An easy way to improve scoring of memory span tasks: The edit distance, beyond "correct recall in the correct serial position"

Corentin Gonthier

An easy way to improve scoring of memory span tasks:

The edit distance, beyond "correct recall in the correct serial position"

Corentin Gonthier [1, 2]

University of Nantes

[1] Nantes Université, LPPL UR 4638, Nantes, France

[2] Institut Universitaire de France

Correspondence concerning this article should be addressed to Corentin Gonthier, Laboratoire de psychologie des Pays de la Loire (LPPL UR 4638), Nantes Université, 44000, Nantes, France

E-mail: corentin.gonthier@univ-nantes.fr

Word count : 7.906 words excluding title page, abstract, figures, tables, and references

I thank Nicolas Loye for his technical assistance in analyzing the data, Jean-Luc Roulin for his interest in discussing this topic, and Jason S. Tsukahara and colleagues for making their data freely available.

Abstract

For researchers and psychologists interested in estimating a subject's memory capacity, the current standard for scoring memory span tasks is the partial-credit method: subjects are credited with the number of stimuli that they manage to recall correctly in the correct serial position. A critical issue with this method, however, is that intrusions and omissions can radically change the scores depending on where they occur. For example, when recalling the sequence ABCDE, "ABCD" is worth 4 points but "BCDE" is worth 0 point. This paper presents an improved scoring method based on the edit distance, meaning the number of changes required to edit the recalled sequence into the target. Edit-distance scoring gives results close to partial-credit scoring, but without the corresponding vulnerability to positional shifts. A reanalysis of memory performance in two large datasets ($N = 1093$ and $N = 758$) confirms that in addition to being more logically consistent, edit-distance scoring demonstrates similar or better psychometric properties than partial-credit, with comparable validity, a small increase in reliability, and a substantial increase of test information (measurement precision in the context of item response theory). Test information was especially improved for harder items and for subjects with ability in the lower range, whose scores tend to be severely underestimated by partial-credit scoring. Code to compute edit distance scores with various software is made available at https://osf.io/wdb83/.

Keywords

Short-term memory ; Working memory ; Serial recall ; Scoring; Partial-credit scoring; Edit distance; Damerau-Levenshtein distance

Memory span or serial recall tasks are a staple of research and clinical practice in psychology. Span tasks tend to follow the same basic structure: a list of stimuli is presented in a certain order, and the subject is asked to recall these stimuli in the same order. In psychological research, these tasks are routinely used to measure individual differences in short-term memory or working memory capacity; in particular, a considerable corpus of literature has studied the determinants of individual differences in working memory (e.g. Brose et al., 2012; Cowan et al., 1998; Engle & Kane, 2004; Kane et al., 2004; Oberauer et al., 2003; Unsworth & Engle, 2007), and their relations with high-level cognition and intelligence (Ackerman et al., 2005). In clinical practice, memory span tasks are a part of everyday assessments, owing largely to their inclusion in Wechsler's intelligence scales (WISC and WAIS).

For both research and clinical purposes, it is therefore of utmost interest to obtain a precise and consistent measure of a subject's ability. For most practical applications relevant to task design, this needs to be a single score that summarizes memory performance, in a way that can be taken to reflect the "memory capacity" of the subject. This measure needs to quantify the degree of similarity between what was presented and what the subject remembered, while avoiding systematic underestimation or overestimation of performance, and it needs to be as neutral as possible regarding theory. Critically, the quality of a measure is contingent on scoring, and scoring is not quite straightforward in memory span tasks. There are multiple options to score recall performance, and the question of which method is more appropriate has continued to draw attention in recent years (e.g. Conway et al., 2005; Giofrè & Mammarella, 2014; Weitzner et al., 2021). Examples of the major possible scoring methods are given in Table 1, including the prevailing methods of all-or-nothing scoring and partial-credit scoring.

Table 1. *Overview of the main possible scoring methods for memory span tasks*

Scoring method	Examples of recalls for the target sequence : ABCDE						
	ABCDE	ABCD	BCDE	ABXCDE	ACBDE	EBACD	BADE
All-or-nothing scoring	5	0	0	0	0	0	0
Partial-credit scoring	5	4	0	2	3	1	0
Edit-distance scoring	5	4	4	4	4	2	3
Partial-credit ignoring order (lenient)	5	4	4	5	5	5	4
Longest correct sequence	5	4	4	3	2	2	2
Relative-order scoring	5	4	4	5	4	3	3
Input-output order correspondence	1	1	1	1	0.75	0.5	0.66

The overarching goal of this paper is to present two arguments supporting an improved method to obtain focal ability estimates: scoring based on the edit distance (Damerau, 1964; Levenshtein, 1966). The first argument in favor of this method relates to logical inconsistencies with the current standard of partial-credit scoring in the context of positional shifts (for an example, compare the second and third columns of Table 1). In the next sections, I review the move from the historical solution of all-or-nothing scoring to partial-credit scoring, the shortcomings of partial-credit scoring when it comes to positional shifts in a recalled sequence, and the ways in which edit-distance scoring solves this problem. The second argument in favor of this method is empirical: in the rest of the paper, I compare the psychometric properties of memory span scores obtained with partial-credit scoring and edit-distance scoring, to show that edit-distance scoring yields substantially more accurate measurement. Edit-distance scoring applies to all serial recall tasks, whether they measure short-term memory or working memory; the two constructs and the corresponding tasks are considered indiscriminately in the following sections.

From All-or-nothing scoring to Partial-credit scoring

The method of all-or-nothing scoring was the most widely used in early working memory studies (e.g. Case et al., 1979, 1982; Daneman & Carpenter, 1980; Waters & Caplan, 1996), and is still employed in many clinical tests, including the WISC and WAIS. Consider

a simple example: the subject is presented with the sequence ABCDE, and recalls the sequence ABCD. In all-or-nothing scoring, a point is given if the recalled sequence is perfectly identical to the target, and no point is given otherwise: the recall ABCD would be scored 0. The intuitive justification for this method is the idea that any individual has a "maximal span" corresponding to the maximal number of items they can remember at once; as long as a trial is below this threshold, the subject should be able to recall all items, and get full credit. There are a few variants of all-or-nothing scoring: the most common solution is to count the total number of sequences perfectly recalled during the task, but some studies have also used the highest set size that the subject is able to recall perfectly (Della Sala et al., 1995, 2010; Friedman & Miyake, 2005).

Despite its historical prevalence and its continued use in clinical tests, all-or-nothing scoring has crippling downsides. Conceptually, the "maximal span" of an individual depends on factors that can vary across different tasks or across testing sessions, which means it is not a stable construct (see Conway et al., 2005). Besides, memory span can also be viewed as the result of the allocation of a continuous pool resources (e.g. Barrouillet et al, 2004; Just & Carpenter, 1992), which does not necessarily translate into a span of discrete size (for an example with visual memory, see Ma et al., 2014). Pragmatically, all-or-nothing scoring has low discriminating power (see Ferguson, 1949; Thurlow, 1950). In other words, it can only produce a very limited range of different scores - only two possible scores (0 or 1) for a given trial, and 11 possible scores for a 10-trials task. This means that much information about individual differences is lost in the process: a subject who recalls ABCD certainly demonstrates better memory performance than one who recalls AB (see Conway et al., 2005), but both responses are scored 0. This loss of information is even more pronounced for longer list lengths, where few subjects can manage perfect recall (especially in working memory tasks; see Unsworth & Engle, 2007). As a consequence, all-or-nothing scoring has been

shown in the case of working memory tasks to yield scores with lower reliability (Conway et al., 2005), and lower correlations with other measures (Friedman & Miyake, 2005; Unsworth & Engle, 2007), than other alternatives.

Due to these issues, most researchers interested in individual differences have now turned to partial-credit scoring (Conway et al., 2005; Redick et al., 2012). In partial-credit scoring, performance within a given trial is scored based on "how many stimuli were correctly recalled in the correct serial position". This definition allows for two variants: partial-credit load scoring, which counts the *number* of stimuli correctly recalled in the correct serial position (for the target sequence ABCDE, the recall ABCD would be scored 4), and partial-credit unit scoring, which counts the *proportion* of stimuli correctly recalled in the correct serial position (for the target sequence ABCDE, the recall ABCD would be scored 0.80, i.e. 80% correct). Both variants are used in practice (Conway et al., 2005), and there is usually little difference between the two. The rest of the current article is based on load scoring, both for consistency with Tsukahara et al. (2020), and because it is a more general solution that can be used even when different subjects complete trials of different set sizes, as is the case in adaptive tasks (Gonthier et al., 2017).

