Teaching English as a Second Language Quarterly (TESLQ)
(Formerly Journal of Teaching Language Skills)
41(4), Fall 2022, pp. 113-172
10.22099/tesl.2022.42703.3084
20.1001.1.20088191.2022.41.4.1.8

Online ISSN: 2717-1604
Print ISSN: 2008-8191
Research Paper

Methodological Synthesis of Working Memory Capacity Measures in Relative Clause Attachment Ambiguity Resolution Studies

Karim Vafaee Seresht [*] **Hamideh Marefat** [**]

Abstract

This synthesis reviews the methodological issues in empirical studies investigating the effect of working memory capacity (WMC) on relative clause ambiguity resolution where results have failed to be consistent cross-linguistically and even intra-linguistically. This discrepancy might have occurred due to 'methodological inconsistencies' in the design (Liu & Brown, 2015), administration, and scoring of WMC measures. This study aimed to investigate the aggregative and developmental status of the methodological practices of WMC measures and describe how transparently such practices have been reported. Based on a comprehensive search, 39 experiments were retrieved from 25 studies, culminating in a collection of studies with a time span of 22 years from 1999 to 2021, and coded for 46 features. Results revealed that although over the past 22 years, the field has witnessed significant improvements in the employment of WMC tests, there are still a lot of variations and inconsistencies calling for attempts to raise methodological awareness among researchers to afford more attention to quality and transparency in reporting WMC tests. The article concludes with a call for reform in standardizing WMC tests and a number of other recommendations for future primary and secondary research.

Keywords: Methodological Awareness, Methodological Synthesis, Methodological Transparency, Relative Clause Attachment Ambiguity Resolution, Working Memory Capacity

Received: 02/02/2022 Accepted: 24/04/2022

[*] Ph.D. Candidate, University of Tehran, kvafaee@gmail.com, Corresponding Author
[**] Professor, University of Tehran, marefat@ut.ac.ir

How to cite this article:
Vafaee Seresht, K., & Marefat, H. (2022). Methodological Synthesis of Working Memory Capacity Measures in Relative Clause Attachment Ambiguity Resolution Studies. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills), 41*(4), 113-172. doi: 10.22099/tesl.2022.42703.3084

Relative clause (RC) attachment ambiguity resolution has been extensively researched in the first language (L1, Cheng et al., 2021; Cuetos & Mitchell, 1988; Fernández, 2003) and second language (L2, Cheng et al., 2021; Papadopoulou & Clahsen, 2003, Dussias & Sagarra, 2007). Such ambiguities occur in sentences like *A customer frowned at the assistant*[NP1] *of the pharmacist*[NP2] *who was looking for a pen* in which the RC has two potential attachment sites in the preceding complex NP: NP1 (high attachment, HA) or NP2 (low attachment, LA). Disappointingly, the results of this domain of research have failed to be consistent cross-linguistically and even intra-linguistically.

As for *cross-linguistic inconsistency*, research has documented that native speakers (L1ers) of *English* (Cuetos & Mitchell, 1988; Fernández, 2003; Frazier, 1979), *Arabic* (Abdelghany & Fodor, 1999), *Romanian*, *Swedish*, and *Norwegian* (Ehrlich et al., 1999) have LA preference while *Spanish* (Carreiras, 1992; Cuetos & Mitchell, 1988; Fernández 2003), *French* (Frenck-Mestre & Pynte, 1997), *German* (Hemforth et al., 2000), *Japanese* (Kamide & Mitchell, 1997), *Dutch* (Brysbaert & Mitchell, 1996), *Italian* (De Vincenzi & Job, 1993), and *Persian* (Arabmofrad & Marefat, 2008) L1ers have a HA preference.

As regards *within-language inconsistency*, in English, there are studies evidencing HA (Hopp, 2014), LA (Felser et al., 2003; Kim & Christianson, 2017), and no robust resolution towards either attachment site (Carreiras & Clifton, 1999). One potential factor for such variation in results, particularly within-language differences, is postulated to be WMC (Hopp, 2014; Kim & Christianson, 2017). However, even research investigating the interaction of WMC with RC attachment preference has yielded conflicting results. In *online* tasks, some studies evidenced that LA in English increases as a function of higher WMC (Hocking, 2003), while some other studies reported the opposite pattern (Felser et al., 2003), and still some others reported no WMC effect (James et al., 2018). Likewise, in *offline* tasks, some studies reported that LA in English increases as a function of higher WMC (Cheng et al., 2021; James et al., 2018; Payne et al., 2014; Swets et al., 2007); however, no reliable WMC effect is observed in some other studies (Felser et al., 2003; Hocking, 2003; Traxler, 2009).

This discrepancy might have occurred due to 'methodological inconsistencies' in the design, administration, and scoring of WMC measures. Unlike single primary research, synthetic research and other forms of secondary research provide a higher level of evidence (Bigby, 2009; McClean et al., 2019) with a higher degree of generalizability (Loewen & Plonsky, 2015). Despite the fact that almost sufficient attention has been

devoted to conducting substantive syntheses in the field (e.g., Norris & Ortega, 2000, 2006; Ortega, 2003; Oswald & Plonsky, 2010; Shakki et al., 2020), scarce attention has been afforded to conducting *methodological syntheses* (Farsani & Babaii, 2020; Liu & Brown, 2015; Marsden, Thompson et al., 2018; Plonsky, 2014; Plonsky et al., 2020; Plonsky & Ghanbar, 2018; Plonsky & Gonulal, 2015; Plonsky & Kim, 2016; Sok et al., 2019; Zhang & Plonsky, 2020). Particularly, the design, administration, and scoring of WMC measures as employed in the investigation of RC ambiguity resolution in both L1 and L2 research, to the best of our knowledge, have hardly received any attention. This necessitates the need to undertake a principled methodological synthesis.

Methodological Synthesis

Unlike substantive syntheses, which "seek to aggregate the results of primary studies and reach conclusions" (Li & Wang, 2018, p. 312), methodological synthesis focuses "on the methods that have produced them" (Marsden, Thompson et al., 2018, p. 6). In essence, methodological syntheses focus on "methodological aspects of the primary research with a view to evaluating whether current practices meet certain criteria and what improvements can be made" (Li & Wang, 2018, p. 132). Moreover, concerning their primary objectives, methodological syntheses are used to *describe*, *evaluate*, *identify relationships*, or *provide chronological changes or improvements* (Plonsky & Gonulal, 2015).

The number of methodological syntheses is experiencing a burgeoning growth in the field. In one such study, Liu and Brown (2015) conducted a methodological synthesis on the effectiveness of corrective feedback in L2 writing. They investigated 32 published studies and 12 dissertations with a focus on both strengths and weaknesses of the retrieved primary studies. Their synthesis shows a number of praiseworthy design features like the use of 'classroom-based research tradition' and 'inclusive coverage of common corrective feedback strategies'. However, what they underscore more are some methodological limitations, including (a) poorly reported research context, methodology, and statistical analyses, (b) experimental designs of low generalizability, (c) the use of split-plot designs that make it impossible to find out the meaningful effects of feedback, and (d) the use of varying measures that make comparability of the results difficult.

In a substantive and methodological synthesis, Plonsky (2014) examined changes over time in research and reporting practices of 606 primary studies from the journal of *Language Learning* and *Studies in Second Language Acquisition*. In the methodological

synthesis part, he mainly focused on 'design preferences', revealing that experimental research makes up a substantial ratio of quantitative studies and that observational research was still in the majority. The results also showed a move towards more internally valid experimental research designs, which could be viewed as "an indication of the maturity of our domain" (p. 463).

Likewise, Plonsky and Kim (2016) synthesized substantive and methodological practices in task-based learner production. They retrieved 85 primary studies of task-based language production published from 2006 to 2015 and coded for the methodological features of study designs, sampling, analyses, and reporting practices, about which they point out a number of concerns (e.g., adopting a pretest-posttest design in more recent studies, sampling mainly highly educated young adult users of English, etc.).

In a methodological synthesis of self-paced reading (SPR), Marsden, Thompson et al. (2018) synthesized the rationales, study contexts, and methodological decision-making of 74 SPRs used in L2 research. They coded each instrument along 121 features. Facing too much variability in the SPR instruments employed, they call for "an urgent need to standardize the use and reporting of this technique" (p. 861), hoping to elevate our understanding of language processing, reading, and learning in L2 (Samavarchi & Rezai, 2014), and ultimately to reduce the impact of methodological issues on findings.

Moreover, in an attempt to describe and evaluate the use of regression analysis in the field of L2 research, Plonsky and Ghanbar (2018) synthesized a total of 541 regression analyses in 171 primary studies. They coded the studies for different statistical models, variables, procedures, reporting practices, and overall variance explained (R^2), and obtained a number of methodological inconsistencies and a lack of transparency in reporting practices.

In order to illustrate the scope of inquiry of collaborative writing in face-to-face L2 settings, Zhang and Plonsky (2020) conducted a methodological and substantive synthesis of 94 quantitative primary studies. As for the methodological synthesis, they coded each study for features like research design, analyses, and reporting practices affiliated with transparency. The findings revealed a strong tendency towards testing mean differences, a reliance on homemade prompts with occasional reporting of the piloting procedure, and inadequate reporting of pre-task training and reliability estimates.

In the reviewed methodological syntheses, different methodological aspects have been examined. One such aspect which is commonly addressed in almost all syntheses is what Marsden (2020) labels 'methodological transparency'.

Methodological Transparency

The importance of methodological transparency is underscored by Plonsky and Gass (2011) who declare that progress in any field of inquiry "depends on sound research methods, principled data analysis, and transparent reporting practices" (p. 325), which require what Plonsky (2017) refers to as "methodological awareness" (p. 508), or Whitney and Budd (1999) label as 'methodological power'. But what is meant by 'transparent reporting practices', and why is it important? These questions are succinctly answered by Marsden (2020) when she asserts

> Methodological transparency can involve all aspects of the research process, from initial design, through peer review, to dissemination of findings. It means making the research process fully transparent so that reviewers and readers can understand exactly what the researchers did to elicit, analyze, and understand their data; that is, how they moved from their research aims to data to findings to interpretation. (p. 15)

Furthermore, along with variations in "aspects of the materials" and "variation in the participants" (Fernández & Sekerina, 2015), 'methodological inconsistencies' (Liu & Brown, 2015) in the design, administration, and scoring of measures can be considered as modulating variables in research practices. Such methodological inconsistencies stem either from methodological advances or researchers' *imperfect replications* of previous research, which may emerge from non-transparent reporting practices (Marsden, 2020). Since methodological transparency is an indicator of methodological quality (Derrick, 2016), most researchers address it when doing methodological syntheses. This aspect is addressed in RQ2 in the current study.

As stated above, the incongruence between the results in RC attachment preferences may have stemmed from variations in the design, administration, and scoring of WMC measures. Moreover, to make a call for standardization and to raise scholarly awareness in the methodological practices and transparency of the features of WMC measures, following other *descriptive* syntheses (e.g., Farsani et al., 2021; Hou & Aryadoust, 2021;

Liu & Brown, 2015; Sok, et al., 2018), we conducted the current methodological synthesis to provide a *descriptive* synthesis of the multifarious methodological features of WMC measures. This study also aims at promoting the methodological rigor of future research by scrutinizing methodological features of WMC measures. With such objectives in mind, we developed the following research questions.

RQ1. How have WMC tests been designed, administered, and scored in the literature on RC ambiguity resolution?

RQ2. To what extent have the WMC tests been reported transparently in the literature on RC ambiguity resolution?

RQ3. Has there been any improvement in the design, administration, and scoring features of the WMC tests used in the literature on RC ambiguity resolution over the past 22 years?

