*, 97-111. doi:10.1016/j.jml.2009.02.004
Hitch, G. J., Halliday, M. S., Dodd, A., & Littler, J. (1989). Development of rehearsal in short-term memory: Differences between pictorial and spoken stimuli. *British Journal of Developmental Psychology, 7*, 347-362.
Hitch, G. J., Halliday, S., Schaafstal, A. M., & Schraagen, J. M. C. (1988). Visual working memory in young children. *Memory & Cognition, 16*, 120-132.
Hitch, G. J., Woodin, M. E., & Baker, S. (1989). Visual and phonological components of working memory in children. *Memory & Cognition, 17*, 175-185.
Hogan, R. M. (1975). Interitem encoding and directed search in free recall. *Memory & Cognition, 3*, 197-209.
Hollingworth, A., & Maxcey-Richard, A. M. (2013). Selective maintenance in visual working memory does not require sustained visual attention. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 1047-1058. doi:10.1037/a0030238
Horton, N. J., Hay, D. C., & Smyth, M. M. (2008). Hebb repetition effects in visual memory: The roles of verbal rehearsal and distinctiveness. *Quarterly Journal of Experimental Psychology, 61*, 1769-1777.
Hoving, K. L., Morin, R. E., & Konick, D. S. (1970). Recognition reaction time and size of the memory set: A developmental study. *Psychonomic Science, 21*, 247-248.
Howard, M. C., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 923-941.
Howard, M. C., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language, 46*, 85-98. doi:10.1006/jmla.2001.2798
Hudjetz, A., & Oberauer, K. (2007). The effects of processing time and processing rate on forgetting in working memory: Testing four models of the complex span paradigm. *Memory & Cognition, 35*(7), 1675-1684.
Hughes, R. W., Hurlstone, M. J., Marsh, J. E., Vachon, F., & Jones, D. M. (2013). Cognitive control of auditory distraction: impact of task difficulty, foreknowledge, and working memory capacity supports duplex-mechanism account. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 539-553.
Hughes, R. W., Vachon, F., & Jones, D. M. (2007). Disruption of short-term memory by changing and deviant sounds: support for a duplex-mechanism account of auditory distraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 1050-1061.

Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language, 30*, 685-701. doi:10.1016/0749-596X(91)90032-F

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1217-1232. doi:10.1037/0278-7393.23.5.1217

Hulme, C., Thomson, N., Muir, C., & Lawrence, A. (1984). Speech rate and development of short-term memory span. *Journal of Experimental Child Psychology, 38*, 241-253.

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

Humphreys, M. S., Maguire, A. M., McFarlane, K. A., Burt, J. S., Bolland, S. W., Murray, K. L., & Dunn, R. (2010). Using maintenance rehearsal to explore recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 147-159.

Hunt, R. R. (1995). The subtlety of distinctiveness - what vonrestorff really did. *Psychonomic Bulletin & Review, 2*(1), 105-112.

Hunt, R. R., & Worthen, J. B. E. (2006). Distinctiveness and Memory. *Canada: Oxford University Press*.

Hurlstone, M. J., & Hitch, G. J. (2015). How is the serial order of a spatial sequence represented? Insights from transposition latencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 295-324. doi:10.1037/a0038223

Hurlstone, M. J., & Hitch, G. J. (2017). How is the serial order of a visual sequence represented? Insights from transposition latencies. *Journal of Experimental Psychology: Learning, Memory and Cognition*. doi:10.1037/xlm0000440

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological Bulletin, 140*, 339-373. doi:10.1037/a0034221

Hutchison, K. A. (2011). The interactive effects of listwide control, item-based control, and working memory capacity on Stroop performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 851-860.

Jackson, M. C., & Raymond, J. E. (2008). Familiarity enhances visual working memory for faces. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 556-568.

Jaeggi, S. M., Schmid, C., Buschkuehl, M., & Perrig, W. J. (2008). Differential age effects in load-dependent memory processing. *Aging, Neuropsychology, and Cognition, 16*, 80-102. doi:10.1080/13825580802233426

Jahnke, J. C. (1969). The Ranschburg Effect. *Psychological Review, 76*(6), 592-560S.

Jalbert, A., Saint-Aubin, J., & Tremblay, S. (2008). Visual similarity in short-term recall for where and when. *Quarterly Journal of Experimental Psychology, 61*(3), 353-360.

Jarrold, C., & Citroën, R. (2013). Reevaluating key evidence for the development of rehearsal: phonological similarity effects in children are subject to proportional scaling artifacts. *Developmental psychology, 49*(5), 837

Jarrold, C., Cocksey, J., & Dockerill, E. (2008). Phonological similarity and lexicality effects in children's verbal short-term memory: concerns about the interpretation of probed recall data. *Quarterly Journal of Experimental Psychology, 61*, 324-340.

Jarrold, C., Hall, D., Harvey, C. E., Tam, H., Towse, J. N., & Zarandi, A. L. (2015). What can we learn about immediate memory from the development of children's free recall? *Quarterly Journal of Experimental Psychology, 68*, 1871-1894. doi:10.1080/17470218.2014.995110

Jarrold, C., Tam, H., Baddeley, A. D., & Harvey, C. E. (2010). The nature and the position of processing determines why forgetting occurs in working memory tasks. *Psychonomic Bulletin & Review, 17*, 772-777.

Jarrold, C., Tam, H., Baddeley, A. D., & Harvey, C. E. (2011). How does processing affect storage in working memory tasks? Evidence for both domain-general and domain-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 688-705.

Jefferies, E., Frankish, C., & Lambon Ralph, M. A. (2006). Lexical and semantic binding in verbal short-term memory. *Journal of Memory and Language, 54*, 81-98.

Jeneson, A., Mauldin, K. N., & Squire, L. R. (2010). Intact working memory for relational information after medial temporal lobe damage. *Journal of Neuroscience, 13*, 13624-13629.

Jeneson, A., & Squire, L. (2012). Working memory, long-term memory, and medial temporal lobe function. *Learning & Memory, 19*, 15-25.

Jenkins, J. J., & Russell, W. A. (1952). Associative clustering during recall. *Journal of Abnormal and Social Psychology, 47*, 818-821. doi:10.1037/h0063149

Jones, D. M. (1993). Objects, streams and threads of auditory attention. In A. D. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness and control* (pp. 87-104). Oxford: Clarendon.

Jones, D. M., Alford, D., Bridges, A., Tremblay, S., & Macken, B. (1999). Organizational factors in selective attention: The interplay of acoustic distinctiveness and auditory streaming in the irrelevant sound effect. *Journal of Experimental Psychology: Learning, Memory & Cognition, 25*, 464-473.

Jones, D. M., Farrand, P., Stuart, G., & Morris, N. (1995). Functional equivalence of verbal and spatial information in serial short-term memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 21*, 1008-1018.

Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 19*, 369-381.

Jones, D. M., & Macken, W. J. (1995). Phonological similarity in the irrelevant speech effect: Within- or between-stream similarity? *Journal of Experimental Psychology: Leaming, Memory, and Cognition, 21*, 103-115.

Jones, D. M., Madden, C., & Miles, C. (1992). Privileged access by irrelevant speech to short-term memory: The role of changing state. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 44A*, 645-669.

Jones, D. M., Miles, C., & Page, J. (1990). Disruption of reading by irrelevant speech: Effects of attention, arousal or memory? *Journal of Applied Cognitive Psychology, 4*, 89-108.

Jones, D. M., & Tremblay, S. (2000). Interference in memory by process or content? A reply to Neath (2000). *Psychonomic Bulletin & Review, 7*, 550-558.

Jonides, J., Schumacher, E. H., Smith, E. E., Lauber, E. J., Awh, E., Minoshima, S., & Koeppe, R. A. (1997). Verbal working memory load affects regional brain activation as measured by PET. *Journal of Cognitive Neuroscience, 9*, 424-475.

Jonides, J., Smith, E. E., Marshuetz, C., Koeppe, R. A., & Reuter-Lorenz, P. (1998). Inhibition in verbal working memory revealed by brain activation. *Proceedings of the National Academy of Science, USA, 95*, 8410-8413.

Jost, K., Bryck, R. L., Vogel, E. K., & Mayr, U. (2010). Are old adults just like low working memory young adults? Filtering efficiency and age differences in visual working memory. *Cerebral Cortex, 21*, 147-1154.

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

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition, 24*, 103-109.

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: a noisy exemplar approach. *Vision Research, 42*, 2177-2192.

Kahana, M. J., Zhou, F., Geller, A. S., & Sekuler, R. (2007). Lure similarity affects visual episodic recognition: detailed tests of a noisy exemplar model. *Memory & Cognition, 35*, 1222-1232.

Kail, R. (2002). Developmental change in proactive interference. *Child Development, 73*, 1703-1714.