Intuitively, partial-credit scoring contains the same basic information as all-or-nothing scoring, with improved discriminating power due to also incorporating information from trials that were not fully correct (Unsworth & Engle, 2007). For this reason, this method demonstrates better reliability than all-or-nothing scoring, and higher correlations with other measures (Conway et al., 2005; Friedman & Miyake, 2005). This can affect results to the extent that switching from all-or-nothing scoring to partial-credit scoring has encouraged rethinking of a major theoretical model of working memory (Unsworth & Engle, 2007). This method also has the advantage of simplicity: partial-credit scoring is easy to implement in any software, or even to calculate without any software for pen-and-paper clinical tests.

The Critical Issue of Partial-credit scoring with Positional Shifts

Despite representing a major improvement over all-or-nothing scoring and being the current standard for memory span tasks, partial-credit scoring does have a serious flaw that has gone largely unnoticed. Because it credits only stimuli recalled precisely in the correct serial position, partial-credit scoring is highly vulnerable to positional shifts in the recalled sequence due to intrusions and omissions. For example, XABCDE is almost a perfect recall of the target ABCDE, but would receive 0 point, since none of the stimuli is in the correct serial position (A is in second position instead of first, etc). The same is true for BCDE. In other words, an intrusion or an omission in the first serial position is enough to invalidate the whole sequence[1]. This is a severe issue, given that omissions and intrusions often make up the majority of errors in memory spans (Unsworth & Engle, 2006).

Even more problematic is the fact that this issue is not logically consistent, because the penalty for a positional shift depends on where the intrusion or omission occurs. For the target sequence ABCDE, a recall of XABCDE is worth 0 point, AXBCDE is worth 1 point, ABCXDE is worth 3 points, and ABCDEX is worth a full 5 out of 5 points - despite the fact that these four responses are logically identical: in each case, the subject has correctly recalled the target ABCDE with a single intrusion. Thus, different subjects recalling the same number of correct items in the same order can obtain completely different scores. (It is also noteworthy that the degree of inconsistency scales with set size: the longer the sequence, the more opportunities to make errors in different serial positions.) Of course, the same is true for omissions: BCDE is worth 0 point, ABDE is worth 2 points, and ABCD is worth 4 points,

[1] Note that forcing subjects to recall a sequence of the same length as the target (e.g. by typing answers into a fixed set of boxes) does not control the problem of omissions occurring in the first serial positions: for the target ABCDE, a subject can always recall BCDEX. Order transpositions, such as recalling BACDE, are less of a problem than omissions or intrusions because they do not invalidate the rest of the sequence - as long as the transposition results in swapping two adjacent items, which is usually the case in adults (Henson et al., 1996; Unsworth & Engle, 2006). Confusion errors, such as recalling ABXDE, pose no problem at all because they do not shift the position of other items in the recalled sequence.

although the subject has in each case correctly recalled four out of five stimuli in the correct order.

In short, partial-credit scoring penalizes intrusions and omissions much more severely if they occur at the beginning than at the end of the sequence. Note that this is an unintended side effect of the scoring method, not a reasoned design choice. Critically, this aspect of scoring is not neutral in terms of the assessed mechanisms, because recall of the first and last items of a sequence can reflect different psychological processes. In particular, the recall of items presented at the end of a sequence may proceed from short-term or primary memory, whereas items presented at the beginning of a sequence may be more likely to be displaced from primary memory, having to be retrieved from long-term or secondary memory. This can be true both for long lists in immediate free recall tasks (Atkinson & Shiffrin, 1968; Glanzer & Cunitz, 1966), where newer to-be-encoded items can displace older items (Davelaar et al., 2005); and for working memory tasks such as complex spans, where a distracting secondary task can capture attention and prevent continued maintenance of items in primary memory (Unsworth & Engle, 2007). Thus it could be said, provocatively, that *short-term memory* span tasks penalize forgetting information stored in *long-term memory* (an omission at the beginning of the sequence) to a greater extent than information stored in *short-term memory* (an omission at the end).

This vulnerability of partial-credit scoring to positional shifts would be enough to question its sensitivity to the correct psychological processes, but this issue is further compounded by the fact that it can interact with individual differences (Unsworth & Engle, 2006). Indeed, not all individuals make the same proportion of errors at the beginning and end of a list. As reported by Unsworth and Engle (2006), adults with a low working memory span tend to make proportionally more omission errors than other subjects, and these errors are much more often located towards the beginning of the sequence; when they make

intrusion errors, these are more often located in the first serial position; and when they make transposition errors, these subjects more often displace items by more than one position. All three features make them more likely to make severe positional shifts in the entire recalled sequence, which suggests that partial-credit scoring is more likely to misestimate performance for low-performing subjects.

More generally, there can also be meaningful between-groups differences in error patterns, at least in short-term memory tasks. For example, older adults tend to make more omissions towards the end of the list in immediate serial recall tasks (Maylor et al., 1999). Children also make comparatively more intrusion and omission errors than adults in serial recall, and when making order transposition errors, they tend to displace items by more than one position (McCormack et al., 2000); both can lead to large positional shifts that completely invalidate a recalled sequence. Lastly, and regardless of ability, serial position curves may also be subject to strategic variability: some subjects appear to prioritize recall of the first or last serial positions in immediate free recall tasks (Unsworth et al., 2011). Performance estimates would be less vulnerable to positional shifts for subjects who choose to emphasize the first serial positions, which means the scoring method favors subjects who choose to use one strategy over the other.

A Better Alternative to Partial-credit scoring: The Edit Distance Method

The fact that the predominant method of partial-credit scoring unwillingly rates performance in an inconsistent way depending on the position of errors during recall, that this can reflect different psychological processes, and that the extent of misestimation can interact with individual and group differences, is reason enough to search for an alternative scoring method.

The literature has explored several possibilities, some of which are illustrated in Table 1. A breakdown of the major scoring methods, and whether they meet reasonable

criteria for a general estimate of memory performance, is also given in Table 2. As can be seen, many of the alternatives are useful to provide insight into select mechanisms of performance, but are hardly defensible as general solutions to provide an ability estimate, mostly because they tend to systematically underestimate or overestimate performance.

Table 2. *Comparison of the major scoring methods*

Requirement	Edit-distance	Partial-credit	All-or-nothing	Lenient scoring	Longest-correct sequence	Relative order	Input-output order correspondence
Depends on correct recall	Yes	Yes	Yes	Yes	Yes	Yes	***No***
Includes information from partial recalls	Yes	Yes	***No***	Yes	Yes	Yes	Yes
Penalizes intrusions	Yes	Yes	Yes	***No***	Yes	***No***	***No***
Depends on correct order	Yes	Yes	Yes	***No***	Yes	Yes	Yes
Does not require strict match with serial position	Yes	***No***	***No***	Yes	Yes	Yes	Yes
Insensitive to serial position of errors	Yes	***No***	Yes	Yes	***No***	Yes	Yes

For example, methods that ignore recall order altogether (e.g. Case et al., 1979, 1982) give a perfect score to both ABCDE and EADCB, which omits relevant information regarding the subject's performance, and may be too generous for a serial recall task (this method has been called "lenient scoring"; Chen & Cowan, 2005). Crediting the longest sequence of consecutively correct items is highly vulnerable to a single error in the middle of the sequence. A more balanced solution is to count the number of correctly recalled items in the correct serial *order* (rather than *position*), for example by crediting a recalled item only if it was presented in a later serial position than the item recalled immediately before ("relative

order scoring", Drewnowski & Murdock, 1980; for other examples, see Addis & Kahana, 2004; Klein et al., 2005). However, this method does not penalize intrusions at all. Assessing only the proportion of correspondence between item order during presentation and at recall ("input-output order correspondence", Asch & Ebenholtz, 1962; for other examples, see Nairne et al., 1991; DeLosh & McDaniel, 1996; McDaniel et al., 1995) ignores the number of items actually recalled and can give higher scores to subjects recalling less items.

There does appear to be a better alternative to partial-credit scoring, however. This alternative relies on the edit distance, a method to quantify the dissimilarity between two sequences of characters. The edit distance has an intuitive definition, as the number of operations that are required to change (edit) the recalled sequence into the target sequence. For example, two operations are required to change the response BACD into the target sequence ABCDE (add E, move B by one position), which means the distance between target and response is two. Likewise, a single operation is required to change the response BCDE into the target sequence ABCDE (add A), which means the distance is one. In other words, the edit distance directly quantifies the number of errors made during recall, taking order into account but without imposing a strict match between each item and its expected serial position (see also Kalm & Norris, 2016).