Method

Study Retrieval

To identify and retrieve the relevant primary studies for the current synthesis, we considered no a priori starting date; however, it included studies published through the end of 2021. This culminated in a collection of studies with a time span of 22 years, beginning in 1999 (Mendelsohn & Pearlmutter, 1999) and ending in 2021 (Cheng et al., 2021).

To provide a representative and inclusive collection of studies, we aimed for all 'peer-reviewed research' and 'fugitive literature', unpublished literature in the form of master's and doctoral theses, and papers in conference proceedings. This approach could help us avoid 'publication bias' (Nakanishi, 2014; Pigott, 2012) or 'file drawer problem' (Rosenthal, 1979), which might arise from including only published research at the exclusion of fugitive literature.

To embark, following Plonsky and Oswald (2015), we exercised a comprehensive and exhaustive keyword search in databases and journals including *Linguistics and Language Behavior Abstracts* (LLBA), *Education Resources Information Center* (ERIC), *Academic Search Ultimate* (ASU) *Scholar Google, ScienceDirect, PsycINFO, IRIS database* (Marsden et al., 2016), *ProQuest, Academia.edu, ResearchGate.net, IRANDOC,* an Iranian local database of M.A. theses and Ph.D. dissertations (http://www.irandoc.ac.ir), and the *Central Library of the University of Tehran* (http://utdlib.ut.ac.ir, UTDLIB). Next, to find any missing relevant research, we applied

"citation chaining" (Ziegler, 2016) through Google Scholar by following the "Cited by" link below the relevant research and "ancestry chasing" (Li & Wang, 2018) by browsing and mining the references sections of the pertinent research.

After a few trials with the keyword search, we came up with the following search terms: ("working memory" + "relative clause attachment"), ("working memory" + "relative clause ambiguity resolution"), and ("working memory" + "relative clause resolution"). Each search yielded many studies, which totaled 1,720. After removing duplicates, the results were reduced to 1,073 ones. Moreover, through *citation chaining* and *ancestry chasing*, the number of these studies rose to 1,079 potential studies. Having read the 'titles' and 'abstracts' of the studies to see if they were relevant to the synthesis, we experienced a sharp reduction in the number of potential research down to 53 studies. However, we did not rely solely on reading titles and abstracts: When we suspected that a study might be pertinent, we searched through the full text for the relevant keywords (i.e., for 'working memory', 'WM', and 'span'), and then perused the method sections. This eventuated in a further reduction of studies down to 25 ones, including 39 experiments (Figure 1). Since the domain of study was narrowed down to experiments conducted on the effect of WMC on RC attachment ambiguity resolution, the number of experiments was limited to 39 ones.

These studies included 12 journal articles, 2 experimental book chapters, 3 conference proceedings, 12 theses and dissertations (four of which were published as journal articles, and hence excluded).

Inclusion and Exclusion Criteria

Both published and fugitive literature were included to avoid the 'file drawer problem' (Rosenthal, 1979) and to have a more comprehensive and exhaustive sample for the synthesis. Furthermore, since the methodology used to measure children's WMC is different from that employed for tapping adults', studies conducted on children (14 years old and younger, Brown, 2014) were excluded. This resulted in the exclusion of one study (Felser et al., 2003).

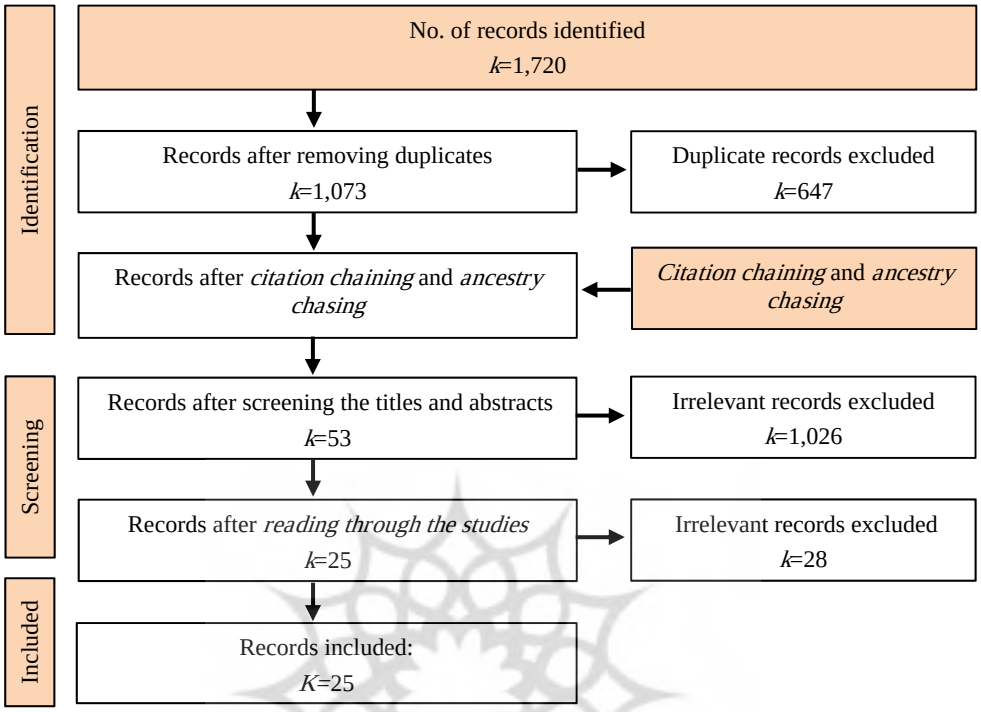

Figure 1. *The Retrieval Process*

Coding Scheme

Having decided on the inclusion and exclusion criteria, we needed to have a coding scheme as our data collection instrument for better categorization and listing of the features of the included studies. Inspired by the coding schemes from the IRIS database (iris-database.org; see Marsden et al., 2016), we developed a topical coding sheet for the studies to be examined. As for the development of the coding scheme, we read through all the studies and added features to the coding scheme based on the criticisms of scholars in the field and based on authors' descriptions of the design, administration and scoring features of the WMC measure they employed in their study. This was not a one-shot task: as we read through other studies, we iteratively went back to read the studies again for the newly detected features. Also, at first, the features were not so neatly neatly classified based on the three categories of 'design', 'administration' and 'scoring' features. As we proceeded and added the features, we perceived that the features could be categorized. Since we did not intend to do a meta-analysis in the current project, we included some

statistically related issues like reliability under the scoring features to make the categories more inclusive (see Appendix for the final coding scheme).

Next, the first author piloted the coding scheme on 5 studies, which resulted in adding some more criteria to consider. The revised scheme was then employed to code all the studies. However, as the first author perused the pertinent studies for the relevant details, he came across some procedures and features that were either unavailable in the previous studies or had gone unnoticed. This made us continually add some new criteria for consideration: The features of every single study inspired us to re-read all the studies recurrently and to add new criteria for consideration. This iterative process eventually resulted in 46 features.

When the first author finished the coding process, the second author coded 13 out of the 39 experiments (33.33%). Cohen's kappa was calculated to avoid potential reliability overestimation that might arise from a chance agreement between coders when using the 'agreement coefficient' (Ary et al., 2019), which yielded sufficient agreement between the coders (κ=.968). Afterward, we arranged a consensus-approaching meeting when we discussed the codes. We managed to resolve all cases of disagreement in the coding. Since the codes were totally topical and the *presence* or *absence* of the features in the experiments were considered (not the interpretation of the current synthesists), the cases of disagreement were resolved by making reference to the exact parts in which the features were referred to.

Results

RQ1. Design, Administration, and Scoring of WMC Tests

To address RQ1, in what follows, we present the results of the synthesis in three categories: design, administration and scoring of the WMC measures. A first critical *design* feature appertains to the 'memory components' in the task. Of the 39 reviewed experiments (Table 1), 5.13% (k=2) employed simple span tests (tests that tap only the storage component of WM), and 94.87% (k=37) used complex ones (tests that gauge both the storage and processing components of WM).

By operationalizing, Baddeley and Hatch's (1974) multi-component model of WM, Daneman and Carpenter (1980) developed a complex reading span test. Standing on the shoulders of these giants, other researchers have developed some variants of the test. Table 1 charts the origins of WMC tests in the retrieved literature. Researchers reported the origins for 87.18% (k=34) of the tests used. Moreover, when a test is not already

developed and validated, researchers should pilot it and report the procedure and the results (Derrick, 2016). In this respect, interestingly, only a single (2.56%) study (Soleimani, 2018) had piloted the WMC, though the piloting procedure and results had not been detailed.

Another critical design feature concerns the use of single vs. multiple measures for tapping WMC. In a multiple-measure, psychometric approach, more than one measure is employed to investigate the WMC construct and to ensure the internal validity of measurement (see Appendix; Swets et al., 2007). Of the 25 studies, only 12% (*k*=3) espoused a multiple-measure, psychometric approach, while 88% (*k*=22) adopted a single-measure approach (Table 1).

Table 1 also exhibits that the reading span test ranked as the most widely employed test (*k*=25, 64.1%), followed by the operation span test (a language neutral WMC test which is evidenced not to correlate with and not to be confounded by the construct of language experience – see Appendix; James et al., 2018; Li et al., 2019; MacDonald & Christiansen, 2002; Najjari & Mohammadi, 2017; Shin, 2020), employed in 15.38% (*k*=6) of tests. The other types of tests have been scantly used.

Apropos of *language* of WMC tests, as shown in Table 1, 61.54% (*k*=24) of tests were in English for both English L1ers and L2ers, indicative of the fact that the effect of WMC on RC ambiguity resolution is predominantly investigated in English. The next ranking test (*k*=5, 12.82%) was the operation span test, a mathematical, language-independent test (see Shin, 2020).

A feature specific to reading and listening span tests is *controlling sentence length and complexity*. Ariji et al. (2003) and Omaki (2005) argue that the sentences in the Daneman and Carpenter (1980) test were randomly taken from magazines and thus were not controlled and evenly distributed with regard to their length and complexity across different sets, which in turn might confound the results. However, this threat to validity is addressed by Ariji et al. (2003) and Omaki (2005), who include only four types of sentences that are controlled for their length and complexity and that are evenly distributed across all conditions of their WMC tests. In the retrieved literature, as exhibited in Table 1, this feature applied to 69.23% (*k*=27) of tests, from which only 15.38% (*k*=6) addressed both sentence length and complexity, and 10.26% (*k*=4) addressed only sentence length. The remaining 43.59% (*k*=16) either did not consider this feature (*k*=4, 10.26%) or did not report any pertinent information (*k*=12, 33.33%).

Table 1.

Some Design Features of WMC Tests

Feature	k	%
WM Task		
Complex	37	94.87
Simple	2	5.13
Origins of WMC Tests		
Already developed	22	56.41
Researcher-developed	6	15.38
Researcher-adapted	3	7.69
Translated version	3	7.69
No information	5	12.82
Approach to WMC Measurement		
Single-Measure	22	88.00
Multiple-Measure	3	12.00
WMC Measure		
Reading Span	26	64.10
Operation Span	6	15.38
Alphabet Span	2	5.13
Minus Digit Span	2	5.13
Word Span	1	2.56
Listening Span	1	2.56
Spatial Span	1	2.56
No Information	1	2.56
Language of WMC Test		
English	25	61.54
OST (Language Neutral)	5	12.82
Turkish	4	10.26
Korean	2	5.13
Chinese Russian	1	2.56
Japanese	1	2.56
Dutch	1	2.56
Addressing Sentence Length and Complexity	12	30.77
Not applicable	6	15.38
Both sentence length and complexity	4	10.26
Neither sentence length nor complexity	4	10.26
Just sentence length	13	33.33
No information		

Concerning recall task type, which engages and taps the storage component, as Table 2 displays, the most commonly used tasks, in rank order, were 'letters after sentences or equations' (*k*=10, 25.64%,), 'final words of the sentence' (*k*=8, 20.51%), 'enhanced words in non-final positions' (17.95%, *k*=5), and 'words after the sentence' (*k*=4, 10.26%).