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working memory capacity: Individual differences in memory span and the control of visual orienting. *Journal of Experimental Psychology: General, 130*, 169-183.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the N-back task: A cautionary tale of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 47-70.

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

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

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle. *Psychological Bulletin, 131*, 66-71.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133(2)*, 189-217.

Kane, M. J., Meier, M. E., Smeekens, B. A., Gross, G. M., Chun, C. A., Silvia, P. J., & Kwapil, T. R. (2016). Individual differences in the executive control of attention, memory, and thought, and their associations with schizotypy. *Journal of Experimental Psychology: General, 145*, 1017-1048. doi:10.1037/xge0000184

Keating, D., & Bobbitt, B. (1978). Individual and developmental differences in cognitive-processing components of mental ability. *Child Development, 49*, 155-167.

Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning & Verbal Behavior, 1*, 153-161.

Kerr, J. R., Avons, S. E., & Ward, G. (1999). The effect of retention interval on serial position curves for item recognition of visual patterns and faces. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1475-1494.

Keshvari, S., Van den Berg, R., & Ma, W. J. (2012). Probabilistic computation in human perception under variability in encoding precision. *PLoS ONE, 7*(6), e40216.

Keshvari, S., Van den Berg, R., & Ma, W. J. (2013). No evidence for an item limit in change detection. *PLoS Comp Biol, 9*(2), e1002927.

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

Kiyonaga, A., & Egner, T. (2014). The working memory Stroop effect: When internal representations clash with external stimuli. *Psychological Science, 25*, 1619-1629.

Klatte, M., Kilcher, H., & Hellbrück, J. (1995). The effects of temporal structure of background noise on working memory. *Zeitschrift für Experimentelle Psychologie, 42*, 517-544.

Klatte, M., Lachmann, T., Schlittmeier, S., & Hellbruck, J. (2010). The irrelevant sound effect in short-term memory: Is there developmental change? *European Journal of Cognitive Psychology, 22*, 1168-1191.

Klauer, K. C., & Zhao, Z. (2004). Double dissociations in visual and spatial short-term memory. *Journal of Experimental Psychology: General, 133*, 355-381.

Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Increased brain activity in frontal and parietal cortex underlies the development of visuospatial working memory capacity during childhood. *Journal of Cognitive Neuroscience, 14*, 1-10.

Kopelman, M. D., & Stanhope, N. (1997). Rates of forgetting in organic amnesia following temporal lobe, diencephalic, or frontal lobe lesions. *Neuropsychology, 11*, 343-356.

Koppenol-Gonzalez, G. V., Bouwmeester, S., & Vermunt, J. K. (2014). Short-term memory development: Differences in serial position curves between age groups and latent classes. *Journal of Experimental Child Psychology, 126*, 138-151. doi:10.1016/j.jecp.2014.04.002

Korsnes, M. S., & Magnusson, S. (1996). Age comparisons of serial position effects in short-term memory. *Acta Psychologica, 94*, 133-143.

Kübler, A., Murphy, K., Kaufman, J., Stein, E. A., & Garavan, H. (2003). Co-ordination within and between verbal and visuospatial working memory: network modulation and anterior frontal recruitment. *NeuroImage, 20*, 1298-1308.
Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance in monkeys. *Journal of Neurophysiology, 34*(3), 337-347.
Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*, 389-433.
Laming, D. (1999). Testing the idea of distinct storage mechanisms in memory. *International Journal of Psychology, 34*, 419-426.
Landman, R., Spekreijse, H., & Lamme, V. A. F. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research, 43*(149-164).
Lange, E. B. (2005). Disruption of attention by irrelevant stimuli in serial recall. *Journal of Memory and Language, 53*, 513-531.
Lange, E. B., Cerella, J., & Verhaeghen, P. (2011). Ease of access to list items in short-term memory depends on the order of the recognition probes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(3), 608-620.
Lange, E. B., & Verhaeghen, P. (2009). No age differences in complex memory search: Older adults search as efficiently as younger adults. *Psychology and Aging, 24*, 105-115.
Langerock, N., Vergauwe, E., & Barrouillet, P. (2014). The maintenance of cross-domain associations in the episodic buffer. *Journal of Experimental Psychology: Learning, Memory & Cognition, 40*, 1096-1109.
LaPointe, L. B., & Engle, R. W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 1118-1133.
Lara, A. H., & Wallis, J. D. (2012). Capacity and precision in an animal model of short-term memory. *J Vision, 12*(3), 1-12.
LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding attended information in short-term memory: an EEG study. *Journal of Cognitive Neuroscience, 25*(1), 127-142. doi:10.1162/jocn_a_00305
Larsen, J. D., & Baddeley, A. D. (2003). Disruption of verbal STM by irrelevant speech, articulatory suppression, and manual tapping: Do they have a common source? *Quarterly Journal of Experimental Psychology, 56A*, 1249-1268.
LeCompte, D. C. (1994). Extending the irrelevant speech effect beyond serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1396-1408.
Lee, C. L., & Estes, W. K. (1977). Order and position in primary memory for letter strings. *Journal of Verbal Learning & Verbal Behavior, 16*, 395-418.
Lepsien, J., Griffin, I. C., Devlin, J. T., & Nobre, A. C. (2005). Directing spatial attention in mental representations: Interactions between attentional orienting and working-memory load. *NeuroImage, 26*, 733-743. doi:10.1016/j.neuroimage.2005.02.026
Lewandowsky, S., Brown, G. D. A., & Thomas, J. L. (2009). Traveling economically through memory space: Characterizing output order in memory for serial order. *Memory & Cognition, 37*, 181-193.
Lewandowsky, S., Brown, G. D. A., Wright, T., & Nimmo, L. M. (2006). Timeless memory: Evidence against temporal distinctiveness models of short-term memory for serial order. *Journal of Memory and Language, 54*, 20-38.
Lewandowsky, S., Duncan, M., & Brown, G. D. A. (2004). Time does not cause forgetting in short-term serial recall. *Psychonomic Bulletin & Review, 11*, 771-790.
Lewandowsky, S., & Farrell, S. (2008). Phonological similarity in serial recall: constraints on theories of memory. *Journal of Memory and Language, 58*, 429-448.
Lewandowsky, S., & Farrell, S. (2011). *Computational modeling in cognition: Principles and practice*. Thousand Oaks, CA: Sage.

Lewandowsky, S., Geiger, S. M., Morrell, D. B., & Oberauer, K. (2010). Turning simple span into complex span: time for decay or interference from distractors? *Journal of Experimental Psychology: Learning Memory and Cognition, 36*(4), 958-978.

Lewandowsky, S., Geiger, S. M., & Oberauer, K. (2008). Interference-based forgetting in verbal short-term memory. *Journal of Memory and Language, 59*(2), 200-222.

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review, 96*, 25-57.

Lewandowsky, S., Nimmo, L. M., & Brown, G. D. A. (2008). When temporal isolation benefits memory for serial order. *Journal of Memory and Language, 58*, 415-428.

Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2011). Neural evidence for a distinction between short-term memory and the focus of attention. *Journal of Cognitive Neuroscience, 24*, 61-79.

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

Li, K. Z. H. (1999). Selection from working memory: On the relationship between processing and storage components. *Aging, Neuropsychology, and Cognition, 6*, 99-116.

Li, Q., & Saiki, J. (2015). Different effects of color-based and location-based selection on visual working memory. *Attention, Perception & Psychophysics, 77*, 450-463. doi:10.3758/s13414-014-0775-3

Li, S. C., & Lewandowsky, S. (1993). Intralist distractors and recall direction: Constraints on models of memory for serial order. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 895-908.

Liefooghe, B., Barrouillet, P., Vandierendonck, A., & Camos, V. (2008). Working memory costs of task switching. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 478-494.

Lilienthal, L., Hale, S., & Myerson, J. (2014). The effects of environmental support and secondary tasks on visuospatial working memory. *Memory & Cognition, 42*, 1118-1129. doi:10.3758/s13421-014-0421-2

Lippman, L. G. (1980). Perceptual isolation and position judgment. *American Journal of Psychology, 93*(4), 683-695.

Lippman, L. G., & Lippman, M. Z. (1978). Isolation and similarity effects in a serial reconstruction task. *American Journal of Psychology, 91*(1), 35-50. doi:10.2307/1421820

Loess, H. (1964). Proactive inhibition in short-term memory. *Journal of Verbal Learning and Verbal Behavior, 3*, 362-368.

Logie, R. H. (2011). The functional organization and capacity limits of working memory. *Current Directions in Psychological Science, 20*, 240-245. doi:10.1177/0963721411415340

Logie, R. H., Della Sala, S., Wynn, V., & Baddeley, A. D. (2000). Visual similarity effects in immediate verbal serial recall. *Quarterly Journal of Experimental Psychology, 53*, 626-646.