The edit distance was initially developed in the field of computer science to analyze spelling errors (Damerau, 1964), and is routinely employed in other fields such as signal processing. Variants of the edit distance based on the idea of string alignment are prominently used in biology to compare the similarity of DNA or protein sequences (e.g. Needleman & Wunsch, 1970; Sellers, 1974). There are several types of edit distances, which differ in the operations that are allowed to edit the sequence and their weighting (for overviews, see Boytsov, 2011; Navarro, 2001; van der Loo, 2014). The solution retained for the present work, and which I recommend for scoring serial recall tasks, is the Damerau-

Levenshtein distance (Damerau, 1964; see also Levenshtein, 1966). Most other alternatives have limited applicability to serial recall tasks (e.g. the Hamming distance requires the recalled sequence to be the same length as the target; the longest common substring method does not allow for substitutions of characters in the sequence), require careful calibration for a particular dataset (e.g. the string alignment algorithm used by Mathy & Varré, 2013), have less desirable features (e.g. the Levenshtein distance used by Kalm et al., 2013, differs from the Damerau-Levenshtein distance by the fact that it does not allow for transposition of adjacent characters: this makes less intuitive sense, counting ABDCE as two errors), and/or make little difference in practice (in Dataset 1, the Levenshtein and Damerau-Levenshtein methods were correlated at $r = .98$, but scoring based on Damerau-Levenshtein provided 4% to 15% more test information across subtests).

Scoring Based on the Damerau-Levenshtein Edit Distance

The Damerau-Levenshtein distance allows four operations: insertion, deletion, substitution, and transposition of adjacent characters. These four operations fit nicely with the four major types of errors in a memory span task: omissions, intrusions, confusions, and order transposition of targets (see Henson, 1998; Unsworth & Engle, 2006). By default, these operations are not weighted: every insertion, deletion, substitution or transposition counts as one operation and increases the distance between target and response by one. Note that transposition is only performed between two adjacent characters, which means a target wrongly transposed by one serial position requires a single operation to edit (a swap with the adjacent response), and a target transposed by more than one serial position always requires two (a deletion at its current serial position and an insertion in the correct position). In other words, omissions, intrusions, confusions, and swaps between adjacent characters are all counted as one error.

The Damerau-Levenshtein edit distance can be easily converted into a scoring method (hereafter called edit-distance scoring) by subtracting the number of operations required to match the target, from the total sequence length. In other words, edit-distance scoring is just the number of stimuli in a trial (the maximum possible score), minus the number of errors made by the subject. For example, the target ABCDE is of length five and the response BACD requires two editions using the Damerau-Levenshtein distance, which yields a score of 5-2 = 3. Examples of edit-distance scoring for various types of errors are given in Table 3. As with partial-credit scoring, edit-distance scoring can also be scored in terms of proportion correct. This only requires adding a further step of dividing the result by the sequence length: BACD would thus be scored (5-2)/5 = 0.60.

Table 3. *Edit distance scores for various types of errors, and comparison with partial-credit scores*

Types of errors	Examples of recalls (target ABCDE)	Required editions for the response to match the target	Edit-distance scoring	Partial-credit scoring
No error	ABCDE	0 (No edition required)	(5-0) = 5	5
Omission errors	ABCD	1 (Add E)	(5-1) = 4	4
	BCDE	1 (Add A)	(5-1) = 4	0
	ACE	2 (Add B, D)	(5-2) = 3	1
Intrusion errors	ABCDEX	1 (Remove X)	(5-1) = 4	5
	XABCDE	1 (Remove X)	(5-1) = 4	0
	XABCDEY	2 (Remove X, Y)	(5-2) = 3	0
Confusion errors	XBCDE	1 (Change X into A)	(5-1) = 4	4
	XBYDE	2 (Change X into A, Y into C)	(5-2) = 3	3
Transposition errors	BACDE	1 (Move A by one position)	(5-1) = 4	3
	BCADE	2 (Remove A, insert A at the beginning)	(5-2) = 3	2
	BCDEA	2 (Remove A, insert A at the beginning)	(5-2) = 3	0
Further examples	BAXE	3 (Move A by one position, change X into C, add D)	(5-3) = 2	0
	EDCBA	4 (Change E into A, D into B, B into D, A into E)	(5-4) = 1	1
	VWXYZ	5 (Change all five letters)	(5-5) = 0	0

As can be seen, edit-distance scoring has a number of desirable properties. It yields a score that is on the same scale as partial-credit scoring, ranging between 0 (the whole sequence needs to be changed to match the target sequence[2]) and the number of stimuli in the trial (no changes are required), or between 0 and 1 when scoring as a proportion. Edit-distance scoring penalizes omissions, intrusions, confusions and transpositions to the same extent, and is also completely invariant to the position where these errors occur. This provides for a robust and consistent alternative to partial-credit scoring.

In general, edit distance scores can be expected to be similar to partial-credit scores, with both methods being based on the number of stimuli correctly recalled, taking serial position into account. The two methods will yield identical results when the errors are confusions, or intrusions or omissions in the last serial position; edit distance scores will always be higher when the errors are transpositions, or intrusions or omissions made earlier in the recalled sequence. As a consequence, the discrepancy between the two scoring methods will tend to be larger for subjects with lower performance (who make more errors) and for more difficult trials (where longer sequences increase the range of different positional shifts that can occur due to an intrusion or omission), where partial-credit scoring will tend to misestimate performance.

Empirical Test of the Psychometric Properties of Edit Distance Scoring

The edit distance appears to be a logical improvement on partial-credit, but it remains to be directly tested whether edit-distance scoring performs as well as partial-credit scoring. A handful of authors have used the edit distance or similar approaches to score human performance (especially a normalized version of the Levenshtein distance: Fonollosa et al.,

[2] This is true unless a subject makes more intrusion errors than there were stimuli in the to-be-remembered sequence. This can yield negative scores, which should not be allowed; fortunately this situation seems rare in actual datasets. I return to this point in the discussion.

2015; Kalm et al., 2013; Kalm & Norris, 2016; Norris et al., 2020; and a conceptually similar but more complex algorithm based on string alignment: Mathy & Varré, 2013), and at least one study found the edit-distance scoring to be more sensitive to learning than partial-credit scoring (Kalm & Norris, 2016). However, edit-distance scoring has never been used in a large dataset or in an established memory task, and there has never been a systematic investigation of its psychometric properties. This is the purpose of the empirical section of this study.

The next section reanalyzes two large memory span datasets to show that edit-distance scoring performs not only well, but better than the current standard of partial-credit scoring. Other methods such as all-or-nothing scoring, relative-order scoring or lenient scoring (see Table 1 and Table 2) are not usually recommended or not usually employed when designing memory span tasks, and were thus of less interest. A test of these other alternatives is detailed as supplemental materials available at https://osf.io/wdb83/. None of them had better all-around psychometric properties than edit-distance scoring.

The two datasets presented in the next section concern participants completing working memory span tasks. The first dataset was collected in the process of developing a shortened battery of complex span tasks, the Composite Complex Span (CCS; Gonthier et al., 2016). The second dataset was collected by Tsukahara and colleagues (2020) in the context of a study of the relation between working memory capacity and attention control, using working memory tasks developed by Draheim and colleagues (2018). It was more difficult to find suitable datasets concerning short-term memory than working memory, which reflects the large number of studies using span tasks to estimate working memory capacity (see e.g. Ackerman et al., 2005; Redick et al., 2012). However, the principles at play are the same for short-term memory than working memory, and conclusions are expected to generalize. An example with a smaller short-term memory dataset based on the forward digit span (N = 54;

Bosen & Barry, 2020; Bosen et al., 2021) is provided as supplemental materials at https://osf.io/wdb83/. The results with this short-term memory task were very similar to those of the two datasets included in the main text and led to the same conclusions.

Edit-distance scoring was expected to perform better than partial-credit scoring, especially due to improved precision in the lower range of scores and for more difficult trials. Classical test theory is not well-equipped to detect this type of difference, because it only assesses aggregate reliability at the test level: measurement error cannot be investigated at the item level and is considered constant across all ability levels, which can mask large differences for a particular range of participant ability. Moreover, rank-ordering of total scores is expected to be relatively similar with partial-credit scoring and edit-distance scoring, leading to similar reliability and convergent validity at the test level, which can mask large differences for some items and some participants. The results were therefore analyzed using item response theory (IRT; for overviews, see Hambleton & Jones, 1993; Embretson, 1996; for an example with complex span tasks, see Draheim et al., 2018). IRT models the relation between participant ability and predicted score on an item, for each item separately. This makes it possible to estimate the amount of information provided by a given item for a given level of ability, indicating the precision with which scores on this item reflect ability for participants of this level (see Fisher, 1925; Thissen, 2000; information is the inverse of the standard error of measurement, and can be converted into a point estimate of reliability for a particular ability level). This measure of information can also be computed for the whole test, which was of particular interest here.