It is claimed that individual differences in recalling the to-be-recalled 'words' may arise from participants' language experience with the words rather than their WMC. Unsworth et al. (2005) suggest using 'letters' presented after the comprehension task to address this issue. However, the use of 'letters' or 'words' in sentence-final positions has been asserted to pose yet another problem. Omaki (2005) and Ariji et al. (2003) argue that this task does not simultaneously tax WM's processing and storage components. They assert that using words in the final position and after element presentation does not simultaneously tax the processing and storage components: for simultaneous taxation of both components, the to-be-recalled information must be introduced somewhere in the middle of the to-be-processed element. Consequently, to help solve this problem, they suggest using non-final words as target words for the recall task.

As shown in Table 2, to engage the processing component of WM, in 46.15% (*k*=18) of tests, 'comprehension questions', in 17.95% (*k*=7), both 'reading aloud and comprehension questions', in 15.38% (*k*=6), 'reading aloud' and in 2.56% (*k*=1), 'Normal/Mirror questions' were employed. However, in 10.26% (*k*=4) of tests, 'no information' was reported, and in 7.69% (*k*=3), 'simple span tests' were employed. It is argued that, in complex span tests, 'reading aloud' alone does not ensure taxing the processing component of WM (Ariji et al., 2003; Omaki, 2005). While 'reading aloud', participants may focus more on correct pronunciation than on processing the task. Thus, to ensure that the processing component is taxed, researchers are recommended to use 'comprehension questions'.

Following Ariji et al. (2003) and Omaki (2005), we classified WMC measures of the literature as 'simultaneous' and 'non-simultaneous' loads on WM. As charted in Table 2, only 28.21% (*k*=11) of experiments employed a design that simultaneously taxed the storage and processing components. These experiments included the to-be-recalled information in non-final positions within the element. In fact, we coded 'words in alphabetical order span test', 'minus digit test', 'direction of tops of letters', and those reading span tests in which the elements were presented in non-final positions as 'simultaneous' since such tests were assumed to be taxing both storage and processing

components of WM simultaneously. The tests that present the to-be-recalled information in the final position of elements or after elements were coded as tests that were assumed to be taxing storage and processing components of WM non-simultaneously (*k*=22, 56.41%). This feature was not applicable to simple span tests that tax only the storage component of WM (*k*=2, 5.13%).

As a very trivial, yet, most probably, modulating factor, unfamiliarity with test instructions may engender variability in results. Only 44% (*k*=17) of the retrieved literature attended to deconfounding the findings by introducing practice items before administering WMC tests (Table 2)

Table 2.
Some Other Design Features of WMC Tests

Feature	k	%
Recall Tasks		
Letters after sentences or equations	10	25.64
Final words of the sentence	8	20.51
Enhanced words in non-final position	7	17.95
Words after the sentence	4	10.26
Words in alphabetical order	2	5.13
Minus two-digit task	1	2.56
Direction of tops of letters	1	2.56
Storage-only: Correctly remembered No. of words	1	2.56
Storage-only: Digits in the same order as heard	1	2.56
No information	4	10.26
Engaging and Tapping the Processing Component		
Comprehension questions	18	46.15
Reading aloud and comprehension questions	7	17.95
Reading aloud	6	15.38
Not applicable	3	7.69
Normal/Mirror questions	1	2.56
No information	4	10.2
(Non-)simultaneous Engagement of WM Components		
Non-simultaneous	22	56.41
Simultaneous	11	28.21
Not applicable	2	5.13
No information	4	10.26
Practice Items		
Included	17	44
Not included	22	56

Table 3 portrays how researchers included different numbers of elements, items and sets in their WMC tests. Such variation may impact participants' performance positively

(e.g., they may develop test-taking strategies) or negatively (e.g., they may get tired when taking the tests), which may, in turn, lead to differences in scores, probable differences in being misclassified as high span or low span, and variability in substantive results (Conway et al., 2005; Fedorova & Yanovich, 2005)

Table 3.
Set Size, Item Size and Total Number of Elements

Feature	k	%
Set Size		
3	1	2.56
4	19	48.72
5	11	28.21
7	1	2.56
No information	7	17.95
Item Size		
2	5	12.82
3	10	25.64
4	1	2.56
5	15	38.46
40	1	2.56
No information	7	17.95
No. of Elements		
36	3	7.69
40	2	5.13
42	6	15.38
60	3	7.69
70	10	25.64
75	2	5.13
80	1	2.56
100	3	7.69
120	1	2.56
175	1	2.56
No information	7	17.95

Note. Following Conway et al. (2005), we call individual sentences and equations ‘elements’, a set of elements ‘an item’ and a set of items ‘a set’.

Table 4 shows variation regarding the shortest and longest item sizes in the reviewed tests, which could also impact substantive results. As seen, the most frequent shortest element is ‘2 elements’ (*k*=29, 74.36%), and the most frequent longest elements, in rank order, are ‘5 elements’ (*k*=17, 43.59%) and ‘6 elements’ (*k*=13, 33.33%).

As shown in Table 4, the predominant comprehension questions for engaging and tapping the processing accuracy were 'True/False questions' (64.1%, *k*=25). This feature was not applicable for storage-only tests (*k*=2, 5.13%).

Some of the studies attempted to address a 'positive response tendency' (Elliott et al., 2009), a tendency for some participants to respond more favorably to some questionnaire items (Table 4). To avoid this problem and hence add to the internal validity of research, in 38.46% (*k*=15) of experiments, question responses were distributed (roughly) equally: Half of the responses were true, and half were false. However, this feature was not applicable for 'word span', 'alphabet span', 'minus digit span', and 'spatial span' tests since they did not capture a True/False dichotomy.

Table 4.
Some More Design Features of WMC Tests

Feature	k	%
Shortest Element		
2	29	74.36
3	4	10.26
No information	6	15.38
Longest Element		
5	17	43.59
6	13	33.33
8	3	7.69
7	1	2.56
No information	5	12.82
Comprehension Question Type		
True/False	25	64.10
Not applicable	2	5.13
Normal/Mirror	1	2.56
No information	11	28.21
Avoiding Positive Response Tendency		
Half true, half false	15	38.46
Not applicable	6	15.38
No information	18	46.15

Table 5 illustrates some administration features of the reviewed WMC tests. As shown, the most widely used presentation instruments, in rank order, were 'screen', *k*=11, 28.21%), 'E-Prime' software (*k*=10, 25.64%), and Microsoft PowerPoint software (*k*=4, 10.26%). As regards presentation type, the most commonly used methods were 'entire element at once' (*k*=16, 41.03%), and 'non-cumulative presentation' (*k*=6, 15.38%). In addition, while ascending order presentation was predominately employed for the tests

(*k*=20, 51.26%), in 10.26% (*k*=4) of tests, the items were presented in randomized order. James et al. (2018) assert that most complex WMC tests are administered in an ascending order, which raises the issue of 'proactive interference' in which memory performance is reduced for recently processed information because of the existence of prior processing of related materials. Thus, they recommend that set sizes be distributed randomly to ensure that only WMC leads to individual differences in test performance (Conway et al., 2005; Lustig et al., 2001).

Further, in 5.13% (*k*=2) of tests, test items were counterbalanced, while in 35.90% (*k*=23), they were not. Swets et al. (2007) argue for the use of non-counterbalanced test items. They remark that variability in counterbalanced designs "runs in direct opposition to the goal of individual differences studies, which is to explain as much of the variance due to individual differences as possible—while minimizing the variance due to task differences" (p. 67). Therefore, to minimize the effect of individual differences other than WMC, they recommend administering materials in the same order to all participants.

Charted in Table 5 is also information about whether researchers discontinued the tests if the participants failed to recall a certain number of words or letters. For example, in some studies, researchers decided to stop the administration process of WMC tests when participants 'made two or more mistakes' in recalling the to-be-recalled information. In some other studies, researchers decided to stop the test administration process when participants 'failed to recall the words from two consecutive items'. Thus, based on certain criteria, the tests were discontinued in 10.26% (*k*=4) of test administrations, while experiments were not in 28.21% (*k*=11) of tests. Among these non-discontinuations, only 7.69% (*k*=3) of experiments justified why the tests were not discontinued, and the rest (*k*=24, 61.54%) did not. James et al. (2018) claim that discontinuing WMC tests "early reduces the data collected from each participant" (p. 162). To address this issue, they recommend having participants complete the entire test. Also, Table 5 shows whether the WMC tests were administered individually or in groups. While in 76.92% (*k*=30) of cases, test administration was individual-based, only a small portion (*k*=4, 10.26%) was group-based.

Table 5.

Some Administration Features of WMC Tests

Feature	k	%
Presentation Instrument		
Screen	11	28.21
E-Prime	10	25.64
Microsoft PowerPoint	4	10.26
DMDX	2	5.13
Experimenter reads	2	5.13
No information	10	25.64
Presentation Type		
Entire element is presented at once	15	38.46
Non-cumulative	6	15.38
Entire element is listened to at once	3	7.69
Not applicable	3	7.69
No information	12	30.77
Presentation Order of Items		
Randomized	4	10.26
Ascending	20	51.28
No information	15	38.46
Counterbalancing		
Yes	2	5.13
No	23	35.90
No information	14	58.97
Test Discontinuation		
Yes	4	10.26
No	11	28.21
No information	24	61.54
Criteria for Discontinuing Tests		
Failing all three items in a set	1	2.56
Failing to recall the words from two consecutive items	1	2.56
Making two or more mistakes in a set	2	5.13
Not applicable because there is no discontinuation	12	30.77
No information	23	58.97
Individual-based or Group-based Administration		
Individual-based	30	76.92
Group-based	4	10.26
No information	5	12.82

The timing of test parts was treated differently by the researchers. As regards the timing of element presentation, a fixed, group-based timing was considered in 30.77% (k=12) of tests (Table 6), whereas an individually calibrated timing (James et al., 2018) was employed in 7.69% (k=3) of tests. As for the questions, there was no time limit in 5.13% (k=2) of tests. In contrast, in 17.95% (k=7) of tests, a fixed, group-based timing, and in 7.69% (k=3), an individually calibrated timing was included. Finally, the tests were

administered with no time limit in 10.26% (k=4) of tests for the to-be-recalled information. On the other hand, in another 10.26% (*k*=4), they were administered following a fixed time limit (Leeser & Sunderman, 2016).

It has been argued that when tapping the processing component, one should address both 'processing accuracy' and 'processing speed/time' (James et al., 2018; Unsworth et al., 2009); otherwise, participants who are faster in sentence processing may have extra time for rehearsing the to-be-recalled information. To address this issue more effectively, James et al. (2018) recommend administering 'individually calibrated' WMC tests in which each individual's processing time is calibrated based on a piloting phase.

Table 6.
Timing Features of WMC Tests

Feature	k	%
Timing for element presentation		
Considered	12	30.77
Individually calibrated timing	3	7.69
No information	24	61.54
Timing for judgment task		
Considered	7	17.95
Individually calibrated timing	3	7.69
No time limit	2	5.13
No information	27	69.23
Timing for recall task		
Considered	4	10.26
No time limit	4	10.26
No information	31	79.49

Table 7 portrays the type of recording for responses in the processing and recall tasks. As for the type of recording for the processing task (i.e., comprehension questions), software recording ranked first (*k*=9, 23.08%). The second most common type of recording was that done by the experimenter (*k*=7, 17.95%). In contrast, when it came to recording the recall task items, the most common strategy was to have the participants write their answers on sheets of paper (*k*=12, 30.77), followed by the experimenter doing that (*k*=8, 20.51%).