Logie, R. H., Gilhooly, K. J., & Wynn, V. (1994). Counting on working memory in arithmetic problem solving. *Memory & Cognition, 22*, 395-410.

Logie, R. H., & Marchetti, C. (1991). Visuo-spatial working memory: Visual, spatial or central executive? In R. H. L. M. Denis (Ed.), *Mental images in human cognition* (pp. 105-115). Amsterdam: North Holland.

Logie, R. H., Saito, S., Morita, A., Varma, S., & Norris, D. (2016). Recalling visual serial order for verbal sequences. *Memory & Cognition, 44*, 590-607. doi:10.3758/s13421-015-0580-9

Logie, R. H., Zucco, G. M., & Baddeley, A. D. (1990). Interference with visual short-term memory. *Acta Psychologica, 75*, 55-74. doi:10.1016/0001-6918(90)90066-O

Logothetis, N. K. (2003). The underpinnings of the BOLD functional magnetic resonance imaging signal. *Journal of Neuroscience, 23*(10), 3963-3971.

Longoni, A. M., Richardson, J. T. E., & Aiello, A. (1993). Articulatory rehearsal and phonological storage in working memory. *Memory & Cognition, 21*, 11-22.

Loosli, S. V., Rahm, B., Unterrainer, J. M., Weiller, C., & Kaller, C. P. (2014). Developmental change in proactive interference across the life span: Evidence from two working memory tasks. *Developmental psychology, 50*, 1060-1072. doi:10.1037/a0035231

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279-281.

Lundqvist, M., Herman, P., & Lansner, A. (2011). Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. *Journal of Cognitive Neuroscience, 23*(10), 3008-3020. doi:10.1162/jocn_a_00029

Lundqvist, M., Rose, J., Herman, P., Brincat, S. L., Buschman, T. J., & Miller, E. K. (2016). Gamma and Beta Bursts Underlie Working Memory. *Neuron, 90*(1), 152-164. doi:10.1016/j.neuron.2016.02.028

Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience and Biobehavioral Reviews, 62*, 100-108. doi:10.1016/j.neubiorev.2016.01.003

Macken, B., Taylor, J. C., Kozlov, M. D., Hughes, R. W., & Jones, D. M. (2016). Memory as embodiment: The case of modality and serial short-term memory. *Cognition, 155*, 113-124.

Mackworth, J. F. (1962). Presentation rate and immediate memory. *Canadian Journal of Psychology, 16*, 42-47.

Mackworth, J. F. (1963). The duration of the visual image. *Canadian Journal of Psychology, 17*, 62-81.

Mackworth, J. F. (1964). Auditory short-term memory. *Canadian Journal of Psychology, 18*, 292-303.

Macnamara, B. N., Moore, A. B., & Conway, A. R. A. (2011). Phonological similarity effects in simple and complex span tasks. *Memory & Cognition, 39*, 1174-1186.

Madigan, S. A. (1971). Modality and recall order interactions in short-term memory for serial order. *Journal of Experimental Psychology, 87*(294-296).

Magimairaj, B. M., & Montgomery, J. W. (2012). Children's verbal working memory: Relative importance of storage, general processing speed, and domain-general controlled attention. *Acta Psychologica, 140*, 196-207. doi:10.1016/j.actpsy.2012.05.004

Majerus, S., Leclercq, A.-L., Grossman, A., Billard, C., Touzin, M., Van der Linden, M., & Poncelet, M. (2009). Serial order short-term memory capacities and specific language impairment: No evidence for a causal association. *Cortex, 45*, 708-720.

Majerus, S., & Van der Linden, M. (2003). Long-term memory effects on verbal short-term memory: A replication study. *British Journal of Developmental Psychology, 21*, 303-310.

Makovski, T. (2012). Are multiple visual short-term memory storages necessary to explain the retro-cue effect? *Psychonomic Bulletin & Review, 19*, 470-476.

Makovski, T., Sussman, R., & Jiang, Y. (2008). Orienting attention in visual working memory reduces interference from memory probes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 369-380.

Mandler, G. (1967). Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The Psychology of Learning and Motivation* (pp. 328–372). Oxford: Academic Press.

Manning, S. K., & Geenhut-Wertz, J. (1990). Visual and auditory modality and suffix effects in young and elderly adults. *Experimental Aging Research, 16*, 3-9.

Manoach, D. S., Schlaug, G., Siewert, B., Darby, D. G., Bly, B. M., Benfield, A., . . . Warach, S. (1997). Prefrontal cortex fMRI signal changes are correlated with working memory load. *NeuroReport, 8*, 545-549.

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition, 122*, 346-362.

Matthews, M. L., & Henderson, L. (1970). Fast presentation rates and the recall of item and order information. *Nature, 226*, 374-376.

Maybery, M. T., Parmentier, F. B. R., & Jones, D. M. (2002). Grouping of list items reflected in the timing of recall: implications for models of serial verbal memory. *Journal of Memory and Language, 47*, 360-385.

Maylor, E. A., & Henson, R. A. (2000). Aging and the Ranschburg effect: No evidence of reduced response suppression in old age. *Psychology and Aging, 15*, 657-670.

Maylor, E. A., Vousden, J. I., & Brown, G. D. A. (1999). Adult age differences in short-term memory for serial order: Data and a model. *Psychology & Aging, 14*, 572-594.

McCabe, J., & Hartman, M. (2008). Working memory for item and temporal information in younger and older adults. *Aging, Neuropsychology, and Cognition, 15*, 574-600. doi:10.1080/13825580801956217

McCormack, T., Brown, G. D. A., Vousden, J. I., & Henson, R. N. A. (2000). Children's serial recall errors: Implications for theories of short-term memory development. *Journal of Experimental Child Psychology, 76*, 222-252.

McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 46, pp. 155-200). San Diego: Academic Press.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: The time course of recognition. *Journal of Experimental Psychology: General, 118*, 346-373.

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 196-204.

McVay, J. C., & Kane, M. J. (2012). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *Journal of Experimental Psychology: General, 141*, 302-320.

Meier, M. E., & Kane, M. J. (2013). Working memory capacity and Stroop interference: Global versus local indices of executive control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 748-759. doi:10.1037/a0029200

Meier, M. E., Smeekens, B. A., Silvia, P. J., Kwapil, T. R., & Kane, M. J. (2018). Working memory capacity and the antisaccade task: A microanalytic-macroanalytic investigation of individual differences in goal activation and maintenance. *Journal of Experimental Psychology: Learning, Memory and Cognition, 44*, 68-84. doi:10.1037/xlm0000431

Meiser, T., & Klauer, K. C. (1999). Working memory and changing-state hypothesis. *Journal of Experimental Psychology: Learning, Memory & Cognition, 25*, 1272-1299.

Mercer, T. (2014). The loss of short-term visual representations over time: decay or temporal distinctiveness? *Journal of Experimental Psychology: Human Perception and Performance, 40*, 2281-2288. doi:10.1037/a0038141

Mercer, T., & Duffy, P. (2015). The loss of residual visual memories over the passage of time. *Quarterly Journal of Experimental Psychology, 68*, 242-248. doi:10.1080/17470218.2014.975256

Meudell, P. R. (1977). Effects of length of retention interval on proactive interference in short-term visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 3*, 264-269.

Miles, C., Jones, D. M., & Madden, C. A. (1991). Locus of the irrelevant speech effect in short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(3), 578-584.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J Neurosci, 16*(16), 5154-5167.

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

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart & Winston.

Miller, G. A., & Selfridge, J. A. (1950). Verbal context and the recall of meaningful material. *American Journal of Psychology, 63*, 176-185.

Miller, S., McCulloch, S., & Jarrold, C. (2015). The development of memory maintenance strategies: training cumulative rehearsal and interactive imagery in children aged between 5 and 9. *frontiers in Psychology, 6*. Retrieved from doi:10.3389/fpsyg.2015.00524

Miozzo, M., Petrov, A., Fischer-Baum, S., & Peressotti, F. (2016). Serial position encoding of signs. *Cognition, 154*, 69-80.