Test of Edit-distance scoring in two Datasets

Dataset 1: The Composite Complex Span

Method

The task, stimuli and sample are described in detail in the original publication for the CCS (Gonthier et al., 2016); a brief summary is given here. The CCS is a battery of three complex span tasks: a reading span, a symmetry span, and an operation span (see Kane et al., 2004; Redick et al., 2012; Unsworth et al., 2005). The principle of complex span tasks is to present a sequence of simple problems to solve, alternating with stimuli to memorize. At the end of a sequence, subjects have to recall all to-be-remembered stimuli in the correct serial order. In the reading span task, subjects decide whether sentences are correct, and memorize digits; in the symmetry span task, they decide whether pictures are vertically symmetrical, and memorize spatial locations in a 4x4 grid; in the operation span, they decide whether math operations are correct, and memorize letters. The CCS was designed to obtain a domain-general estimate of working memory capacity by combining short versions of these three tasks. Set sizes range from 4 to 8 stimuli to remember for the reading span, from 3 to 6 for the symmetry span, and from 3 to 7 for the operation span (with one trial for the lowest and highest set sizes, and two trials for the others).

As part of the initial validation of the CCS, a sample of 1093 undergraduate students completed the task for course credit (mean age = 20.79, $SD = 4.61$, 142 males). 109 participants were excluded because they failed to perform adequately on the problem solving aspect of one or more of the three subtests (Unsworth & Engle, 2005; we excluded participants scoring in the bottom 5th percentile for accuracy on the problems of any subtest), leaving a dataset of $n = 984$ for analysis (for a total of 21648 trials). A subset of 303 participants completed the task on two occasions, allowing for an estimation of test-retest

reliability. Another subset of 405 participants also completed Raven's Advanced Progressive Matrices (APM; Raven, 1998), allowing for an estimation of concurrent validity.

Each trial of the CCS was scored using both edit-distance scoring (as illustrated in Table 3, and computed as the set size of the trial minus the Damerau-Levenshtein distance between the target and the participant's recall), and partial-credit load scoring (computed as the number of items recalled in the correct serial position). Total scores on each task were obtained by summing all trials. Total scores on each task were also averaged, after standardization, to yield a composite working memory capacity estimate (Gonthier et al., 2016). Damerau-Levenshtein distances were computed using *stringdist* (van der Loo, 2014) for R (R Core Team, 2021).

For each scoring method, we examined the distribution of scores, their internal consistency (computed using Cronbach's alpha for each task, and McDonald's omega total coefficient for the composite score; see Cronbach, 1951; Zinbarg et al., 2005), their test-retest reliability (computed as the bivariate correlation between test and retest), and their validity (computed as the bivariate correlation between the CCS and the APM). IRT analyses were performed using the package *mirt* (Chalmers, 2012) for *R* (R Core Team, 2021), using a generalized partial credit model estimating both difficulty and discrimination parameters.

Results

The two scoring methods yielded the same score on 75% of trials, with edit-distance scoring crediting higher scores on 25% of trials. The distributions of scores for the three subtests are represented in Figure 1. Edit-distance scoring yielded distributions comparable in shape to partial-credit scoring, but there were less scores in the lower ranges – as expected given that edit-distance scoring penalizes partly correct responses with a positional shift to a lesser extent than partial-credit scoring. The difference was especially visible for the reading span and operation span subtests; edit-distance scoring had less impact on scores on the

symmetry span, where set sizes were lower and where the serial position of an error made less of a difference.

As expected, switching from partial-credit to edit-distance scoring made more difference for subjects with low performance: average performance increased more for subjects scoring in the bottom quartile (+6.29 points in the reading span, +1.78 points in the symmetry span, +4.56 points in the operation span) than for subjects scoring in the top quartile (+1.93 points in the reading span, +0.54 points in the symmetry span, +1.40 points in the operation span). Likewise, edit-distance scoring increased performance to a greater extent for trials with a larger set size (e.g. for trials of length 8 in the reading span task, average performance increased from 4.06 out of 8 to 5.22 out of 8) than for shorter trials (e.g. for trials of length 3 in the reading span, average performance increased from 2.73 out of 3 to 2.78 out of 3).

Overall reliability was comparable for the two scoring methods, but slightly better for edit-distance scoring, as summarized in Table 4. Correlations between scores obtained with the two scoring methods were very high (r = .95 for the reading span, r = .98 for the symmetry span, r = .97 for the operation span, and r = .98 for composite scores). Concurrent validity with the RAPM was identical for the two scoring methods, r = .38, p < .001 for both composite scores.

The results of the IRT analysis are represented in Figure 2 and Table 5. At the test level, edit-distance scoring substantially increased test information, especially in the lower range of scores and especially for the reading span (46% increase) and operation span (32% increase). This improvement is represented in Figure 2. For the reading span, peak test information increased from 3.81 (corresponding to .74 reliability, computed as 1-1/information) to 5.16 (.81 reliability); for the symmetry span, peak information increased from 3.36 (.70 reliability) to 3.59 (.72 reliability); and for the operation span, peak

information increased from 4.75 (.79 reliability) to 6.59 (.85 reliability). At the item level, edit-distance scoring led to higher information for all trials of all tasks, except for the easiest trial of the symmetry span task. The fit of IRT models was generally good for both partial-credit and edit-distance scoring ($p > .02$ for all items and RMSEA <.026 for all items).

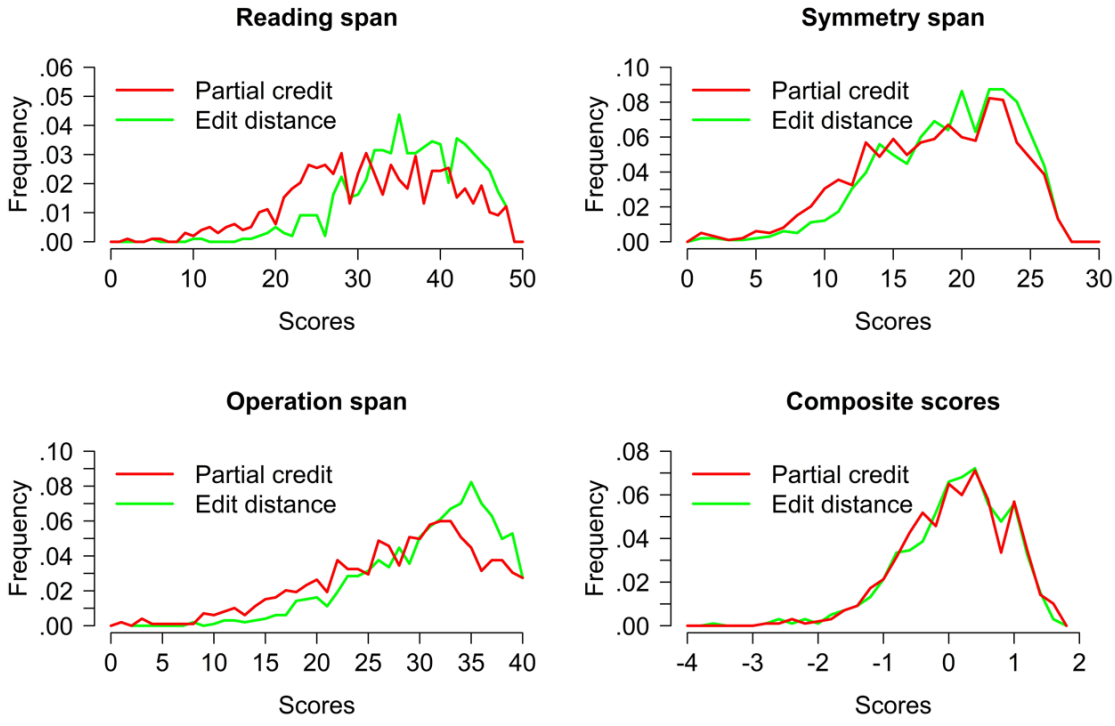

Figure 1. Distributions of scores for Dataset 1 as a function of scoring method.