Table 7.

Recording of Responses

Feature	k	%
Recording responses to comprehension questions		
Participants wrote responses on answer sheets	3	7.96
Experimenter wrote responses on answer sheets	7	17.95
Software recorded responses	9	23.08
Not applicable	2	5.13
No information	18	46.18
Recording the to-be-recalled information		
Participants wrote responses down on answer sheets	12	30.77
Participants typed responses	3	7.69
Experimenter wrote responses down on answer sheets	8	20.51
No information	16	41.03

The WMC tests and other tests of the experiments were mostly administered in a single session, as illustrated in Table 8, with the WMC tests preceding or following the other tests.

Table 8.

Sessions and Sequence of Tests

Sessions and Sequence of Tests	k	%
A single session with no information about sequencing	3	7.69
One session: WMC test administered 'before' RC attachment task	6	15.38
One session: WMC test administered 'after' RC attachment task	7	17.95
Two sessions: WMC test administered before RC attachment task	2	5.13
Two sessions without reference to task sequence	3	7.69
No information	18	46.15

Regarding scoring features of the WMC tests, as portrayed in Table 9, the first critical feature which was under-reported in the reviewed studies was reliability. This critical feature was addressed only in 23.08% (k=9) of tests. In three experiments (James et al., 2018), 'inter-rater reliability' and split-half reliability' were investigated. In three other experiments (Swets et al., 2007), coefficient alphas were investigated. The Pearson correlation was employed in two experiments (Kaya, 2012) to estimate reliability. Finally, in one experiment (Marefat et al., 2015), KR-21 was used. Moreover, reliability was discussed in 15.38% (k=6) of experiments, but a coefficient was not reported.

As for the measurement scale, as charted in Table 9, WMC was treated as a nominal variable in 46.15% (k=18) of tests and continuous in 48.72% (k=19) of tests. In this regard, James et al. (2018) indicate that considering variables as continuous reflects the

full range of variables across individuals, is more powerful than nominal scale and "yields more accurate estimates of effect size and lower rates of Type I error" (p. 170).

Also, we examined the way the experiments assigned a cut-off score. This feature was not applicable when WMC was considered continuous (*k*=19, 48.7%). Nevertheless, as shown in Table 9, a cut-off point was determined differentially in the reviewed literature, with 'mean as the cut-off point' being the most frequent method (*k*=6, 15.4%).

Table 9.
Some Scoring Features of WMC Tests

Feature	k	%
Reliability Coefficients		
Reported	9	23.08
Not reported	6	15.38
No information	24	61.54
WMC as continuous or nominal?		
Continuous	19	48.72
Nominal	18	46.15
No information	2	5.13
Cut-off score		
Not applicable for a continuous variable	19	48.72
Mean as the cut-off point	6	15.38
High span (span $\geq$ 4) or low span (span < 4)	2	5.13
30th and 70th percentiles as cut-off points	1	2.56
No information	11	28.21

In dual-component WMC tests, there are two sources of data for *scoring*: one from the storage/recall component and one from the processing component (Conway et al., 2005). This is further complicated when scores are assigned to recall order and partial performance. As exhibited in Table 10, there was great variability in how performance on WMC tests was scored. While some studies credited scores only to the recall component, some considered both recall and processing components in scoring. Some others granted scores for recall orders, as well. Still, others devoted some scores for partial performance on test parts. Finally, some experiments considered a 'threshold level' for comprehension questions below which the data from participants were excluded from all analyses (Leeser & Sunderman, 2016).

Table 10.

Features of the Scoring System

Feature	k	%
Detailed features of scoring system		
[element level, +recall, +processing, -order, -fractional]	6	15.38
[element level, +recall, -processing, -order, -fractional]	5	12.82
[element level, +recall, +85%processing for all the test, -order, -fractional]	1	2.56
[item level, +recall, -processing, +order, -fractional]	12	30.77
[item level, +recall, +processing, +order, -fractional]	3	7.69
[item level, +recall, -processing, +order, +fractional]	1	2.56
[item level, +recall, +85%processing for all the test, +order, -fractional]	1	2.56
[set level, +recall, +75% processing for each set, -order, +fractional]	2	5.13
[set level, +recall, -processing, -order, -fractional]	2	5.13
[set level, +recall, -processing, -order, +fractional]	1	2.56
No information	5	12.82
Threshold level		
Not used	36	92.31
Used	3	7.69
Scoring method		
Inferred to be partial-credit load scoring	12	30.77
Inferred to be absolute scoring method*	4	10.26
Inferred to be partial-credit unit scoring	3	7.69
Composite scoring of recall and processing	6	15.38
Partial-credit unit scoring	4	10.26
Composite scoring of recall, processing, and order	3	7.69
All-or-nothing unit scoring	1	2.56
All-or-nothing load scoring	1	2.56
No information	5	12.82

Note. *Absolute scoring method is the one that affords one score when a 'whole set' is responded correctly (Conway et al., 2005).

RQ2. Transparency of Reporting WMC Tests

As stated previously, RQ2 aimed to investigate the extent to which the WMC tests have been reported transparently in the literature of RC ambiguity resolution. Tables 11-13 depict how transparency was addressed in the design, administration and scoring features of the retrieved literature.

This synthetic study showed that the most transparent design feature is 'WMC task type' (*k*=38, 97.44%), as displayed in Table 11. On the other hand, the three least transparent design features, in rank order, are 'inclusion of practice items' (*k*=17, 43.59%) and 'avoiding positive response tendency' (*k*=21, 53.85%).

Table 11.

Transparency Information for Design Features

Information is provided for …	k	%
WMC task type	38	97.44
Recall task type	35	89.76
Engaging and tapping the processing component of WM	35	89.76
Simultaneous and non-simultaneous loading of WM	35	89.76
Origins of WMC tests	34	87.18
Longest item size	34	87.18
Shortest item size	33	84.62
Information about set size	32	82.05
Information about item size	32	82.05
Information about total number of elements	32	82.05
Comprehension question type	28	71.79
Addressing sentence length and complexity	26	66.67
Avoiding positive response tendency	21	53.85
Inclusion of practice items	17	43.59
Mean		79.12

Table 12 provides information on what percentage of studies supplies information about administration features. The most transparent administration feature is 'individual- vs. group-based test administration' (*k*=34, 87.18%). The second most transparent administration feature is the 'presentation instrument' (*k*=29, 74.36%). By contrast, the least transparent administration feature is 'timing for the to-be-recalled information' (*k*=8, 20.51%), followed by 'timing for processing questions' (*k*=12, 30.77%).

Table 12.

Transparency Information for Administration Features

Information is provided for …	K	%
Individual- vs. group-based test administration	34	87.18
Presentation instrument	29	74.36
Presentation type	27	69.23
Presentation order of items	24	61.54
Recording the to-be-recalled information	23	58.97
Sessions and sequence of tests	21	53.85
Recording comprehension questions	21	53.85
Counterbalancing	16	41.03
Criteria for discontinuing the tests	16	41.03
Test discontinuation	15	38.46
Timing for element presentation	15	38.46
Timing for processing questions	12	30.77
Timing for the to-be-recalled information	8	20.51
Mean		51.48

As shown in Table 13, the two most transparent scoring features for which direct or indirect mention is provided in the experiments, in rank order, are 'measurement scale of WMC' (*k*=37, 94.87%) and 'scoring method' (*k*=34, 87.18%). In contrast, 'reliability reporting' is the least reported scoring feature (*k*=15, 38.66%).

Table 13.
Transparency Information for Scoring Features

Information is provided for …	k	%
Measurement scale of WMC	37	97.87
Scoring method	34	87.18
Cut-off scores	28	71.79
Reliability reporting	15	38.66
Mean		73.87

Two more transparency features relate to whether sample test elements were provided in the materials section and whether full WMC tests were provided in the appendix or supplementary materials of the study. Sample items were provided in 43.59% (k=17) of experiments, and only in 20.51% (*k*=8) of experiments full WMC tests were provided in the appendix sections or supplementary materials.

Table 14.
Two More Transparency Features

Feature	k	%
Sample test items in the materials section		
Provided	17	43.59
Not provided	22	56.41
Supplementary materials (e.g., in the appendix)		
Provided	8	20.51
Not provided	31	79.49

RQ3. Changes in WMC Tests over Time

To investigate whether the WMC tests in the literature of RC ambiguity resolution have gone through any changes or improvements over the past 22 years, we split the period into three roughly seven-year periods with 13 cases of WMC tests in each time span: 1999-2007 (*k*=13), 2008-2014 (*k*=13), and 2015-2021 (*k*=13).

The first design feature to consider was the 'origins of WMC tests'. Figure 2 shows a trivial change in how the WMC tests were designed and employed in the three-time

spans. For instance, in the three consecutive time spans, 69.23% (*k*=9), 53.85% (*k*=7), and 46.15% (*k*=6) of tests employed were 'already developed'. This evinces that as time passed, there grew a tendency not to employ 'already developed' WMC tests. In the other categories of origins of WMC tests, no noticeable pattern of change was detected.

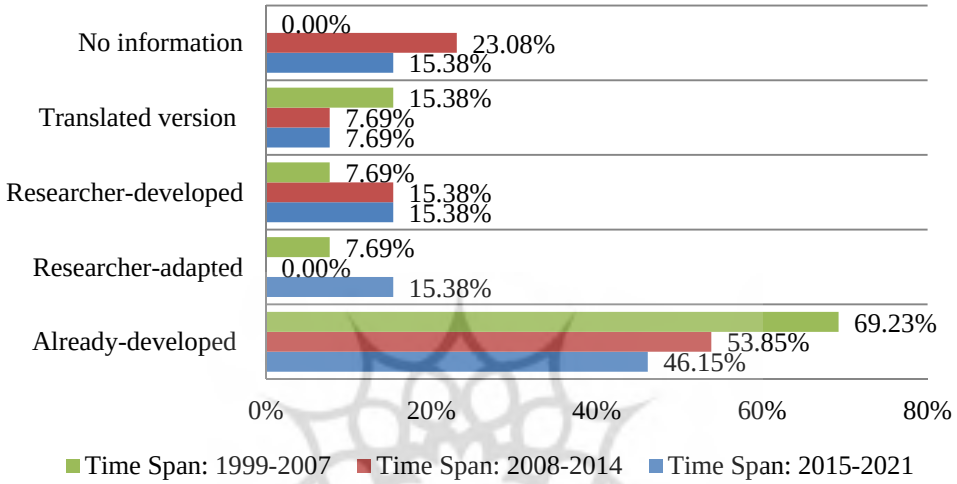

Figure 2. *Origins of WMC Tests*

As for the WMC test type, Figure 3 portrays two conspicuous changes. First, a reduction in the use of reading span tests was experienced as a function of the passage of time. In rank order, in 1999-2007 period, 92.31% (*k*=12), in 2008-2014 span, 53.85% (*k*=7), and in 2015-2021 period, 46.15% (*k*=6) of tests were 'reading span tests'. Second, of the 13 WMC tests employed in 2015-2021 span, 38.46% (*k*=5) were 'operation span tests', illustrating a marked increase in the use of such tests.