Miyake, A., & Shah, P. (1999). *Models of working memory. Mechanisms of active maintenance and executive control*. Cambridge: Cambridge University Press.
Mohr, H. M., Goebel, R., & Linden, D. E. (2006). Content- and task-specific dissociations of frontal activity during maintenance and manipulation in visual working memory. *Journal of Neuroscience, 26*(17), 4465-4471. doi:10.1523/JNEUROSCI.5232-05.2006
Mongillo, G., Barak, O., & Tsodyks, M. (2008). Synaptic theory of working memory. *Science, 319*(5869), 1543-1546. doi:10.1126/science.1150769
Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10*, 465-501.
Morey, C. C., & Bieler, M. (2013). Visual short-term memory always requires general attention. *Psychonomic Bulletin & Review, 20*, 163–170. doi:10.3758/s13423-012-0313-z
Morey, C. C., & Cowan, N. (2004). When visual and verbal memories compete: Evidence of cross-domain limits in working memory. *Psychonomic Bulletin & Review, 11*, 296-301.
Morey, C. C., & Cowan, N. (2005). When do visual and verbal memories conflict? The importance of working-memory load and retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 703-713.
Morey, C. C., Elliott, E. M., Wiggers, J., Eaves, S. D., Shelton, J. T., & Mall, J. T. (2012). Goal-neglect links Stroop interference with working memory capacity. *Acta Psychologica, 141*, 250-260. doi:10.1016/j.actpsy.2012.05.013
Morey, C. C., & Mall, J. T. (2012). Cross-domain interference costs during concurrent verbal and spatial serial memory tasks are asymmetric. *Quarterly Journal of Experimental Psychology, 65*, 1777-1797. doi:http://doi.org/10.1080/17470218.2012.668555
Morey, C. C., Mareva, S., Lelonkiewicz, J. R., & Chevalier, N. (2017). Gaze-based rehearsal in children under 7: A developmental investigation of eye movements during a serial spatial memory task. *Developmental Science*. doi:10.1111/desc.12559
Morey, C. C., & Miron, M. D. (2016). Spatial sequences, but not verbal sequences, are vulnerable to general interference during maintenance in working memory. *Journal of Experimental Psychology: Leaming, Memory, and Cognition, 42*, 1907–1918. doi:10.1037/xlm0000280
Morey, C. C., Morey, R. D., van der Reijden, M., & Holweg, M. (2013). Asymmetric cross-domain interference between two working memory tasks: Implications for models of working memory. *Journal of Memory and Language, 69*, 324-348. doi:http://doi.org/http://dx.doi.org/10.1016/j.jml.2013.04.004
Morin, C., Brown, G. D. A., & Lewandowsky, S. (2010). Temporal isolation effects in recognition and serial recall. *Memory & Cognition, 38*(849-859).
Morin, R. E., DeRosa, D. V., & Ulm, R. (1967). Short-term recognition memory for spatially isolated items. *Psychonomic Science, 9*, 617-618.
Mosse, E. K., & Jarrold, C. (2008). Hebb Learning, verbal short-term memory, and the acquisition of phonological forms in children. *Quarterly Journal of Experimental Psychology, 61*, 501-514.
Mottaghy, F. M., Gangitano, M., Sparing, R., Krause, B. J., & Pascual-Leone, A. (2002). Segregation of areas related to visual working memory in the prefrontal cortex revealed by rTMS. *Cereb Cortex, 12*(4), 369-375.
Muller, N. G., & Knight, R. T. (2006). The functional neuroanatomy of working memory: contributions of human brain lesion studies. *Neuroscience, 139*(1), 51-58. doi:10.1016/j.neuroscience.2005.09.018
Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*, 482-488.
Murdock, B. B. (1967). Auditory and visual stores in short term memory. *Acta Psychologica, 27*, 316-324.
Murdock, B. B. (1968a). Serial order effects in short-term memory. *Journal of Experimental Psychology Monograph Supplement, 76*, 1-15.
Murdock, B. B. (1968b). Serial order effects in short-term memory. *Journal of Experimental Psychology, 76*, 1-15.

Murdock, B. B., & Okada, R. (1970). Interresponse times in single-trial free recall. *Journal of Experimental Psychology, 86*, 263–267

Murdock, B. B., & vom Saal, W. (1967). Transpositions in short-term memory. *Journal of Experimental Psychology, 74*, 137-143.

Murdock, B. B., & Walker, K. D. (1969). Modality effects in free recall. *Journal of Verbal Learning and Verbal Behavior, 8*, 665-676.

Murray, A. M., Nobre, A. C., Clark, I. A., Cravo, A. M., & Stokes, M. G. (2013). Attention restores discrete items to visual short-term memory. *Psychological Science*. doi:1177/0956797612457782

Murray, D. J. (1965). Vocalization at presentation and immediate recall with varying presentation rates. *Quarterly Journal of Experimental Psychology, 17*, 47-56. doi:10.1080/17470216508416407

Murray, D. J. (1966). Effects of loudness and presentation-rate on auditory short-term memory. *Nature, 210*, 226-226. doi:10.1038/210226a0

Murray, D. J. (1968). Articulation and acoustic confusability in short-term memory. *Journal of Experimental Psychology, 78*, 679-684.

Murray, D. J., & Roberts, B. (1968). Visual and auditory presentation, presentation rate, and short-term memory in children. *British Journal of Psychology, 59*, 119-125.

Myerson, J., Hale, S., Rhee, S. H., & Jenkins, L. (1999). Selective interference with verbal and spatial working memory in young and older adults. *Journals of Gerontology: Psychological Sciences, 54B*, P161-P164.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition, 18*, 251-269.

Nairne, J. S. (1991). Positional uncertainty in long-term memory. *Memory & Cognition, 19*(4), 332-340.

Nairne, J. S., Ceo, D. A., & Reysen, M. B. (2007). The mnemonic effects of recall on immediate retention. *Memory & Cognition, 35*(1), 191-199.

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

Naveh-Benjamin, M., Cowan, N., Kilb, A., & Chen, Z. (2007). Age-related differences in immediate serial recall: Dissociating chunk formation and capacity. *Memory & Cognition, 35*, 724-737.

Neath, I. (1997). Modality, concreteness, and set-size effects in a free reconstruction of order task. *Memory & Cognition, 25*, 256-263.

Neath, I. (2000). Modeling the effects of irrelevant speech on memory. *Psychonomic Bulletin & Review, 7*, 403-423.

Neath, I., & Crowder, R. G. (1990). Schedules of presentation and temporal distinctiveness in human memory. *Journal of Experimental Psychology: Leaming, Memory, and Cognition, 16*(2), 316-327.

Neath, I., & Crowder, R. G. (1996). Distinctiveness and very short-term serial position effects. *Memory, 4*, 225-242.

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2013). A meta-analysis of executive components of working memory. *Cerebral Cortex, 23*(2), 264-282. doi:10.1093/cercor/bhs007

Newsome, R. N., Duarte, A., Pun, C., Smith, V. M., Ferber, S., & Barense, M. D. (2015). A retroactive spatial cue improved VSTM capacity in mild cognitive impairment and medial temporal lobe amnesia but not in healthy older adults. *Neuropsychologia, 77*, 148-157. doi:10.1016/j.neuropsychologia.2015.08.017

Ng, H. L. H., & Maybery, M. T. (2002). Grouping in short-term verbal memory: Is position coded temporally? *Quarterly Journal of Experimental Psychology, 55A*, 391-424.

Nicolson, R. (1981). The relationship between memory span and processing speed. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and Learning* (pp. 179-183). New York: NY: Plenum Press.

Nimmo, L. M., & Lewandowsky, S. (2005). From brief gaps to very long pauses: Temporal isolation does not benefit serial recall. *Psychonomic Bulletin & Review, 12*(6), 999-1004.

Nimmo, L. M., & Lewandowsky, S. (2006). Distinctiveness revisited: Unpredictable temporal isolation does not benefit short-term serial recall of heard or seen events. *Memory & Cognition, 34*, 1368-1375.

Nimmo, L. M., & Roodenrys, S. (2005). The phonological similarity effect in serial recognition. *Memory, 13*, 773-784.

Norman, D. (1966). Acquisition and retention in short-term memory. *Journal of Experimental Psychology, 72*, 369-381.

Nosofsky, R. M., & Donkin, C. (2016). Response-time evidence for mixed memory states in a sequential-presentation change-detection task. *Cognitive Psychology, 84*, 31-62. doi:10.1016/j.cogpsych.2015.11.001

Oberauer, K. (2003a). Selective attention to elements in working memory. *Experimental Psychology, 50*(4), 257-269.

Oberauer, K. (2003b). Understanding serial position curves in short-term recognition and recall. *Journal of Memory and Language, 49*(4), 469-483.

Oberauer, K. (2005). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology-General, 134*(3), 368-387.

Oberauer, K. (2006). Is the focus of attention in working memory expanded through practice? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(2), 197-214.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin, 142*, 758-799. doi:10.1037/bul0000046

Oberauer, K., & Hein, L. (2012). Attention to information in working memory. *Current Directions in Psychological Science, 21*, 164-169. doi:10.1177/0963721412444727

Oberauer, K., Jones, T., & Lewandowsky, S. (2015). The Hebb repetition effect in simple and complex memory span. *Memory & Cognition, 43*, 852-865. doi:10.3758/s13421-015-0512-8

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

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

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review, 115*(3), 544-576.