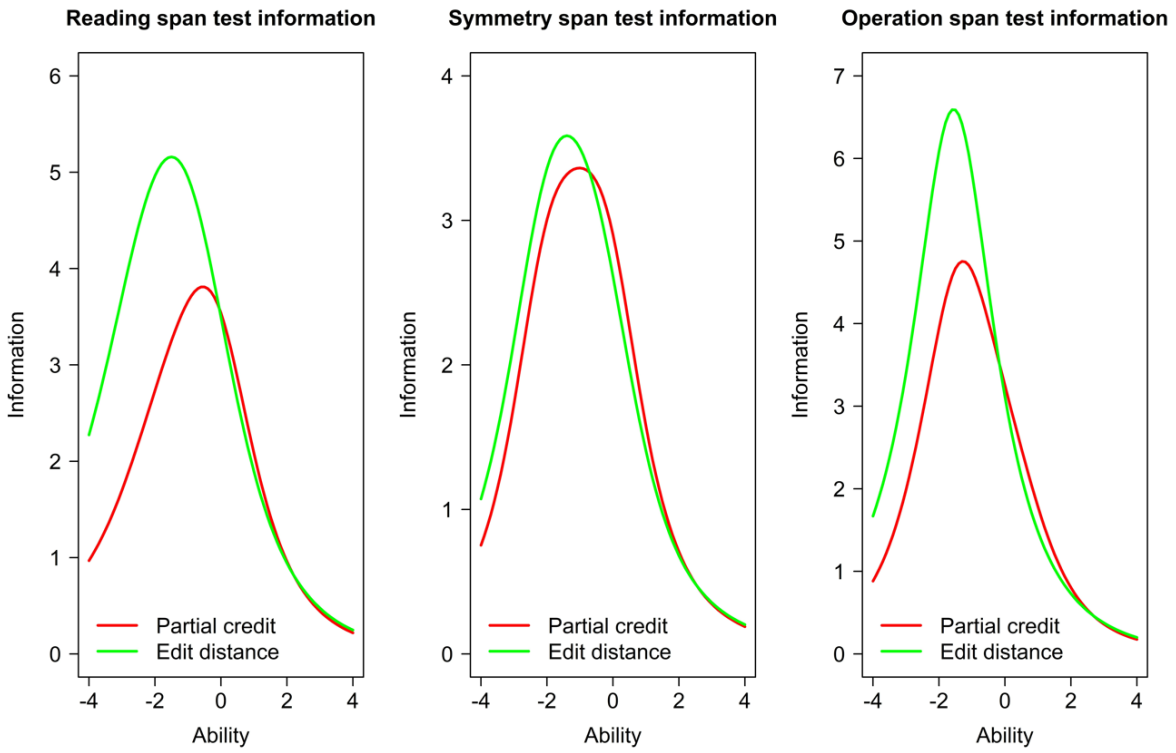

Figure 2. Test information for Dataset 1 as a function of scoring method.

Table 4. *Reliability for Datasets 1 and 2 as a function of scoring method*

Reliability	Scoring	Reading span	Symmetry span	Operation span	Composite score
Internal consistency (Dataset 1)	Partial-credit	.73	.70	.75	.85
	Edit distance	.75	.70	.77	.86
Test-retest stability (Dataset 1)	Partial-credit	.60	.66	.67	.75
	Edit distance	.61	.66	.67	.76
Internal consistency (Dataset 2)	Partial-credit	.82	.82	.87	.93
	Edit distance	.84	.84	.88	.94

Note. The reported coefficients are Cronbach's alphas (internal consistency for the three tasks), McDonald's omega total (internal consistency for the composite score), and correlation coefficients (test-retest stability).

Table 5. *Test and item information for Dataset 1 as a function of scoring method*

Reading span			Symmetry span			Operation span		
Trial set size	Partial-credit	Edit distance	Trial set size	Partial-credit	Edit distance	Trial set size	Partial-credit	Edit distance
4	1.73	2.50	3	2.14	2.05	3	1.77	2.27
5a	1.49	2.31	4a	2.68	2.94	4a	2.49	3.20
5b	2.00	2.64	4b	2.51	2.90	4b	2.18	3.09
6a	2.37	3.62	5a	2.92	2.94	5a	1.72	2.42
6b	2.69	3.59	5b	2.83	3.06	5b	2.32	3.10
7a	3.23	4.32	6	2.48	2.85	6a	2.80	4.09
7b	2.34	4.05				6b	2.43	3.01
8	2.11	3.15				7	3.20	3.69
Total	17.97	26.20	Total	15.55	16.73	Total	18.90	24.88

Dataset 2: Reanalysis of Tsukahara et al. (2020)

Method

Tsukahara and colleagues (2020) collected data using three complex span tasks: a rotation span, a symmetry span and an operation span. Details about the data collection are available in the original publication; the dataset itself is available at https://osf.io/hsqru/. The three tasks are described in detail by Draheim and colleagues (2018, Study 2). The symmetry span and operation span were very similar to the versions used for the CCS task (Gonthier et al., 2016), except that set sizes ranged from two to seven for the symmetry span and from three to eight for the operation span, with two trials per set size (except for set size eight in the operation span which had four trials). The rotation span is also a complex span task, with similar structure: subjects were required to decide whether rotated letters are mirrored reversed, and to memorize the length and direction of radial arrows (eight directions and two lengths, for sixteen possible stimuli). Set sizes for the rotation span ranged from two to seven, with two trials per set size. Additional data were collected using three reasoning tasks – Raven's matrices, letter sets, and number series. The scores on these three tasks were averaged after standardization to create a composite intelligence score, which was used to test for convergent validity.

A total of 758 participants were collected for this dataset (for a total of 29128 trials). All were native English speakers, between 18 and 35 years old (with an average of approximately 23 years old), 44% male, and compensated for their participation. All participants with complete data for the working memory tasks were kept for the current study. Data processing and analyses were identical to Dataset 1, with two exceptions: test-retest reliability was not available here, and IRT analyses constrained item parameters to be equal for the two trials with the same set size (as in Draheim et al., 2018).

Results

The distributions of scores for the three tasks are represented in Figure 3. Edit-distance scoring again yielded distributions comparable in shape to partial-credit scoring, but with less scores in the lower ranges. There was a difference between partial-credit and edit-distance scoring on 29% of trials. As in the case of Dataset 1, average scores increased to a greater extent for trials with a higher set size (e.g. from 4.05 to 5.01 for set size 8 of the operation span) than for shorter trials (e.g. from 2.61 to 2.65 for set size 3 of the same task), and they increased to a greater extent for subjects with lower performance (e.g. in the operation span, average increase of +9.70 points for subjects in the bottom quartile and +3.72 points for subjects in the top quartile).

Reliability, as estimated based on internal consistency, was again comparable for the two methods but slightly higher for edit-distance scoring (see Table 4). Correlations between scores computed using partial-credit and edit-distance scoring were again very high, for the rotation span (r = .96), the symmetry span (r = .97), the operation span (r = .96), and for composite scores (r = .98). The correlations with fluid intelligence were also comparable (r = .65 for partial-credit scoring and r = .64 for edit-distance scoring).

The results of the IRT analyses are represented in Figure 4 and Table 6. As in Dataset 1, edit-distance scoring substantially increased test information, especially in the

lower range of scores, and with particular improvement for the operation span (30% increase). This improvement is represented in Figure 4. For the rotation span, peak test information increased from 6.34 (.84 reliability) to 7.03 (.86 reliability); for the symmetry span, peak information increased from 6.20 (.84 reliability) to 7.75 (.87 reliability); and for the operation span, peak information increased from 10.00 (.90 reliability) to 15.36 (.93 reliability). At the item level, item information was systematically higher for edit-distance scoring than for partial-credit scoring, for trials of all set sizes. The fit of IRT models was excellent for both partial-credit and edit-distance scoring ($p > .070$ for all items and RMSEA <.02 for all items, except for one item in the rotation span; see also Draheim et al., 2018).

Table 6. *Test and item information for Dataset 2 as a function of scoring method*

Rotation span			Symmetry span			Operation span		
Trial set size	Partial-credit	Edit distance	Trial set size	Partial-credit	Edit distance	Trial set size	Partial-credit	Edit distance
2	1.66	1.93	2	2.01	2.32	3	2.07	2.73
3	2.00	2.30	3	2.31	2.66	4	2.45	3.13
4	2.77	3.35	4	2.27	2.79	5	2.09	3.11
5	3.16	3.49	5	3.00	3.41	6	2.66	3.83
6	2.96	3.32	6	2.88	3.20	7	3.09	3.99
7	3.25	3.46	7	2.95	3.46	8	3.22	3.87
Total	31.60	35.67	Total	30.84	35.68	Total	37.60	49.06

Note. Item information is the same for all trials of the same set size (due to item parameters being constrained equal, as in Draheim et al., 2018).