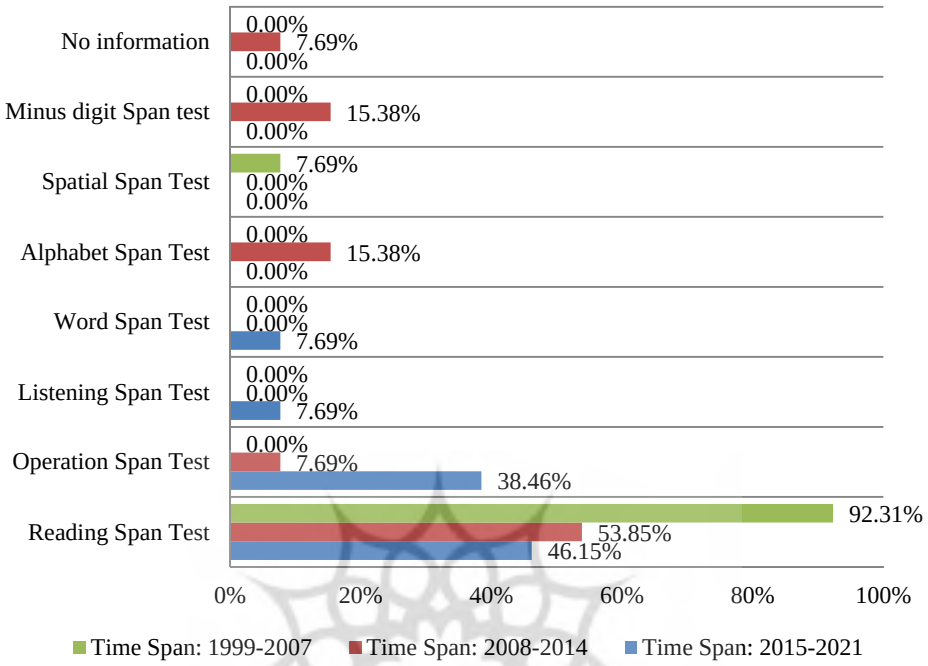

Figure 3. *WMC Test Type*

As illustrated in Figure 4, the language of WMC tests in all time spans was dominantly English. While in 1999-2007, 76.92% (k=10) of tests were in English, in 2008-2014, tests in English experienced a small reduction (*k*=9, 69.23%). But in 2015-2021 span, only 38.46% (*k*=5) of experiments employed WMC tests in English. In the same time span, 30.77% (*k*=4) of experiments employed 'operation span tests', which is language neutral.

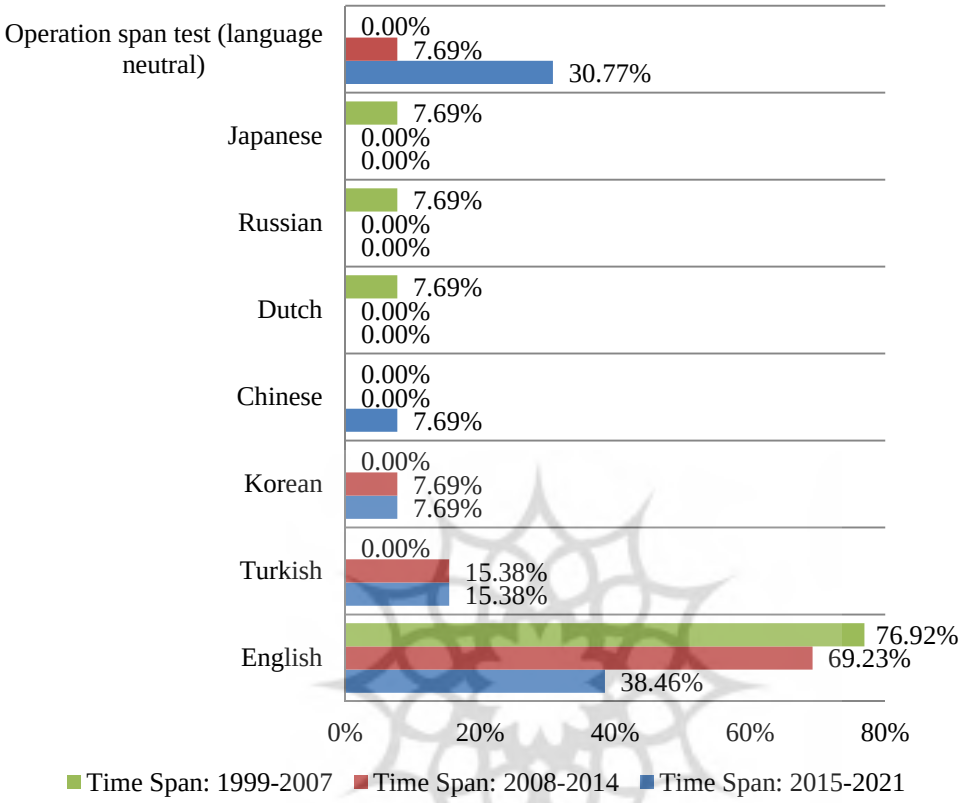

Figure 4. *Language of WMC Tests*

Figure 5 shows a marked change in how the WMC tests addressed sentence length and complexity (Omaki, 2005). In the 1999-2007 period, 23.08% (k=3) of tests addressed sentence length, while 30.77% (k=4) of tests addressed both sentence length and complexity. This picture experienced a total change in the other two time spans. In the 2008-2014 span, sentence length was not addressed, but both sentence length and complexity were addressed in 15.38% (k=2) of tests. Surprisingly, in the 2015-2021 period, sentence length was considered only in 7.69% (k=1) of experiments, and sentence length and complexity were not considered.

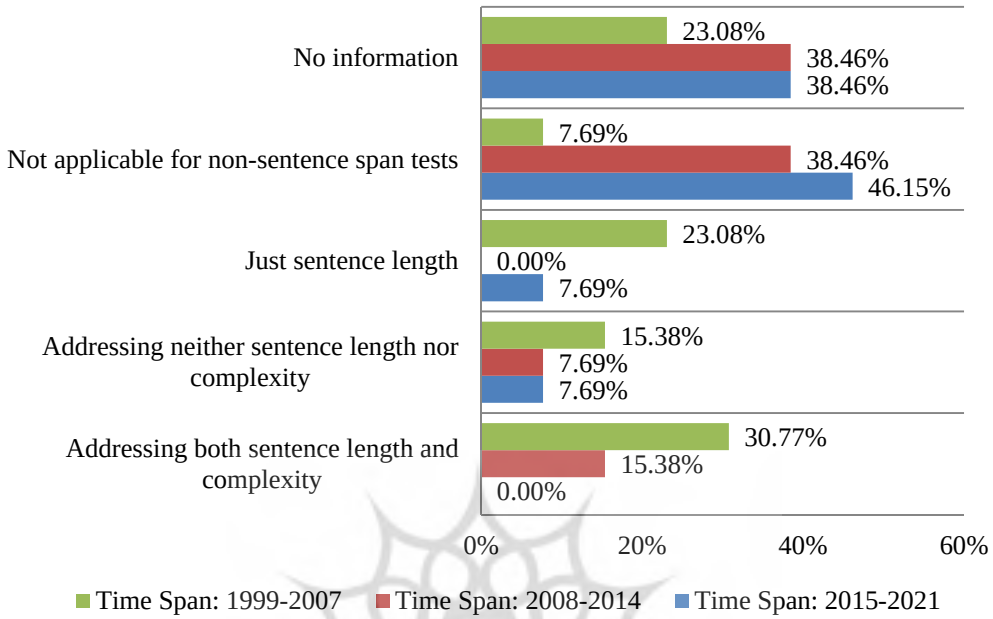

Figure 5. *Addressing Sentence Length and Complexity*

As portrayed in Figure 6, in the 1999-2007 span, more regard was afforded to the use of 'words' *after* sentences (*k*=3, 23.08%), in *final* (*k*=4, 30.77%) and *non-final* (*k*=4, 30.77%) positions. In the 2015-2021 span, this attention switched to the use of letters after sentences or equations (*k*=8, 60.54%, see Unsworth et al., 2005). No conspicuous change was observed in the design of other types of recall tasks.

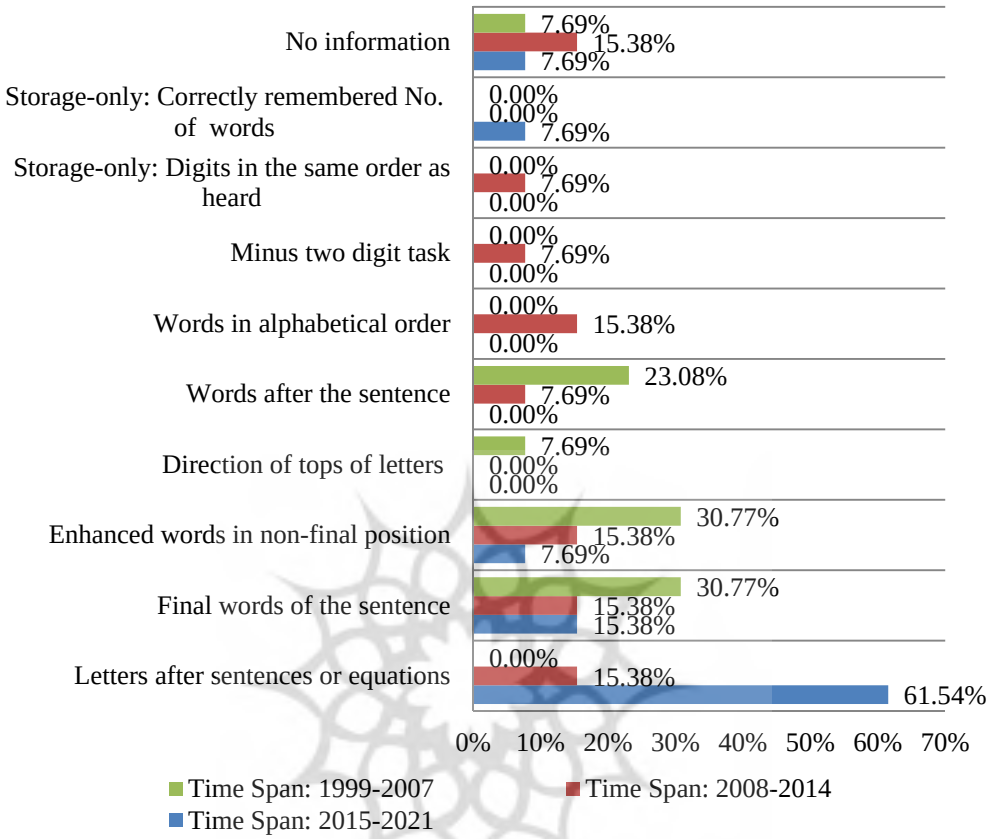

Figure 6. *Recall Task Type*

As for how the processing component has been engaged in all the time spans, the most widely used technique has been the use of comprehension questions (Figure 7). To engage the processing component, in the 1999-2007 span, the majority of the experiments (*k*=8, 61.54%) employed only 'comprehension questions'. This amount experienced a sharp decline for the period of 2008-2014 (*k*=4, 30.77%), but it remained the most widely used technique in this span. Also, in the 2015-2021 period, comprehension questions were most frequent (*k*=6, 46.15%). The second most widely used strategy in 1999-2007 was 'reading aloud' (k=3, 23.08%), but in the 2015-2021 span, attention shifted to using 'reading aloud and comprehension questions' (*k*=5, 38.46%).