Oberauer, K., & Lewandowsky, S. (2013). Evidence against decay in verbal working memory. *Journal of Experimental Psychology: General, 142*, 380-411.

Oberauer, K., & Lewandowsky, S. (2016). Control of information in working memory: Encoding and removal of distractors in the complex-span paradigm. *Cognition, 156*, 106-128. doi:10.1016/j.cognition.2016.08.007

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review, 19*, 779-819.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence--their correlation and their relation: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*, 61-65.

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity - facets of a cognitive ability construct. *Personality and Individual Differences, 29*, 1017-1045.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence, 31*, 167-193.

Oberauer, K., Wendland, M., & Kliegl, R. (2003). Age differences in working memory - The roles of storage and selective access. *Memory & Cognition, 31*(4), 563-569.

Olson, I. R., Page, K., Moore, K. S., Chatterjee, A., & Verfaellie, M. (2006). Working memory for conjunctions relies on the medial temporal lobe. *Journal of Neuroscience, 26*(17), 4596-4601. doi:10.1523/JNEUROSCI.1923-05.2006

Osth, A. F., & Dennis, S. (2015). The fill-in effect in serial recall can be obscured by omission errors. *Journal of Experimental Psychology: Leaming, Memory, and Cognition, 41*, 1447-1455. doi:10.1037/xlm0000113

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Hum Brain Mapp, 25*(1), 46-59. doi:10.1002/hbm.20131

Öztekin, I., Güngör, N. Z., & Badre, D. (2012). Impact of aging on the dynamics of memory retrieval: a time-course analysis. *Journal of Memory and Language, 67*, 285-294. doi:10.1016/j.jml.2012.05.003

Öztekin, I., & McElree, B. (2010). Relationship between measures of working memory capacity and the time course of short-term memory retrieval and interference resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 383-397.

Page, M. P. A., Cumming, N., Norris, D., Hitch, G. J., & McNeill, A. M. (2006). Repetition learning in the immediate serial recall of visual and auditory materials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 716-733.

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review, 105*, 761-781.

Page, M. P. A., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical Transactions of the Royal Society of London, 364*, 3737-3753.

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 332-350.

Palmer, S. (2000). Working memory: A developmental study of phonological recoding. *Memory, 8*(3), 179–193.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology & Aging, 17*, 299-320.

Parmentier, F. B. R., Andrés, P., Elford, G., & Jones, D. M. (2006). Organization of visuo-spatial serial memory: interaction of temporal order with spatial and temporal grouping. *Psychological Research-Psychologische Forschung, 70*(3), 200-217. doi:10.1007/s00426-004-0212-7

Parmentier, F. B. R., King, S., & Dennis, I. (2006). Local temporal distinctiveness does not benefit auditory verbal and spatial serial recall. *Psychonomic Bulletin & Review, 13*(3), 458-465. doi:10.3758/bf03193870

Parmentier, F. B. R., & Maybery, M. T. (2008). Equivalent Effects of Grouping by Time, Voice, and Location on Response Timing in Verbal Serial Memory. *Journal of Experimental Psychology-Learning Memory and Cognition, 34*(6), 1349-1355. doi:10.1037/a0013258

Parmentier, F. B. R., Maybery, M. T., & Jones, D. M. (2004). Temporal grouping in auditory spatial serial memory. *Psychonomic Bulletin & Review, 11*(3), 501-507. doi:10.3758/bf03196602

Paulesu, E., Frith, C. D., & Frackowiak, R. S. (1993). The neural correlates of the verbal component of working memory. *Nature, 362*(6418), 342-345. doi:10.1038/362342a0

Pearson, B., Raškevičius, J., Bays, P. M., Pertzov, Y., & Husain, M. (2014). Working memory retrieval as a decision process. *Journal of Vision, 14*, 1-15. doi:10.1167/14.2.2

Peich, M.-C., Husain, M., & Bays, P. M. (2013). Age-related decline of precision and binding in visual working memory. *Psychology and Aging, 28*, 729-743. doi:10.1037/a0033236

Pelegrina, S., Lechuga, M. T., García-Madruga, J. A., Elosúa, M. R., Macizo, P., Carreiras, M., Fuentes, L. J., & Bajo, M. T. (2015). Normative data on the n-back task for children and young adolescents. *frontiers in Psychology, 6*. Retrieved from doi:10.3389/fpsyg.2015.01544

Penney, C. G. (1989). Modality effects and the structure of short-term and verbal memory. *Memory & Cognition, 17*, 398-422.

Pertzov, Y., Bays, P. M., Joseph, S., & Husain, M. (2013). Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology: Human Learning and Memory, 29*, 1224-1231. doi:10.1037/a0030947

Pertzov, Y., Miller, T. D., Gorgoraptis, N., Caine, D., Schott, J. M., Butler, C., & Husain, M. (2013). Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis. *Brain : a journal of neurology, 136*(Pt 8), 2474-2485. doi:10.1093/brain/awt129

Peteranderl, S., & Oberauer, K. (2017). Serial recall of colors: Two models of memory for serial order applied to continuous visual stimuli. *Memory & Cognition*. doi:10.3758/s13421-017-0741-0

Peterson, L. R., & Johnson, S. T. (1971). Some effects of minimizing articulation on short-term retention. *Journal of Verbal Learning & Verbal Behavior, 10*, 346-354.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58*, 193-198.

Phaf, R. H., & Wolters, G. (1993). Attentional shifts in maintenance rehearsal. *American Journal of Psychology, 106*, 353-382.

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Percept Psychophys, 16*(2), 283-290.

Pickering, S. J., Gathercole, S. E., & Peaker, S. M. (1998). Verbal and visuospatial short-term memory in children: Evidence for common and distinct mechanisms. *Memory & Cognition, 26*, 1117-1130.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472-491.

Plancher, G., & Barrouillet, P. (2013). Forgetting from working memory: Does novelty encoding matter? *Journal of Experimental Psychology: Learning, Memory & Cognition, 39*, 110-125.

Pollack, I., Johnson, L. B., & Knaff, P. R. (1959). Running memory span. *Journal of Experimental Psychology, 57*, 137-146.

Polyn, S. M., Kragel, J. E., McClurey, J. D., & Burke, J. F. (2016). *Altering the flow of mental time: A test of retrieved-context theory*. Unpublished Manuscript, retrieved from https://memory.psy.vanderbilt.edu/files/pubs/PolyEtal_sub.pdf

Portrat, S., Camos, V., & Barrouillet, P. (2009). Working memory in children: a time-constrained functioning similar to adults. *Journal of Experimental Child Psychology, 102*, 368-374.

Portrat, S., Guida, A., Phénix, T., & Lemaire, B. (2016). Promoting the experimental dialogue between working memory and chunking: Behavioral data and simulation. *Memory & Cognition, 44*, 420-434. doi:10.3758/s13421-015-0572-9

Postle, B. R., Berger, J. S., & D'Esposito, M. (1999). Functional neuroanatomical double dissociation of mnemonic and executive control processes contributing to working memory performance. *Proc Natl Acad Sci U S A, 96*(22), 12959-12964.

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology, 17*, 132-138.

Prinzmetal, W., Amiri, H., Allen, K., & Edwards, T. (1998). Phenomenology of attention: 1. Color, location, orientation, and spatial frequency. *J Experimental Psychology: Human Perception and Performance, 24*(1), 261-282.

Puckett, J. M., & Lawson, W. M. (1989). Absence of adult age differences in forgetting in the Brown-Peterson task. *Acta Psychologica, 72*, 159-175.

Puckett, J. M., & Stockburger, D. W. (1988). Absence of age-related proneness to short-term retroactive interference in the absence of rehearsal. *Psychology and Aging, 3*, 342-347.

Quinlan, J. A., Neath, I., & Surprenant, A. M. (2015). Positional uncertainty in the Brown-Peterson paradigm. *Canadian Journal of Experimental Psychology, 69*, 64-71. doi:10.1037/cep0000038

Rama, P., Poremba, A., Sala, J. B., Yee, L., Malloy, M., Mishkin, M., & Courtney, S. M. (2004). Dissociable functional cortical topographies for working memory maintenance of voice identity and location. *Cereb Cortex, 14*(7), 768-780. doi:10.1093/cercor/bhh037

Randall, J. G., Oswald, F. L., & Beier, M. E. (2014). Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin, 140*, 1411-1431. doi:10.1037/a0037428

Ranganath, C., & Blumenfeld, R. S. (2005). Doubts about double dissociations between short- and long-term memory. *Trends in Cognitive Sciences, 9*, 374-380. doi:10.1016/j.tics.2005.06.009

Ratcliff, R., & Murdock, B. (1976). Retrieval processes in recognition memory. *Psychological Review, 83*, 190-214.