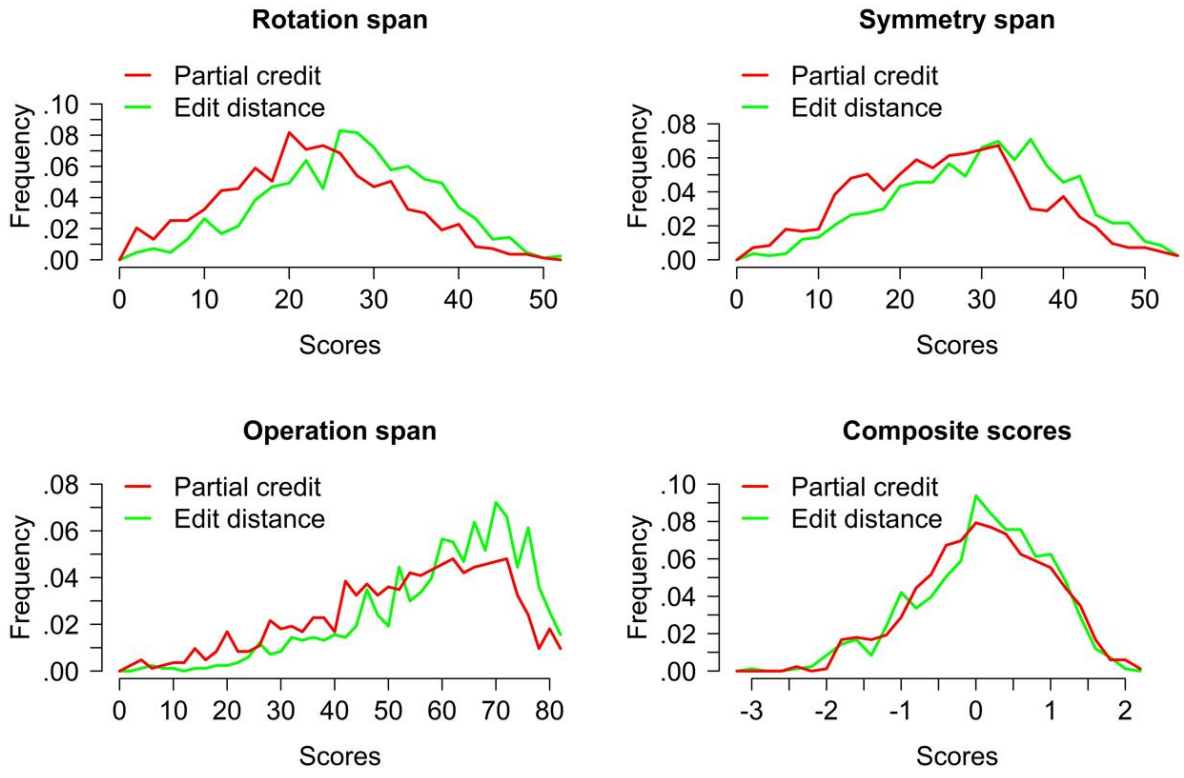

Figure 3. Distributions of scores for Dataset 2 as a function of scoring method.

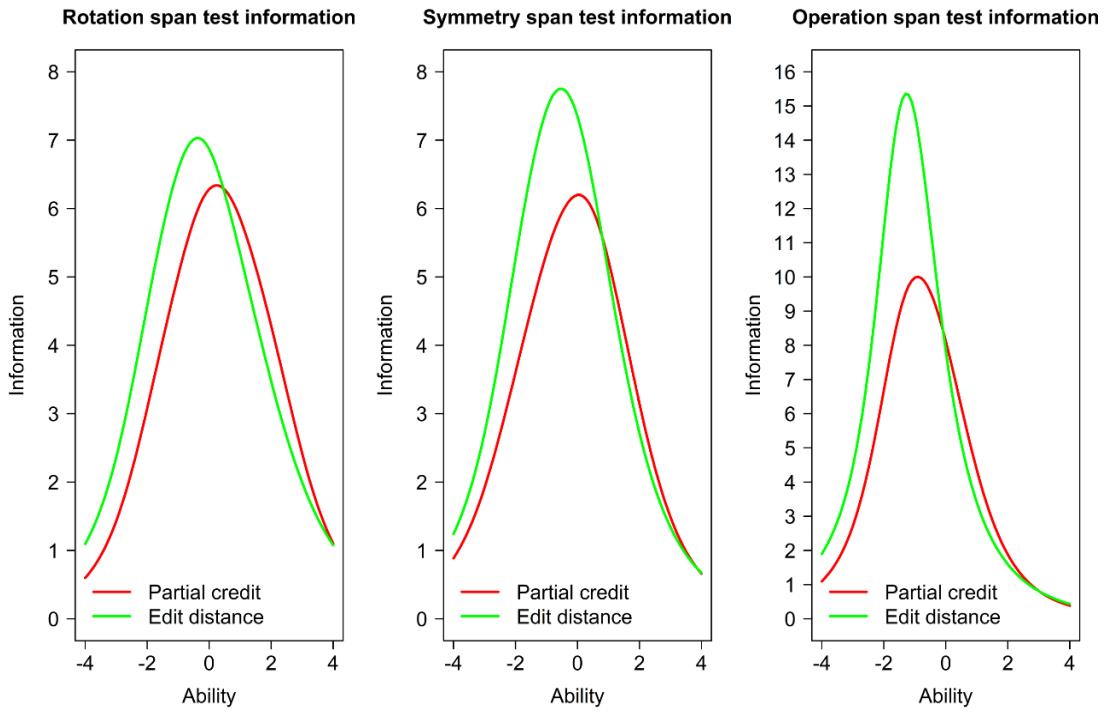

Figure 4. Test information for Dataset 2 as a function of scoring method.

Discussion

In two datasets using working memory span tasks (Gonthier et al., 2016; Tsukahara et al., 2020; and in a dataset using a forward digit span reported as supplemental material), edit-distance scoring demonstrated psychometric properties comparable to or better than traditional partial-credit scoring. Validity was comparable at the test level, and there was a small improvement in reliability for edit-distance scoring; but the major difference was a substantial increase in test information, which was manifest in almost all items. This increase was especially large for participants in the lower ranges of ability – participants who misrecall enough stimuli for edit-distance scoring to make a substantial difference when positional shifts occur. In other words, edit-distance scoring not only led to higher average scores for participants with low ability: it also led to more accurate estimates of these participants' ability.

The better precision of edit-distance scoring, as reflected in the two datasets reanalyzed here, combines with a more logically consistent scoring scheme. Contrary to partial-credit scoring, the edit distance always assigns the same score to the same number of recalled stimuli: it does not place a greater penalty on forgetting from long-term than short-term memory (Atkinson & Shiffrin, 1968; Davelaar et al., 2005; Glanzer & Cunitz, 1966; Unsworth & Engle, 2007), or strategically prioritizing final positions (Unsworth et al., 2011), or making intrusion and omission errors rather than confusion errors; and it will not interact with group differences (Maylor et al., 1999; McCormack et al., 2000) or individual differences (Unsworth & Engle, 2006) in the propensity to make positional shifts.

Contrary to the critical difference between all-or-nothing scoring and partial-credit scoring (e.g. Unsworth & Engle, 2007), this is more of an incremental improvement: in general, the results obtained with edit distance scores over a whole task should not radically differ from partial-credit scores, which means at the level of a sample, the correlations

between the total score and other tasks should be relatively unchanged. On the other hand, there can be a very large difference for a given trial - for high set sizes in particular, e.g. BCDEFGH for ABCDEFGH would be scored 7/8 with edit-distance scoring and 0/8 with partial-credit scoring -, which can have major impact on the score of an individual. Given that it is obviously of interest for both clinical practice and individual differences to obtain estimates with as much precision as possible at the level of each subject, or even at the level of a given trial (Draheim et al., 2018), there is little reason not to use a better method of scoring if it is available.

Although edit-distance scoring should always perform at least as well as partial-credit scoring, its benefits will be particularly visible in certain cases. Edit-distance scoring yields rank-ordering of subjects similar to partial-credit scoring on average, but partial-credit scoring can substantially underestimate the score of a given participant; therefore edit-distance scoring is especially recommended when the performance of a single participant is of interest (such as in clinical settings, job selection, etc). Edit-distance scoring provides more precise estimates for participants with a low level of ability, or a tendency to make large positional shifts (whereas the results will converge with partial-credit scoring when all participants have near-perfect recall); therefore edit-distance scoring can be particularly useful for samples with low working memory capacity (such as in research on working memory deficits, and developmental research).

In terms of task design, edit-distance scoring can make a large difference when using load scoring (Conway et al., 2005), as in the current study: the score on a trial is weighted by its set size, which means more difficult trials contribute more to the total, and edit-distance scoring will have more impact on the total score. When using an adaptive memory task (Gonthier et al., 2017), the difficulty of a trial is adjusted depending on the score on the previous trial: as a result, having a much higher score on a given trial can also impact the next

trials, and potentially change the final ability estimate to a large extent if the number of trials is limited or if this occurs towards the end of the task. A similar effect can occur when using a memory task discontinued based on performance (e.g. Alloway, 2007), where set sizes are presented in ascending order until the participant recalls less than a certain proportion correct. Lastly, edit-distance scoring can make a large difference for tasks designed with many difficult trials, given that performance on high set sizes is particularly susceptible to the position of errors and the positional shifts they create.