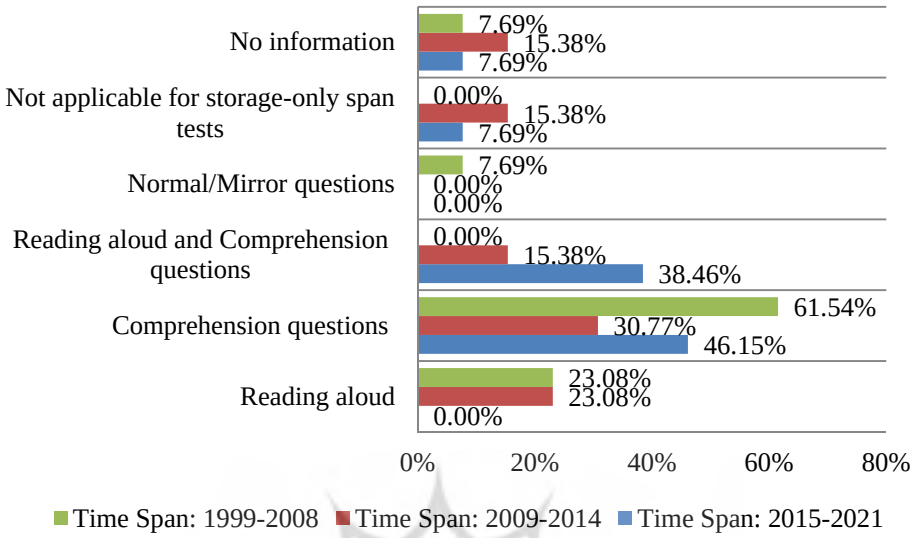

Figure 7. *Engaging the Processing Component*

As for whether the storage and processing components of WM are (non-) simultaneously engaged (Omaki, 2005). Figure 8 shows no noticeable change between 1999-2007 and 2008-2014. The great change was introduced between the time spans of 2008-2014 and 2015-2021. While in the 2008-2014 span, the storage and processing components of WM were equally engaged both simultaneously and non-simultaneously (k=5, 38.46%), in the 2015-2021 period, these two components were engaged differentially, with non-simultaneous engagement receiving more weight (k=10, 76.92%) than simultaneous engagement (k=1, 7.69%).

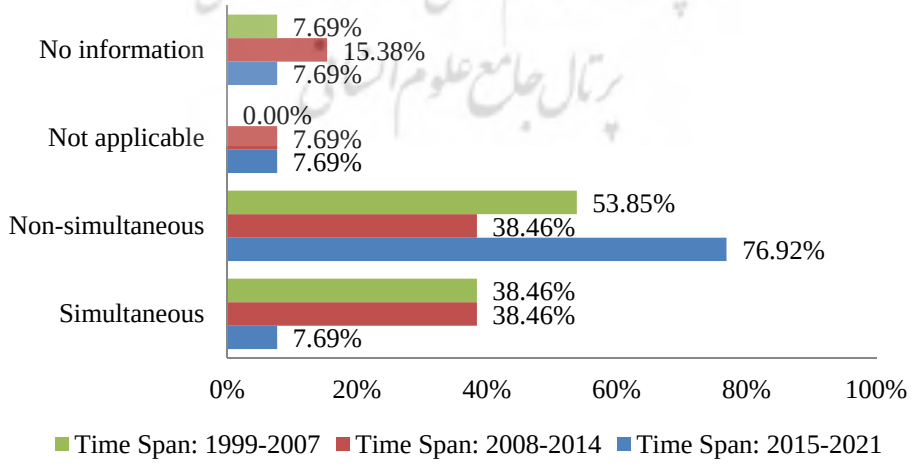

Figure 8. *Engaging both Storage and Processing Components of WM*

As for the presence of practice items (Figure 9), while in the 1999-2007 span, 61.54% (*k*=8) of experiments included practice items, in the 2008-2014 period, only 15.38% (*k*=2) of experiments incorporated practice items. However, this decline was compensated for in 2015-2021, in which 53.85% (*k*=7) of experiments included practice items.

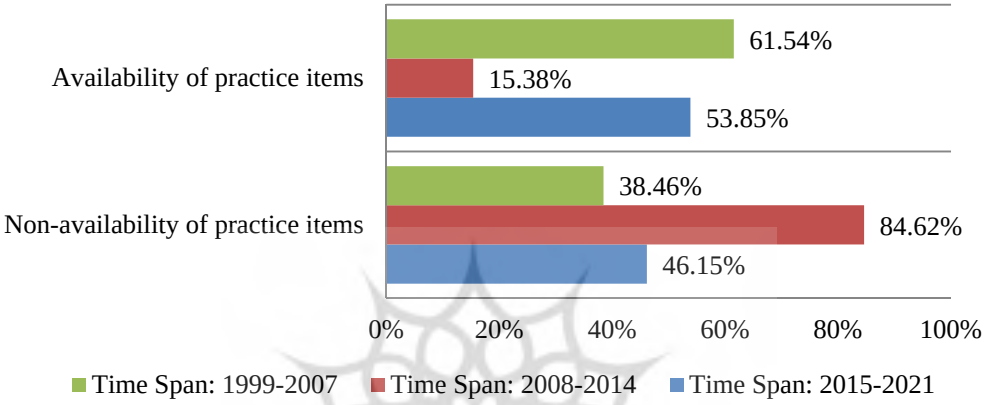

Figure 9. *Availability of Practice Items*

Regarding the number of sets, items, and elements, as charted in Figure 10, the 1999-2007 span provided the most detailed information (*k*=12, 92.31%), and the 2008-2014 span provided the least amount of information (*k*=9, 69.23%). However, this disregard is compensated for in the 2015-2021 period when 84.62% (*k*=11) of experiments provided such information.

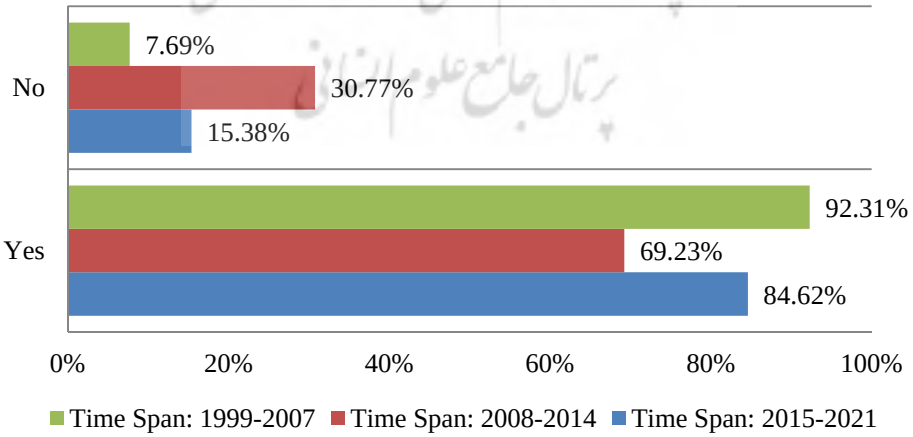

Figure 10. *Information on the Number of Sets, Items and Elements*

To tap the comprehension of the participants, as shown in Figure 11, in the 1999-2007 span, 61.54 (*k*=8) of experiments employed 'True/False questions', while this number declined to 46.15% (*k*=6) for the span 2008-2014. This change seems to be related to the fact that in the latter time span, no information was provided for comprehension question type in 46.15% (*k*=6) of experiments. However, in the 2015-2021 span, a noticeable rise can be witnessed in recounting and the employment of the True/False question type. This time span reported the employment of True/False question types in 84.62% (*k*=11) of experiments.

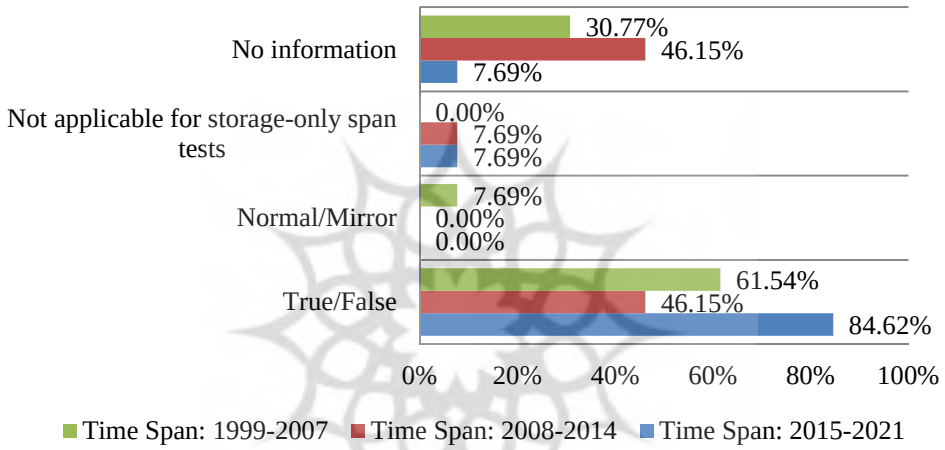

Figure 11. *Comprehension Question Type*

As indicated previously, some experiments included questions with (roughly) equal number of true and false responses to avoid positive response tendency. As shown in Figure 12, the time span of 1999-2007 ranked first in heeding to this feature (*k*=8, 61.54%), followed by the period of 2015-2021 (*k*=6, 46.15%). In the 2008-2014 span, this feature was least attended to in 7.69% (*k*=1) of experiments and not applicable for 30.77% (k=4) of experiments. It was not applicable to the 'Word Span Test', 'Alphabet Span Test', 'Minus digit Span test', and 'Spatial Span Test' for which the True/False dichotomy does not apply.

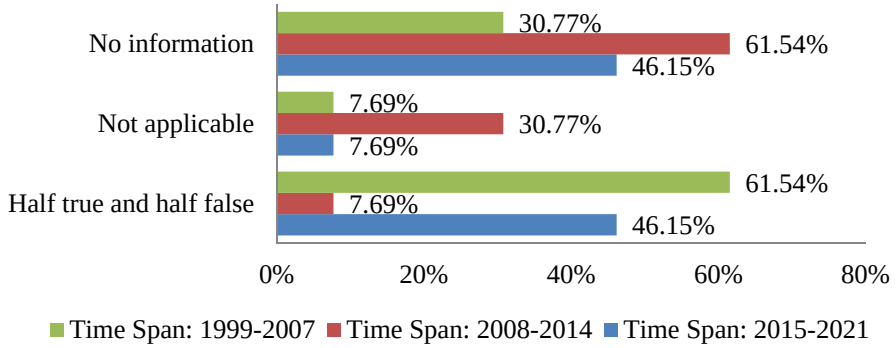

Figure 12. *Avoiding Positive Response Tendency*

Now, it is time to see what changes occurred in the administration features of the WMC tests. Concerning the 'presentation instrument', as displayed in Figure 13, there occurred great diversity in the use of different presentation instruments for the WMC tests across all time spans, but no specific pattern of change was observed as the choice seems to be more idiosyncratic. However, a tendency towards the use of more user-friendly software and instruments (i.e., Microsoft PowerPoint and Screen) could be seen.