Rattat, A.-C. (2010). Bidirectional interference between timing and concurrent memory processing in children. *Journal of Experimental Child Psychology, 106*, 145-162.

Redick, T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., . . . Engle, R. E. (2016). Cognitive predictors of a common multitasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of Experimental Psychology: General, 145*, 1473-1492. doi:10.1037/xge0000219

Rerko, L., & Oberauer, K. (2013). Focused, unfocused, and de-focused information in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1175-1196. doi:10.1037/a0031172

Rerko, L., Oberauer, K., & Lin, H.-Y. (2014). Spatially imprecise representations in working memory. *Quarterly Journal of Experimental Psychology, 67*, 3-15. doi:10.1080/17470218.2013.789543

Ricker, T. J., & Cowan, N. (2010). Loss of visual working memory within seconds: The combined use of refreshable and non-refreshable features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1355-1368.

Ricker, T. J., Spiegel, L. R., & Cowan, N. (2014). Time-based loss in visual short-term memory is from trace decay, not temporal distinctiveness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1510-1523. doi:10.1037/xlm0000018

Ricker, T. J., Vergauwe, E., & Cowan, N. (2016). Decay theory of immediate memory: From Brown (1958) to today (2014). *Quarterly Journal of Experimental Psychology, 69*, 1969-1995. doi:10.1080/17470218.2014.914546

Riggs, K. J., McTaggart, J., Simpson, A., & Freeman, R. P. J. (2006). Changes in the capacity of visual working memory in 5 to 10 year olds. *Journal of Experimental Child Psychology, 95*, 18-26.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology, 92*, 365-372.

Röer, J. P., Bell, R., Marsh, J. E., & Buchner, A. (2015). Age equivalence in auditory distraction by changing and deviant speech sounds. *Psychology and Aging, 30*, 849-855. doi:10.1037/pag0000055

Rohrer, D. (1996). On the relative and absolute strength of a memory trace. *Memory & Cognition, 24*, 188-201.

Rohrer, D., & Wixted, J. T. (1994). An analysis of latency and interresponse time in free recall. *Memory & Cognition, 22*, 511–524

Romo, R., Brody, C. D., Hernandez, A., & Lemus, L. (1999). Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature, 399*(6735), 470-473. doi:10.1038/20939

Roodenrys, S., Hulme, C., Alban, J., Ellis, A. W., & Brown, G. D. A. (1994). Effects of word frequency and age of acquisition on short-term memory span. *Memory & Cognition, 22*, 695-701.

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., . . . Eickhoff, S. B. (2012). Modelling neural correlates of working memory: a coordinate-based meta-analysis. *NeuroImage, 60*(1), 830-846. doi:10.1016/j.neuroimage.2011.11.050

Rouder, J. N., Speckman, P. I., Sun, D., & Morey, R. D. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225-237.

Rouleau, N., & Belleville, S. (1996). Irrelevant speech effect in aging: An assessment of inhibitory processes in working memory. *Journal of Gerontology: Psychologal Sciences, 51B*, P356-P363.

Ryan, J. (1969). Grouping and short-term memory: Different means and patterns of grouping. *Quarterly Journal of Experimental Psychology, 21*, 137-147.

Saito, S., Logie, R. H., Morita, A., & Law, A. S. (2008). Visual and phonological similarity effects in verbal immediate serial recall: A test with kanji materials. *Journal of Memory and Language, 59*, 1-17. doi:10.1016/j.jml.2008.01.004

Sakai, K., & Inui, T. (2002). A feature-segmentation model of short-term visual memory. *Perception, 31*, 579-589.

Sala, J. B., & Courtney, S. M. (2007). Binding of what and where during working memory maintenance. *Cortex, 43*(1), 5-21.

Sala, J. B., Rama, P., & Courtney, S. M. (2003). Functional topography of a distributed neural system for spatial and nonspatial information maintenance in working memory. *Neuropsychologia, 41*(3), 341-356.

Salamé, P., & Baddeley, A. D. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior, 21*, 150-164.

Salamé, P., & Baddeley, A. D. (1989). Effects of background music on phonological short-term memory. *Quarterly Journal of Experimental Psychology, 41A*, 107-122.

Salamé, P., & Baddeley, A. D. (1990). The effects of irrelevant speech on immediate free recall. *Bulletin of the Psychonomic Society, 28*, 540-542.

Sander, M. C., Werkle-Bergner, M., & Lindenberger, U. (2011a). Binding and strategic selection in working memory: a lifespan dissociation. *Psychology and Aging, 26*, 612-624. doi:10.1037/a0023055

Sander, M. C., Werkle-Bergner, M., & Lindenberger, U. (2011b). Contralateral delay activity reveals life-span age differences in top-down modulation of working memory contents. *Cerebral Cortex, 21*, 2809-2819. doi:10.1093/cercor/bhr076

Sanders, A. F., & Schroots, J. J. F. (1969). Cognitive categories and memory span III. Effects of similarity on recall. *Quarterly Journal of Experimental Psychology, 21*, 21-28.

Sanders, A. F., & Willemsen, E. M. (1978). The course of proactive interference in immediate probed recall. *Acta Psychologica, 42*, 133-144.

Sarigiannidis, I., Crickmore, G., & Astle, D. E. (2016). Developmental and individual differences in the precision of visuo-spatial memory. *Child Development, 39*, 1-12.

Schlittmeier, S. J., Weißgerber, T., Kerber, S., Fastl, H., & Hellbrück, J. (2012). Algorithmic modeling of the irrelevant sound effect (ISE) by the hearing sensation fluctuation strength. *Attention, Perception & Psychophysics, 74*, 194-203. doi:10.3758/s13414-011-0230-7

Schlittmeier, S. J., Weisz, N., & Bertrand, O. (2011). What characterizes changing-state speech in affecting short-term memory? An EEG study on the irrelevant sound effect. *Psychophysiology, 48*, 1669-1680.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry, 20*(1), 11-21.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General, 125*, 4-27. doi:http://doi.org/10.1037/0096-3445.125.1.4

Shahabi, S. R., Abad, F. J., & Colom, R. (2014). Short-term storage is a stable predictor of fluid intelligence whereas working memory capacity and executive function are not: A comprehensive study with Iranian schoolchildren. *Intelligence, 44*, 134-141. doi:10.1016/j.intell.2014.04.004

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*, 1248-1284. doi:10.1080/03640210802414826

Shimi, A., & Scerif, G. (2017). Towards an integrative model of visual short-term memory maintenance: Evidence from the effects of attentional control, load, decay, and their interactions in childhood. *Cognition, 169*, 61-83. doi:10.1016/j.cognition.2017.08.005

Shipstead, Z., & Engle, R. E. (2013). Interference within the focus of attention: Working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 277-289. doi:10.1037/a0028467

Shipstead, Z., Lindsey, D. R. B., Marshall, R. L., & Engle, R. E. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language, 72*, 116-141.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366.

Simon, H. A. (1974). How big is a chunk? *Science, 183*, 482-488.

Sims, C. R. (2015). The cost of misremembering: Inferring the loss function in visual working memory. *Journal of Vision, 15*(3).

Sligte, I. G., Scholte, H. S., & Lamme, V. A. F. (2008). Are there multiple visual short-term memory stores? *PLoS ONE, 3*. doi:10.1371/journal.pone.0001699

Smith, E. E., Jonides, J., & Koeppe, R. A. (1996). Dissociating verbal and spatial working memory using PET. *Cerebral Cortex, 6*(1), 11-20.

Smith, M. H., & Stearns, E. G. (1949). The influence of isolation on the learning of surrounding materials. *American Journal of Psychology, 62*(3), 369-381. doi:10.2307/1418278

Smyth, M. M., Hay, D. C., Hitch, G. J., & Horton, N. J. (2005). Serial position memory in the visual-spatial domain: Reconstructing sequences of unfamiliar faces. *Quarterly Journal of Experimental Psychology, 58A*, 909-930.

Smyth, M. M., & Pendleton, L. R. (1990). Space and movement in working memory. *Quarterly Journal of Experimental Psychology, 42A*, 291-304.

Smyth, M. M., & Scholey, K. A. (1994). Interference in immediate spatial memory. *Memory & Cognition, 22*, 1-13.

Smyth, M. M., & Scholey, K. A. (1996). Serial order in spatial immediate memory. *Quarterly Journal of Experimental Psychology, 49A*, 159-177.

Sörqvist, P. (2010). High working memory capacity attenuates the deviation effect but not the changing-state effect: Further support for the duplex-mechanism account of auditory distraction. *Memory & Cognition, 38*, 651–658.