How to Implement of Edit-distance Scoring

Using edit-distance scoring appropriately invites three remarks regarding task design. First, investigators may want to limit the number of stimuli that can be recalled in a given trial to the set size of the trial (for example, preventing a subject from recalling more than 5 stimuli if only 5 stimuli were presented). This is because allowing subjects to recall more stimuli than were actually presented can yield edit distance scores below zero (changing DEFG into the target ABC requires four operations, which means a score of 3-4 = -1). Negative scores could alternatively be rounded to zero. In any case, negative scores should not be allowed: it would be inconsistent with prior literature, it would make descriptive statistics difficult to interpret, and it would penalize to different extents subjects who completely forget the sequence (for the target ABC, it does not seem justified to attribute better memory to a subject recalling XXX, scored 0, than to a subject recalling XXXXX, scored -2). Note that observing negative scores may be rare in adult studies: they occurred on 0.89% of trials in Dataset 2 (where they were rounded to zero; negative trials did not occur at all for Dataset 1, where recall was limited to the length of the target sequence). However, it might conceivably happen more frequently in populations with lower ability.

Second, in applications where a normal distribution of ability is particularly desirable, investigators may want to increase the difficulty of the task by including more trials with high

set sizes. This is because edit-distance scores will be higher than partial-credit scores – not because they are more lax, but because partial-credit scores underestimate the actual performance of low-ability subjects who make positional shifts. Being more sensitive to partial learning (memory for part of the items despite the lack of strict item-position associations; see Kalm & Norris, 2016), edit-distance scoring will thus yield higher scores for subjects in the lower range of ability. As a result, the distribution of total scores will tend to be shifted to the right, possibly with left-side skewness in some cases (see Figures 1 and 3), unless more difficult trials are added to the task. This may not be necessary for samples with low average ability, but it could be useful for samples of young adults with high performance, as was the case in the two datasets reported here.

Third, using edit-distance scoring can simplify task design, by removing the need for the option to indicate a blank in the recalled sequence. Span tasks usually allow subjects to indicate a skipped item in the sequence by giving a "?" response (e.g. AB?DE; for examples, see Chen & Cowan, 2005; Unsworth et al., 2005). With partial-credit scoring, this is necessary because marking an omission allows subjects to keep the end of the recalled sequence in the correct serial positions, and to get credit for these items. However, instructions regarding how to indicate skipped items are "both awkward and unreliable" (Klein et al., 2005); the point of doing so is not necessarily understood by the subjects (to whom the scoring method is not explained), and as a result, it tends to be rarely used (subjects reported a "?" in 4.29% of trials in Dataset 1). Edit-distance scoring removes the need for this option altogether: the same score is given to responses AB?DE and ABDE.

Limitations and Possible Extensions

With no obvious downsides, edit-distance scoring is a clearly better alternative to partial-credit scoring (and also performs better than other methods, as discussed in the supplemental materials at https://osf.io/wdb83/). This complements the list of recent advances

in the measurement of memory spans, along with the introduction of shortened domain-general task batteries (Foster et al., 2015; Gonthier et al., 2016; Oswald et al., 2015), adaptive tasks (Gonthier et al., 2017), and IRT modeling (Draheim et al., 2018). In contrast to partial-credit scoring, the edit distance represents an exhaustive solution to the general problem of quantifying the discrepancy between two series of items, which has driven its adoption by other fields such as biology and computer sciences (e.g. Damerau, 1964; Levenshtein, 1966; Needleman & Wunsch, 1970; Sellers, 1974). It is therefore unlikely that a better general solution for scoring can be found.

In fact, the main limitation of edit-distance scoring may be its slightly greater computational complexity. Implementing the Damerau-Levenshtein distance is somewhat more involved than the simple expression required to compute partial-credit scores. The algorithm is however well-known, and it can be summarized in under 30 lines of code. Given that the code for edit-distance scoring is readily available for R, Python and VBA and can be automatically computed with an Excel macro (see https://osf.io/wdb83/), this should not be a major obstacle for researchers interested in memory span tasks. The edit distance is also comparatively easy to calculate mentally for use in clinical settings (see Table 3; mental computation of the edit distance in fact tends to be easier than its algorithmic implementation).

Note that edit-distance scoring was designed to obtain ability estimates for memory, as unbiased as possible, in situations where a single score is required to summarize performance. This makes up a large share of individual differences studies (e.g. searching for correlations between working memory capacity and other constructs; Ackerman et al., 2005) and applied settings (e.g. clinical practice, job selection, etc). For these situations, edit-distance scoring performs better than partial-credit scoring and there does not seem to be a better general alternative. In other cases, particularly in fundamental research on mechanisms

of memory, it may be helpful to consider other theory-driven alternatives and to combine multiple scoring methods. For example, a study specifically interested in the binding of items to serial positions, or in separating memory for item information from memory for serial order (e.g. Majerus et al., 2006), may be better served by combining a form of scoring that takes strict serial position into account (such as partial-credit scoring) with a form of scoring that does not (such as lenient scoring, which appears to provide reasonable results; see Supplemental materials at https://osf.io/wdb83/), and interpreting the two concurrently (for an example, see Ward et al., 2010).

By the same logic, a further refinement of edit-distance scoring could be imagined. Computation of the Damerau-Levenshtein distance, and by extension edit-distance scoring, makes it possible to assign different weights to different types of errors. For example, it is possible to give less weight (penalize recall to a lesser extent) to transposition errors than to omission errors. Another reasonable possibility would be to count all transpositions as one error, even when characters are transposed by more than one serial position. Taking advantage of this possibility of differential weighting of errors would require a better model of the role and meaning of different types of errors than is available at present. The majority of studies to date have focused on total scores; research interested in patterns of errors in serial recall tasks (such as Unsworth & Engle, 2006) provides an important window into the cognitive processes at play, which could potentially be leveraged to obtain better scoring of recall performance.

Open Practices Statement

Code to compute edit distance scores with various software is available at https://osf.io/wdb83/. The data for Dataset 2 are available at https://osf.io/hsqru/, courtesy of Tsukahara et al. (2020).

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working Memory and Intelligence: The Same or Different Constructs? *Psychological Bulletin, 131*(1), 30-60. https://doi.org/10.1037/0033-2909.131.1.30

Addis, K. M., & Kahana, M. J. (2004). Decomposing serial learning: What is missing from the learning curve? *Psychonomic Bulletin & Review*, *11*(1), 118–124. https://doi.org/10.3758/BF03206470

Alloway, T. P. (2007). *Automated Working Memory Assessment (AWMA)*. Harcourt Assessment.

Asch, S. E., & Ebenholtz, S. M. (1962). The process of free recall: Evidence for non-associative factors in acquisition and retention. *The Journal of Psychology: Interdisciplinary and Applied*, *54*(1), 3–31. https://doi.org/10.1080/00223980.1962.9713093

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence and J. T. Spence (Eds.), *Psychology of learning and motivation, 2* (pp. 89-195). Academic Press. https://doi.org/10.1016/S0079-7421(08)60422-3

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Bosen, A. K., & Barry, M. F. (2020). Serial Recall Predicts Vocoded Sentence Recognition Across Spectral Resolutions. *Journal of Speech, Language, and Hearing Research : JSLHR*, *63*(4), 1282-1298. https://doi.org/10.1044/2020_jslhr-19-00319

Bosen, A. K., Sevich, V. A., & Cannon, S. A. (2021). Forward digit span and word familiarity do not correlate with differences in speech recognition in individuals with cochlear implants after accounting for auditory resolution. *Journal of Speech, Language, and Hearing Research : JSLHR*, *64*(8), 3330-3342. https://doi.org/10.1044/2021_jslhr-20-00574

Boytsov, L. (2011). Indexing methods for approximate dictionary searching: Comparative analysis. *ACM Journal of Experimental Algorithmics*, *16*(1), 1-91. https://doi.org/10.1145/1963190.1963191

Brose, A., Schmiedek, F., Lövdén, M., & Lindenberger, U. (2012). Daily variability in working memory is coupled with negative affect: The role of attention and motivation. *Emotion*, *12*(3), 605-617. https://doi.org/10.1037/a0024436

Case, R., Kurland, M., & Daneman, M. (1979, March). *Operational efficiency and the growth of M-space* [Paper presentation]. Society for Research in Child Development, San Francisco.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology, 33*(3), 386–404. https://doi.org/10.1016/0022-0965(82)90054-6