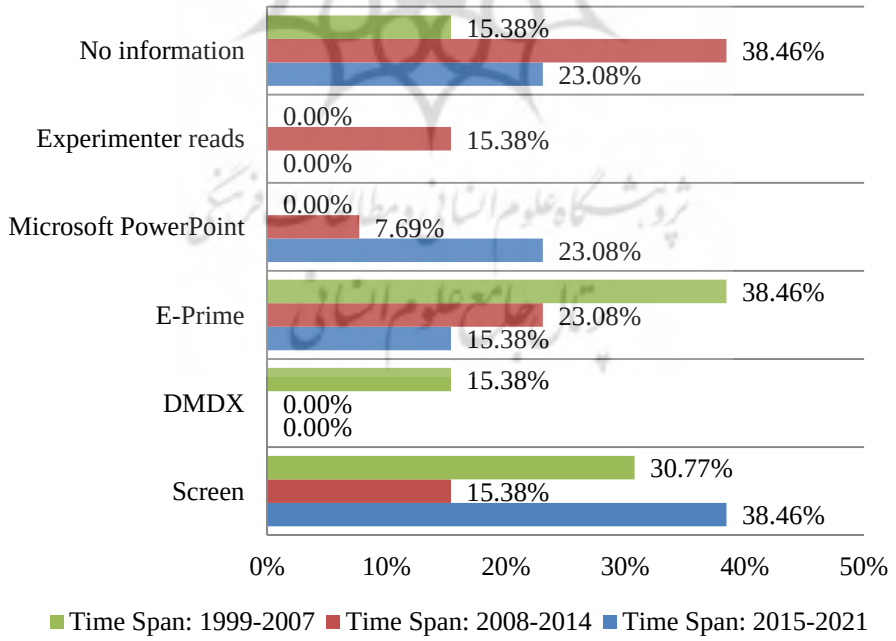

Figure 13. *Presentation Instrument*

As shown in Figure 14, in the 1999-2007 period, primary attention was afforded to presenting entire elements at once (*k*=6, 46.15%), followed by attention to presenting elements noncumulatively (*k*=4, 30.77%). All presentation types gained equal prominence in the 2008-2014 span (*k*=2, 15.38%). However, the time span of 2015-2021 paid attention only to presenting entire elements at once (*k*=8, 61.54%).

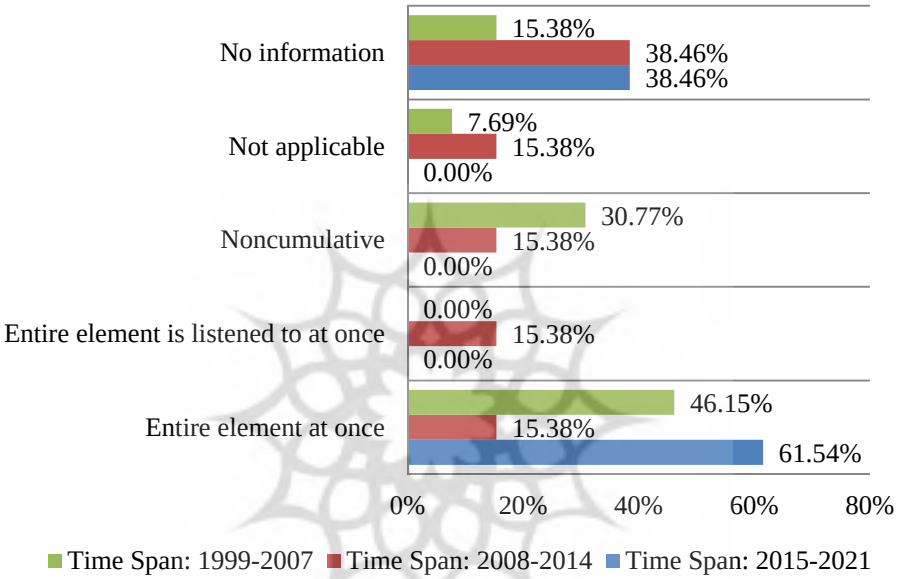

Figure 14. *Presentation Type*

Apropos of the presentation order of the elements (Figure 15), no change was detected between the time spans of 1999-2007 and 2008-2014. But a change occurred in the 2015-2021 span: In 30.77% (*k*=4) of experiments, elements were presented randomly, while in the previous two-time spans, no random presentation was employed (James et al., 2018).

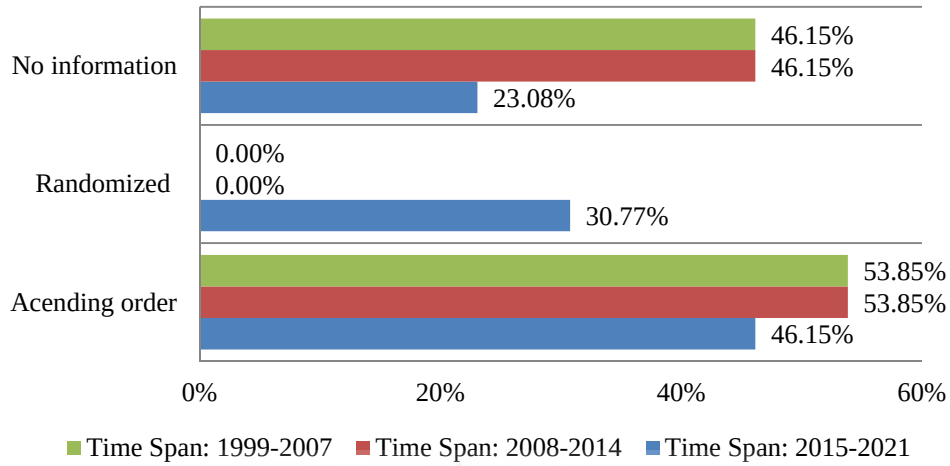

Figure 15. *Presentation Order*

As shown in Figure 16, in the 1999-2007 period, only 15.38% (*k*=2) of experiment elements of WMC tests were counterbalanced. Counterbalancing was not employed in 30.77% (*k*=4) of experiments in the periods of 1999-2007 and 2008-2014 (Swets et al., 2007). Also, in the most recent time span, 46.15% (*k*=6) of experiments did not use counterbalancing of elements. It seems that as time passes, researchers are convinced by Swets et al.'s (2007) argument for not counterbalancing test elements.

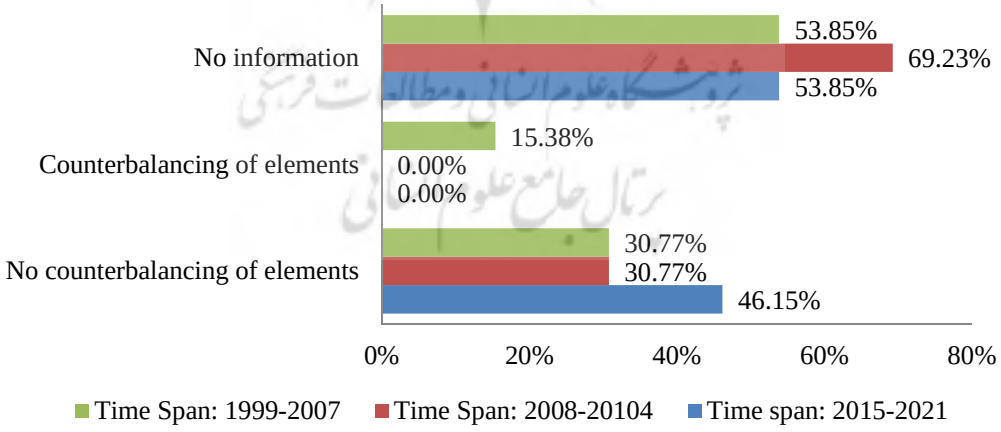

Figure 16. *Counterbalancing of Elements*

As for the administration feature of test discontinuation (Figure 17), in 1999-2007 span, in 46.15% (*k*=6) of experiments, it was directly stated that the administration of the

WMC test was not discontinued while in the 2008-2014 period, seemingly the test was not discontinued at all (*k*=0, 0.00%) based on any criteria. But in the 2015-2021 span, 38.46% (*k*=5) of experiments provided information that the test was not discontinued. As for the discontinuation of the WMC tests, no significant change was seen (James et al., 2018).

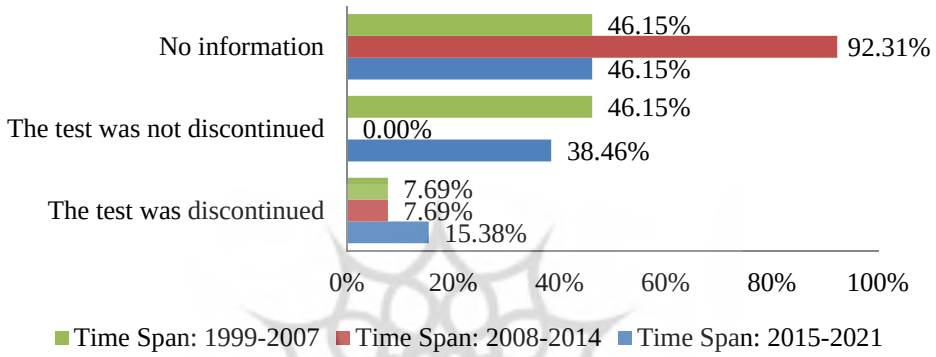

Figure 17. *Test Discontinuation*

Considering the fact that there were only four experiments in which the test was discontinued, in two of the cases (15.38%), the criterion is 'making two or more mistakes in a set'; for one (7.69%), it is 'failing to recall the words from two consecutive items' and for the other one (7.69%), it is 'failing all three items in a set' (Table 5). In other cases, it was either not applicable or no information was provided.

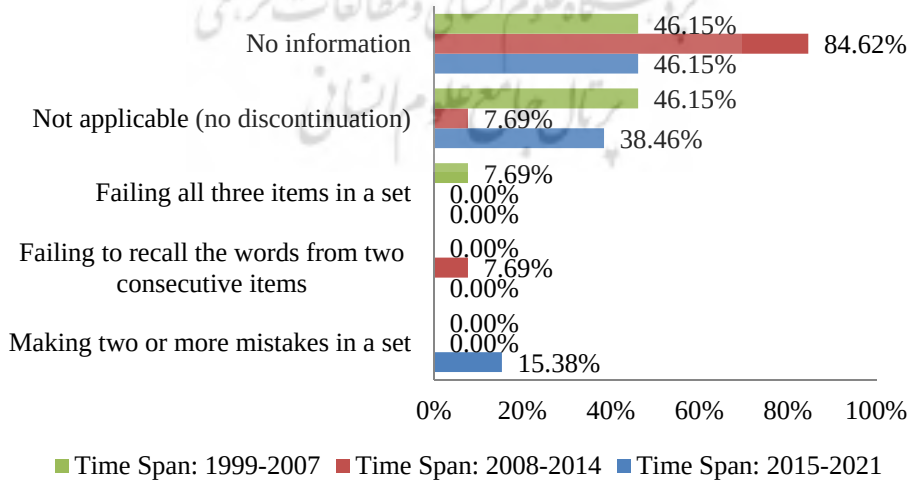

Figure 18. *Test Discontinuation Criteria*

Figure 19 exhibits a noticeable change in how element presentation was timed. Element presentation was timed for the whole group in 53.85% (*k*=7) of experiments in the 1999-2007 period, in 15.38% (*k*=2) of experiments in the 2008-2014 span, and in 23.1% (*k*=3) of experiments in 2015-2021 period. Also, element presentation was individually calibrated in 23.1% (*k*=3) of experiments in the 2015-2021 span. Moreover, in the 1999-2007 span, in 46.15% (*k*=6) of experiments, 'no information' is provided in this regard. However, a conspicuous rise can be witnessed for the period of 2008-2014 (*k*=11, 84.62%). This is, however, followed by a decline for the 2015-2021 span (*k*=7, 53.85%).