Souza, A. S. (2016). No age deficits in the ability to use attention to improve visual working memory. *Psychology and Aging, 31*, 456-470. doi:10.1037/pag0000107

Souza, A. S., & Oberauer, K. (2015). Time-based forgetting in visual working memory reflects temporal distinctiveness, not decay. *Psychonomic Bulletin & Review, 22*, 156-162. doi:10.3758/s13423-014-0652-z

Souza, A. S., Rerko, L., Lin, H.-Y., & Oberauer, K. (2014). Focused attention improves working memory: Implications for flexible-resource and discrete-capacity models. *Attention, Perception, & Psychophysics, 76*, 2080-2102. doi:10.3758/s13414-014-0687-2

Souza, A. S., Rerko, L., & Oberauer, K. (2014). Unloading and reloading working memory: Attending to one item frees capacity. *Journal of Experimental Psychology: Human Perception and Performance, 40*, 1237-1256. doi:10.1037/a0036331

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs, 74 (Whole No. 498)*.

Sperling, G., & Speelman, R. G. (1970). Acoustic similarity and auditory short-term memory: Experiments and a model. *Models of human memory (New York: Academic press)*, 151-202.

Spitzer, S. (1976). The development of visual and auditory recall as a function of presentation and probe modalities, serial position, and series size. *Child Development, 47*, 767-778.

Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring latent visual working memory representations in human cortex. *Neuron, 91*, 694-707. doi:10.1016/j.neuron.2016.07.006

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153*, 652-654.

Sternberg, S. (1969). Memory scanning: Mental processes revealed by reaction-time experiments. *American Scientist, 57*, 421-457.

Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences, 19*(7), 394-405. doi:10.1016/j.tics.2015.05.004

Sugase-Miyamoto, Y., Liu, Z., Wiener, M. C., Optican, L. M., & Richmond, B. J. (2008). Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. *PLoS Computational Biology, 4*(5), e1000073. doi:10.1371/journal.pcbi.1000073

Surprenant, A. M., Kelley, M. R., Farley, L. A., & Neath, I. (2005). Fill-in and infill errors in order memory. *Memory, 13*, 267-273.

Szmalec, A., Duyck, W., Vandierendonck, A., Mata, A. B., & Page, M. P. A. (2009). The Hebb repetition effect as a laboratory analogue of novel word learning. *Quarterly Journal of Experimental Psychology, 62*, 435-443.

Szmalec, A., Vandierendonck, A., & Kemps, E. (2005). Response selection involves executive control: Evidence from the selective interference paradigm. *Memory & Cognition, 33*, 531-541.

Szmalec, A., Verbruggen, F., Kemps, E., & Vandierendonck, A. (2011). Control of interference during working memory updating. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 137-151.

Taconnat, L., Raz, N., Toczé, C., Bouazzaoui, D., Sauzéon, H., Fay, S., & Isingrini, M. (2009). Ageing and organisation strategies in free recall: The role of cognitive flexibility. *European Journal of Cognitive Psychology, 21*, 347-365. doi:10.1080/09541440802296413

Tam, H., Jarrold, C., Baddeley, A. D., & Sabatos-DeVito, M. (2010). The development of memory maintenance: Children's use of phonological rehearsal and attentional refreshment in working memory tasks. *Journal of Experimental Child Psychology, 107*, 306-324.

Tan, L., & Ward, G. (2007). Output order in immediate serial recall. *Memory & Cognition, 35*, 1093-1106.

Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review, 15*, 535-542.

Tehan, G., Hendry, L., & Kocinski, D. (2001). Word length and phonological similarity effects in simple, complex, and delayed serial recall tasks: Implications for working memory. *Memory, 9*, 333-348.

Tehan, G., & Tolan, G. A. (2005). Is spoken duration a sufficient explanation of the word length effect? *Memory, 13*, 372-379.

Thomas, J. G., Milner, H. R., & Haberlandt, K. F. (2003). Forward and backward recall: Different response time patterns, same retrieval order. *Psychological Science, 14*, 169-174

Thorn, A. S. C., & Frankish, C. (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 729-735.

Thorn, A. S. C., Gathercole, S. E., & Frankish, C. R. (2005). Redintegration and the benefits of long-term knowledge in verbal short-term memory: An evaluation of Schweickert's (1993) multinomial processing tree model. *Cognitive Psychology, 50*, 133-158.

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428*(6984), 751-754. doi:10.1038/nature02466nature02466

Todd, J. J., & Marois, R. (2005). Posterior parietal cortex activity predicts individual differences in visual short-term memory capacity. *Cognitive, Affective & Behavioral Neuroscience, 5*(2), 144-155.

Towse, J. N., Cowan, N., Hitch, G. J., & Horton, N. J. (2008). The recall of information from working memory. Insights from behavioural and chronometric perspectives. *Experimental Psychology, 55*, 371-383. doi:10.1027/1618-3169.55.6.371

Towse, J. N., Hitch, G. J., & Hutton, U. (2002). On the nature of the relationship between processing activity and item retention in children. *Journal of Experimental Child Psychology, 82*, 156-184.

Towse, J. N., Hitch, G. J., & Skeates, S. (1999). Developmental sensitivity to temporal grouping effects in short-term memory. *International Journal of Behavioral Development, 23*, 391-411.

Treiman, R., & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 145-152.

Tremblay, S., & Jones, D. M. (1999). Change of intensity fails to produce an irrelevant sound effect: Implications for the representation of unattended sound. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1005-1015.

Tremblay, S., Macken, W. J., & Jones, D. M. (2000). Elimination of the word length effect by irrelevant sound revisited. *Memory & Cognition, 28*, 841-846.

Tremblay, S., Macken, W. J., & Jones, D. M. (2001). The impact of broadband noise on serial memory: Changes in band-pass frequency increases disruption. *Memory, 9*, 323-331. doi:10.1080/09658210143000010

Tremblay, S., Parmentier, F. B. R., Guérard, K., Nicholls, A. P., & Jones, D. M. (2006). A spatial modality effect in serial memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*, 1208-1215.

Tremblay, S., Parmentier, F. B. R., Hodgetts, H., Hughes, R. W., & Jones, D. M. (2012). Disruption of verbal-spatial serial memory by extraneous air-traffic speech. *Journal of Applied Research in Memory and Cognition, 1*, 73-79.

Tresch, M. C., Sinnamon, H. M., & Seamon, J. G. (1993). Double dissociation of spatial and object visual memory: Evidence from selective interference in intact human subjects. *Neuropsychologia, 31*, 211-219.

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review, 69*, 344-354.

Tulving, E., & Arbuckle, T. Y. (1966). Input and output interference in short-term associative memory. *Journal of Experimental Psychology, 72*, 145-150.

Tulving, E., & Colotla, V. A. (1970). Free recall of trilingual lists. *Cognitive Psychology, 1*, 86-98.

Tulving, E., & Patkau, J. E. (1962). Concurrent effects of contextual constraint and word frequency on immediate recall and learning of verbal material. *Canadian Journal of Psychology, 16*, 83-95

Turcotte, J., Gagnon, S., & Poirier, M. (2005). The effect of old age on the learning of supraspan sequences. *Psychology and Aging, 20*, 251-260. doi:10.1037/0882-7974.20.2.251

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

Tzeng, O. J. L., & Hung, D. L. (1973). Intralist organization and subsequent free recalls. *Journal of Experimental Psychology, 98*(1), 119-124. doi:10.1037/h0034310

Unsworth, N., & Engle, R. W. (2006a). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language, 54*, 68-80. doi:10.1016/j.jml.2005.06.003

Unsworth, N., & Engle, R. W. (2006b). A temporal-contextual retrieval account of complex span: An analysis of errors. *Journal of Memory and Language, 54*, 346-362.

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

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

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

Unsworth, N., & Robison, M. K. (2016). The influence of lapses of attention on working memory capacity. *Memory & Cognition, 44*, 188-196. doi:0.3758/s13421-015-0560-0

Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory & Language, 62*, 392-406.

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2010). The contributions of primary and secondary memory to working memory capacity: An individual differences analysis of immediate free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 240-247.

Vallar, G., & Baddeley, A. D. (1982). Short-term forgetting and the articulatory loop. *Quarterly Journal of Experimental Psychology, 34A*, 53-60.

Van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psych Rev, 121*(1), 124-149.

van den Berg, R., & Ma, W. J. (2014). "Plateau"-related summary statistics are uninformative for comparing working memory models. *Attention, Perception & Psychophysics*. doi:10.3758/s13414-013-0618-7

Van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, 109*, 8780-8785.

van Lamsweerde, A. E., Beck, M. R., & Elliott, E. M. (2015). Retrieval from long-term memory reduces working memory representations for visual features and their bindings. *Memory & Cognition, 43*, 237-246. doi:10.3758/s13421-014-0468-0

Vandierendonck, A. (2015). Modality independence of order coding in working memory: Evidence from cross-modal order interference at recall. *Quarterly Journal of Experimental Psychology*. doi:10.1080/17470218.2015.1032987

Veltman, D. J., Rombouts, S. A. R. A., & Dolan, R. J. (2003). Maintenance versus manipulation in verbal working memory revisited: an fMRI study. *NeuroImage, 18*, 247-256. doi:10.1016/S1053-8119(02)00049-6

Vergauwe, E., Barrouillet, P., & Camos, V. (2009). 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, 35*, 1012-1028.