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Charter, R. A., & Feldt, L. S. (1996). Testing the equality of two alpha coefficients. *Perceptual and Motor Skills*, *82*(3, Pt 1), 763–768. https://doi.org/10.2466/pms.1996.82.3.763

Chen, Z., & Cowan, N. (2005). Chunk Limits and Length Limits in Immediate Recall: A Reconciliation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1235–1249. https://doi.org/10.1037/0278-7393.31.6.1235

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks : A methodological review and user's guide. *Psychonomic bulletin & review*, *12*(5), 769-786. https://doi.org/10.3758/BF03196772

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*(2), 141–160. https://doi.org/10.1037/0096-3445.127.2.141

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the Association for Computing Machinery*, *7*(3), 171–176. https://doi.org/10.1145/363958.363994

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450-466. https://doi.org/10.1016/s0022-5371(80)90312-6

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review*, *112*(1), 3–42. https://doi.org/10.1037/0033-295X.112.1.3

Della Sala, S., Baddeley, A., Papagno, C., & Spinnler, H. (1995). Dual-task paradigm: A means to examine the central executive. In J. Grafman, K. J. Holyoak, & F. Boller

(Eds.), *Structure and functions of the human prefrontal cortex*. (Vol. 769, pp. 161–171). New York Academy of Sciences.

Della Sala, S., Foley, J. A., Beschin, N., Allerhand, M., & Logie, R. H. (2010). Assessing dual-task performance using a paper-and-pencil test: Normative data. *Archives of Clinical Neuropsychology, 25*(5), 410–419. https://doi.org/10.1093/arclin/acq039

DeLosh, E. L., & McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1136–1146. https://doi.org/10.1037/0278-7393.22.5.1136

Draheim, C., Harrison, T. L., Embretson, S. E., & Engle, R. W. (2018). What item response theory can tell us about the complex span tasks. *Psychological Assessment, 30*(1), 116–129. https://doi.org/10.1037/pas0000444

Drewnowski, A., & Murdock, B. B. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(3), 319–332. https://doi.org/10.1037/0278-7393.6.3.319

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*, 341–349. https://doi.org/10.1037/1040-3590.8.4.341

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. *Psychology of Learning & Motivation*(44), 145.

Ferguson, G. A. (1949). On the theory of test discrimination. *Psychometrika*, *14*, 61–68. https://doi.org/10.1007/BF02290141

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700-725. https://doi.org/10.1017/S0305004100009580

Fonollosa, J., Neftci, E., & Rabinovich, M. (2015). Learning of chunking sequences in cognition and behavior. *PLoS Computational Biology*, *11*(11). https://doi.org/10.1371/journal.pcbi.1004592

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2015). Shortened complex span tasks can reliably easure working memory capacity. *Memory & Cognition*, *43*(2), 226-236. https://doi.org/10.3758/s13421-014-0461-7

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*(4), 581–590. https://doi.org/10.3758/BF03192728

Giofrè, D., & Mammarella, I. C. (2014). The relationship between working memory and intelligence in children: Is the scoring procedure important? *Intelligence, 46*, 300-310. https://doi.org/10.1016/j.intell.2014.08.001

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning & Verbal Behavior, 5*(4), 351–360. https://doi.org/10.1016/S0022-5371(66)80044-0

Gonthier, C., Aubry, A., & Bourdin, B. (2017). Measuring working memory capacity in children using adaptive tasks: Example validation of an adaptive complex span. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0916-4

Gonthier, C., Thomassin, N., & Roulin, J.-L. (2016). The Composite Complex Span : French validation of a short working memory task. *Behavior Research Methods*, *48*(1), 233-242. https://doi.org/10.3758/s13428-015-0566-3

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

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. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *49A*(1), 80–115. https://doi.org/10.1080/027249896392810

Henson, R. N. A. (1998). Short-term memory for serial order: The Start–End Model. *Cognitive Psychology*, *36*(2), 73–137. https://doi.org/10.1006/cogp.1998.0685

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

Kalm, K., Davis, M. H., & Norris, D. (2013). Individual sequence representations in the medial temporal lobe. *Journal of Cognitive Neuroscience*, *25*(7), 1111–1121. https://doi.org/10.1162/jocn_a_00378

Kalm, K., & Norris, D. (2016). Recall is not necessary for verbal sequence learning. *Memory & Cognition*, *44*, 104–113. https://doi.org/10.3758/s13421-015-0544-0

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. https://doi.org/10.1037/0096-3445.133.2.189

Klein, K. A., Addis, K. M., & Kahana, M. J. (2005). A comparative analysis of serial and free recall. *Memory & Cognition*, *33*(5), 833–839. https://doi.org/10.3758/BF03193078

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

Majerus, S., Poncelet, M., Greffe, C., & Van der Linden, M. (2006). Relations between vocabulary development and verbal short-term memory: The relative importance of short-term memory for serial order and item information. *Journal of Experimental Child Psychology*, *93*(2), 95–119. https://doi.org/10.1016/j.jecp.2005.07.005

Mathy, F., & Varré, J.-S. (2013). Retention-error patterns in complex alphanumeric serial-recall tasks. *Memory*, *21*(8), 945–968. https://doi.org/10.1080/09658211.2013.769607

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 and Aging, 14*(4), 572–594. https://doi.org/10.1037/0882-7974.14.4.572

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*(3), 222–252. https://doi.org/10.1006/jecp.1999.2550

McDaniel, M. A., Einstein, G. O., DeLosh, E. L., May, C. P., & Brady, P. (1995). The bizarreness effect: It's not surprising, it's complex. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 422–435. https://doi.org/10.1037/0278-7393.21.2.422

Nairne, J. S., Riegler, G. L., & Serra, M. (1991). Dissociative effects of generation on item and order retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 702–709. https://doi.org/10.1037/0278-7393.17.4.702

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, *33*(1), 31–88. https://doi.org/10.1145/375360.375365

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–53. https://doi.org/10.1016/0022-2836(70)90057-4

Norris, D., Kalm, K., & Hall, J. (2020). Chunking and redintegration in verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(5), 872–893. https://doi.org/10.1037/xlm0000762

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*(6), 1017-1045. https://doi.org/10.1016/S0191-8869(99)00251-2

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, *31*(2), 167–193. https://doi.org/10.1016/S0160-2896(02)00115-0

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, *47*(4), 1343-1355. https://doi.org/10.3758/s13428-014-0543-2

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual: Section 4, Advanced Progressive Matrices, 1998 edition*. Oxford Psychologists Press.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal Of Psychological Assessment, 28*(3), 164-171. https://doi.org/10.1027/1015-5759/a000123

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. *SIAM Journal on Applied Mathematics*, *26*(4), 787–793. https://doi.org/10.1137/0126070

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer., 2nd ed.* (pp. 159–184). Lawrence Erlbaum Associates Publishers.

Thurlow, W. R. (1950). Direct measures of discriminations among individuals performed by psychological tests. *The Journal of Psychology*, *29*(2), 281-314. https://doi.org/10.1080/00223980.1950.9916033

Tsukahara, J. S., Harrison, T. L., Draheim, C., Martin, J. D., & Engle, R. W. (2020). Attention control: The missing link between sensory discrimination and intelligence. *Attention, Perception, & Psychophysics*, *82*(7), 3445–3478. https://doi.org/10.3758/s13414-020-02044-9

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2011). Inter- and intra-individual variation in immediate free recall: An examination of serial position functions and recall initiation strategies. *Memory*, *19*(1), 67–82. https://doi.org/10.1080/09658211.2010.535658

Unsworth, N., & Engle, R. W. (2006). A temporal-contextual retrieval account of complex span: An analysis of errors. *Journal of Memory and Language*, *54*(3), 346–362. https://doi.org/10.1016/j.jml.2005.11.004

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133*(6), 1038-1066. https://doi.org/10.1037/0033-2909.133.6.1038

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498-505. https://doi.org/10.3758/bf03192720

van der Loo, M. P. J. (2014). The stringdist package for approximate string matching. *R Journal*, *6*(1), 111-122. https://doi.org/10.32614/RJ-2014-011

Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1207–1241. https://doi.org/10.1037/a0020122

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *49A*(1), 51–79. https://doi.org/10.1080/027249896392801

Weitzner, D. S., Calamia, M., Hill, B. D., & Elliott, E. M. (2021). Examining an alternative scoring procedure for a clinical working memory measure. *Assessment*. https://doi.org/10.1177/10731911211032270

Zinbarg, R. E., Revelle, W., Yovel, I., Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123-133. https://doi.org/10.1007/s11336-003-0974-7