Vergauwe, E., Barrouillet, P., & Camos, V. (2010). Do mental processes share a domain-general resource? *Psychological Science, 21*, 384-390. doi:http://doi.org/10.1177/0956797610361340

Vergauwe, E., Dewaele, N., Langerock, N., & Barrouillet, P. (2012). Evidence for a central pool of general resources in working memory. *Journal of Cognitive Psychology, 24*, 359-366. doi:http://doi.org/10.1080/20445911.2011.640625

Vergauwe, E., Hardman, K. O., Rouder, J. N., Roemer, E., McAllaster, S., & Cowan, N. (2016). Searching for serial refreshing in working memory: Using response times to track the content of the focus of attention over time. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-016-1038-1

Vergauwe, E., Hartstra, E., Barrouillet, P., & Brass, M. (2015). Domain-general involvement of the posterior frontolateral cortex in Time-Based Resource-Sharing in working memory: An fMRI study. *NeuroImage, 115*, 104-116. doi:http://doi.org/104-116.10.1016/j.neuroimage.2015.04.059

Vergauwe, E., Langerock, N., & Barrouillet, P. (2014). Maintaining information in visual working memory: memory for bindings and memory for features are equally disrupted by increased attentional demands. *Canadian Journal of Experimental Psychology, 68*, 158-162. doi:http://doi.org/10.1037/cep0000025

Verhaeghen, P., & Basak, C. (2005). Ageing and switching of the focus of attention in working memory: Results from a modified N-back task. *Quarterly Journal of Experimental Psychology*.

Verhaeghen, P., Cerella, J., & Basak, C. (2004). A working memory workout: how to expand the focus of serial attention from one to four items in 10 hours or less. *Journal of Experimental Psychology: Learning, Memory & Cognition, 30*, 1322-1337.

Vitali, P., Minati, L., Chiarenza, G., Brugnolo, A., Girtler, N., Nobili, F., . . . Rodriguez, G. (2006). The Von Restorff effect in ageing and Alzheimer's disease. *Neurological Science, 27*, 166-172. doi:10.1007/s10072-006-0662-3

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature, 428*(6984), 748-751. doi:10.1038/nature02447

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature, 438*(7067), 500-503. doi:10.1038/nature04171

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 1436-1451.

von Restorff, H. (1933). Über die Wirkung von Bereichsbildungen im Spurenfeld. *Psychologische Forschung, 18*, 299-342.

Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory: A meta-analysis. *Cognitive, Affective, and Behavioral Neuroscience, 3*(4), 255-274.

Ward, G. (2002). A recency-based account of the list length effect in free recall. *Memory & Cognition, 30*, 885-892.
Ward, G., Avons, S. E., & Melling, L. (2005). Serial position curves in short-term memory: Functional equivalence across modalities. *Memory, 13*, 308-317.
Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relation between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1207-1241. doi:10.1037/a0020122
Warrington, E. K., & Baddeley, A. D. (1974). Amnesia and memory for visual location. *Neuropsychologia, 12*, 257-263.
Watkins, M. J. (1972). Locus of the modality effect in free recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 644-648.
Watkins, M. J., & Watkins, O. C. (1973). The postcategorical status of the modality effect in serial recall. *Journal of Experimental Psychology, 99*, 226-230. doi:10.1037/h0034651
Watkins, M. J., Watkins, O. C., & Crowder, R. G. (1974). The modality effect in free and serial recall as a function of phonological similarity. *Journal of Verbal Learning and Verbal Behavior, 13*, 430-447.
Welch, G. B., & Burnett, C. T. (1924). Is primacy a factor in association-formation? *American Journal of Psychology, 35*, 396-401. doi:10.2307/1414018
Wickelgren, W. A. (1964). Size of rehearsal group and short-term-memory. *Journal of Experimental Psychology, 68*(4), 413-419. doi:10.1037/h0043584
Wickelgren, W. A. (1965). Acoustic similarity and intrusion errors in short-term memory. *Journal of Experimental Psychology, 70*(1), 102-108.
Wickelgren, W. A. (1967). Rehearsal grouping and hierarchical organization of serial position cues in short-term memory. *Quarterly Journal of Experimental Psychology, 19*, 97-&
Wickelgren, W. A., Corbett, A. T., & Dosher, B. A. (1980). Priming and retrieval from short-term memory: a speed accuracy trade-off analysis. *Journal of Verbal Learning and Verbal Behavior, 19*, 387-404.
Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *frontiers in Psychology, 4*. Retrieved from doi:10.3389/fpsyg.2013.00433
Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*, 1120-1135.
Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin, 138*, 102-125. doi:10.1037/a0025715
Wingard, J. A. (1980). Life-span developmental changes in mnemonic organization: a multimethod analysis. *International Journal of Behavioral Development, 3*, 467-487.
Wolff, M. J., Ding, J., Myers, N. E., & Stokes, M. G. (2015). Revealing hidden states in visual working memory using electroencephalography. *frontiers in Systems Neuroscience, 9*. Retrieved from doi:10.3389/fnsys.2015.00123
Woods, D. L., Wyma, J. M., Herron, T. J., & Yund, E. W. (2016). An improved spatial span test of visuospatial memory. *Memory, 24*, 1142-1155. doi:10.1080/09658211.2015.1076849
Woodward, A. J., Macken, W. J., & Jones, D. M. (2008). Linguistic familiarity in short-term memory: A role for (co-)articulatory fluency? *Journal of Memory and Language, 58*, 48-65.
Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature, 440*(7080), 91-95. doi:nature0426210.1038/nature04262
Zeintl, M., & Kliegel, M. (2010). Proactive and coactive interference in age-related performance in a recognition-based operation span task. *Gerontology, 56*, 421-429. doi:10.1159/000237875
Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*, 23-35.

Table 1: Brief Descriptions of Experimental Paradigms for Studying Working Memory

Name	Description
Serial Recall (SR)	Reproduction of a sequentially presented list of items in the order of presentation
Free Recall (FR)	Reproduction of a list of items in free order
Probed Recall (PR)	Recall of items in response to a retrieval cue uniquely identifying that item (e.g., its ordinal list position, or its spatial location)
Reconstruction of order (ROO)	Reproduction of the order of presentation of a list of items by placing each item in its correct ordinal list position (e.g., by moving the item with the mouse into a spatial place-holder for its list position)
Complex Span (CS)	Presentation of a list of items is interleaved with brief episodes of a distractor processing task; at the end the items have to be recalled (usually in serial order)
Brown-Peterson (BP)	Recall of a short list of items after a retention interval filled with a distractor processing task
Change Detection (CD)	An array of objects with simple visual features (e.g., colors, orientations) is presented briefly; after a brief delay the entire array is presented again, and the person decides whether or not one feature has been changed. Sometimes only one object is presented at test in the location of one original array object.
Continuous reproduction (a.k.a. delayed estimation) (CR)	An array of objects with simple visual features (e.g., colors, orientations) is presented briefly. After a brief delay one object is marked (usually by its location), and the person reproduces its feature on a continuous response scale (e.g., selecting its color on a color wheel), enabling the measure of memory precision on a continuous scale
N-Back (NB)	A long series of stimuli is presented sequentially, and the person decides for each stimulus whether it matches the one presented *n* steps back.

Running memory span (RM)	A series of stimuli of unpredictable length is remembered, and when it stops, the person is asked to recall the last *n* list elements, or as many list elements as possible from the end of the list.
Item Recognition (IRec)	A sequentially presented list, or simultaneously presented array of items is remembered briefly, followed by a single probe; the person decides whether that probe was contained in the memory set.
Relational Recognition (a.k.a. Local Recognition) (RRec)	Each item of a memory set is presented in a unique location, or in a relation with another unique stimulus, such as color. At test, a probe is presented in one of the locations (or in conjunction with one of the unique stimuli), and the person decides whether the probe matches the memory item in that location (or with that unique stimulus).
Memory Updating (MU)	Presentation of a set of initial items (e.g., digits, spatial locations of objects) is followed by instructions to update individual items, either through transformation (e.g., adding or subtracting some value from a digit, or shifting an object to a new spatial location) or through replacement (e.g., presenting a new digit, or presenting an object in a new location).

Note: Each paradigm name is accompanied by the acronym that this paradigm is given in the reference table in Appendix C.