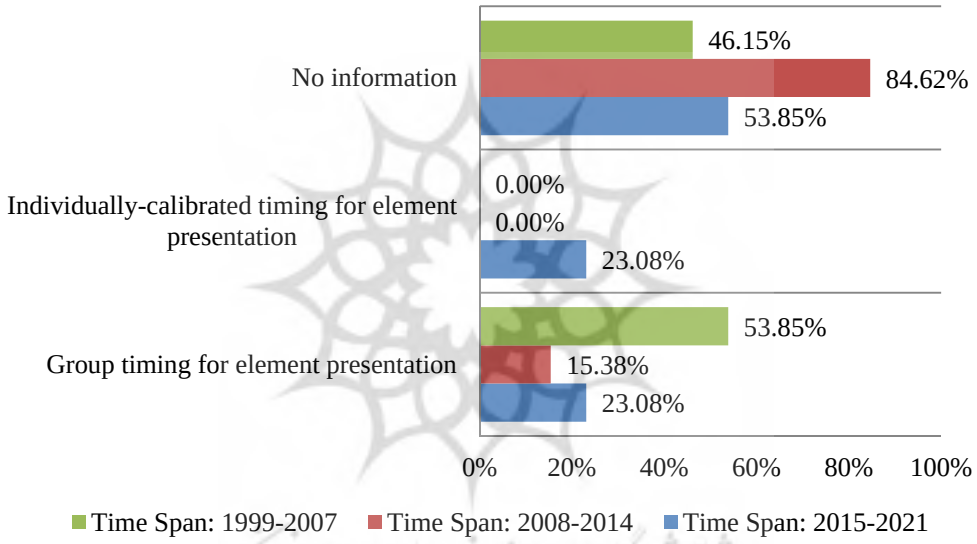

Figure 19. *Timing for Element Presentation*

As for whether any changes were made across the time spans on the timing for the processing accuracy task, Figure 20 displays two moderately noticeable changes. First, this timing was individually calibrated only in the 2015-2021 span (*k*=3, 23.08%). Second, while in the 1999-2007 period, 30.77% (*k*=4) of experiments included timing for administering the test, much less attention was devoted to this timing in the periods of 2008-2014 (*k*=1, 7.69%) and 2015, 2021 (*k*=2, 15.38%).

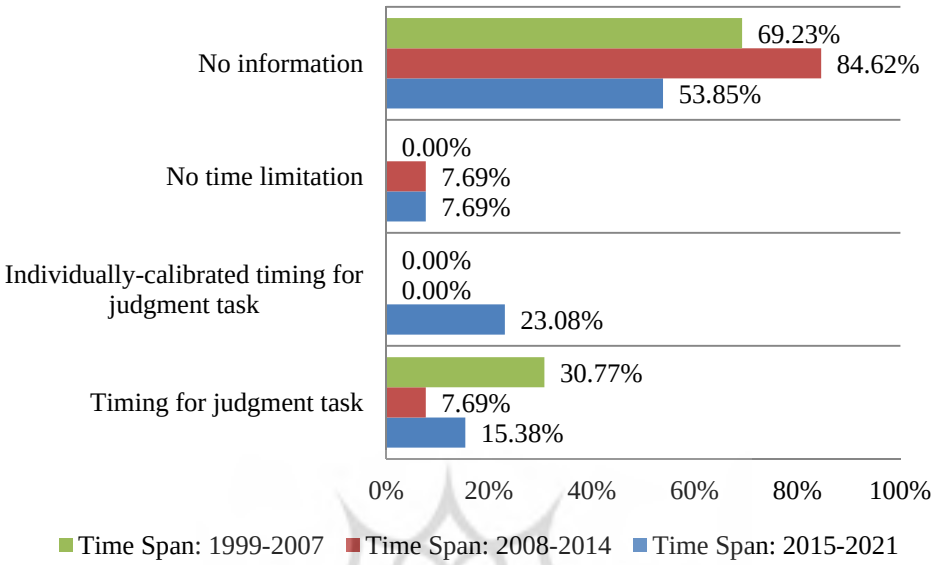

Figure 20. *Timing for the Processing Accuracy Task*

As seen in Figure 21, in the 1999-2007 span, the timing for the recall task was considered in 23.08% (k=3) of experiments. This experienced a reduction down to one experiment (7.69%) in the 2008-2014 span and no experiment in the 2015-2021 period. Moreover, in the 1999-2007 span, 30.77% (*k*=4) of experiments were not timed for their recall task. In the other two time spans, no information was provided for the rest of the experiments as to whether their recall tasks were timed or not.

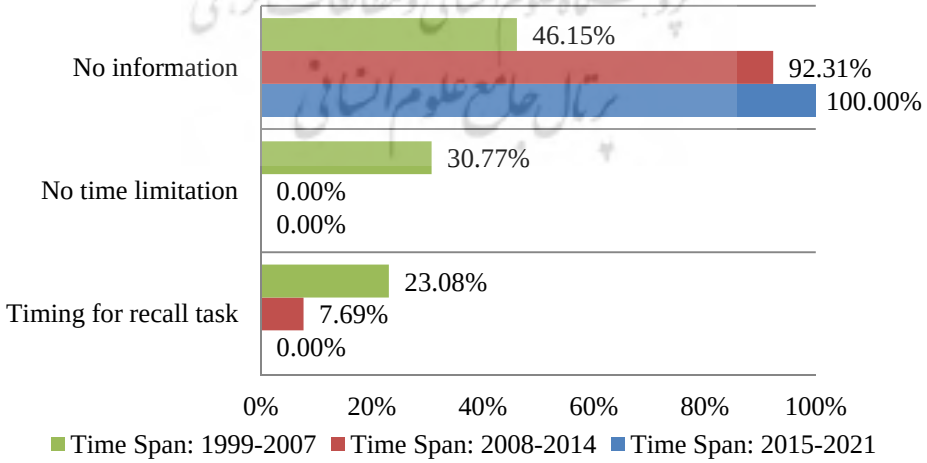

Figure 21. *Timing for Recall Task*

Figure 22 illustrates answers to comprehension questions were recorded with great diversity. In the 1999-2007 span, only 23.08% (*k*=3) of experiments had the participants write their responses to comprehension questions on an answer sheet and only 15.38% (*k*=2) used some software. Also, in the 2008-2014 period, in 30.77% (*k*=4) of experiments, 'the experimenter wrote the responses on an answer sheet' and only in 7.69% (*k*=1) of experiments some software was used. What is noticeable in Figure 22 is that in the 2015-2021 span, researchers have frequently employed some software to record the participants' responses to comprehension questions (*k*=6, 46.15%), and only in 23.08% (*k*=3) of experiments, the experimenters themselves have recorded the responses. But still, in this time span, in 30.77% (*k*=4) of experiments, they have provided 'no information' in this regard.

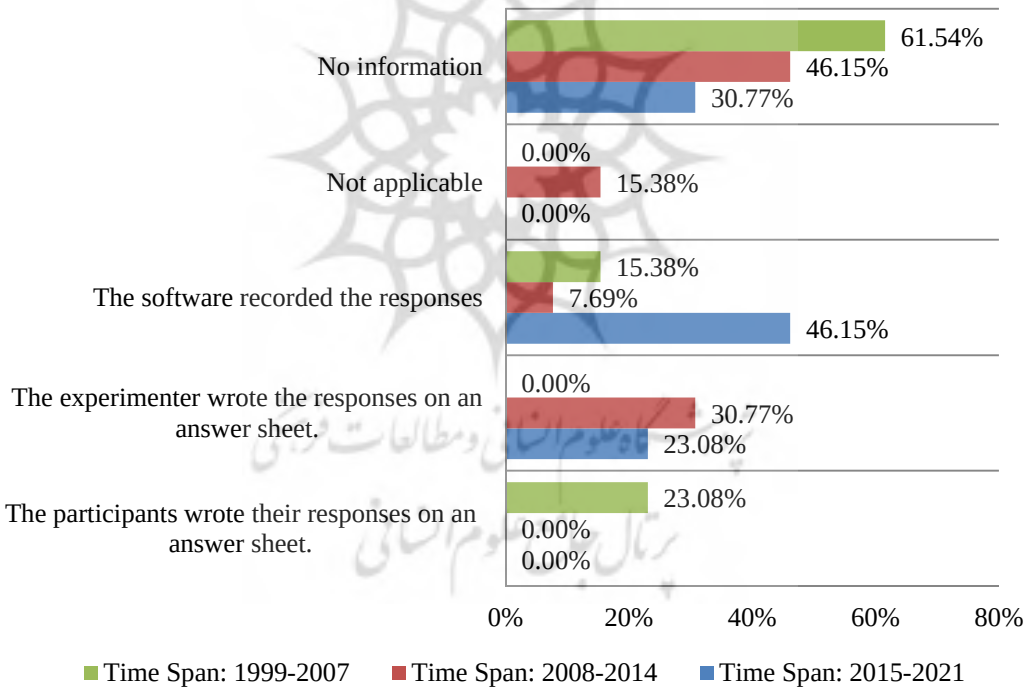

Figure 22. *Recording Answers to Comprehension Questions*

In order to record answers to recall tasks (Figure 23), the experiments in the 1999-2007 period mainly had the participants write their answers down on sheets of paper (*k*=7, 53.85%) or in some experiments, the experimenter himself/herself wrote the responses down (*k*=3, 23.08%). These two strategies were employed in the 2008-2014 span, but

with a decrease in the number of experiments: 23.08% (*k*=3) for the former strategy and 15.38% (*k*=2) for the latter strategy. In the 2015-2021 span, the strategy of 'having the participants type their responses' (*k*=3, 23.08%) was added to the two previous strategies of 'having the participants write down their responses' (*k*=2, 15.38%) and 'having the experimenter write down the responses (*k*=3, 23.08%). However, still, 38.46% (*k*=5) of experiments did not provide any information in this respect.

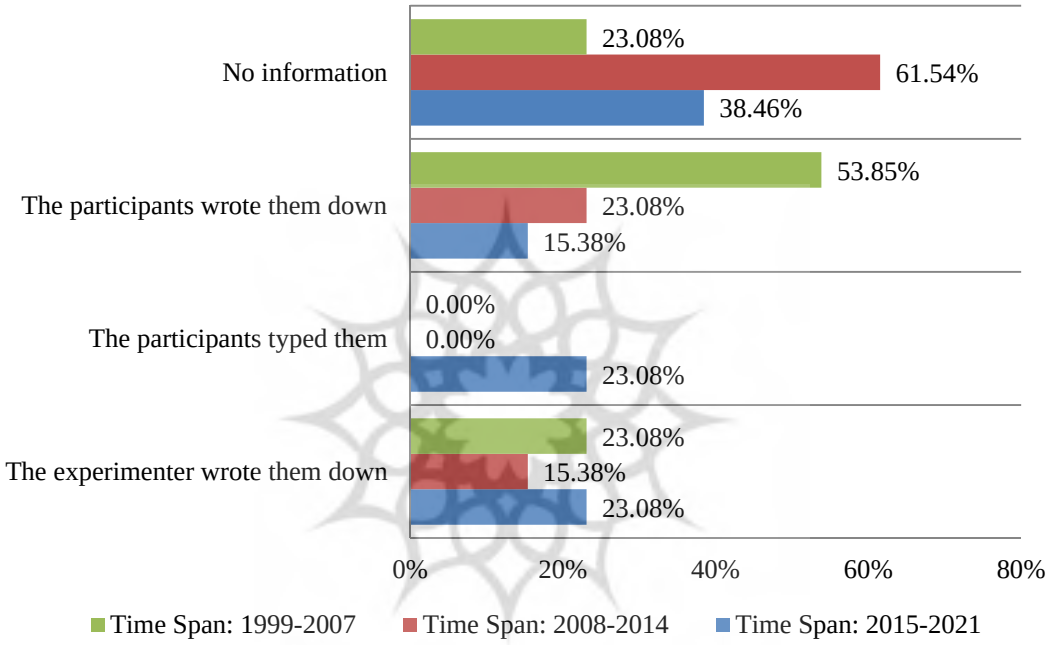

Figure 23. *Recording Answers to Recall Tasks*

Figure 24 shows whether the WMC tests were administered in the same session as the RC attachment task and, if yes, whether the test was administered before or after the RC attachment task. But unfortunately, even in the recent time span of 2015-2021, 53.85% (*k*=7) of experiments provided no information in this respect, but the tendency to administer the tests in two different sessions has increased.

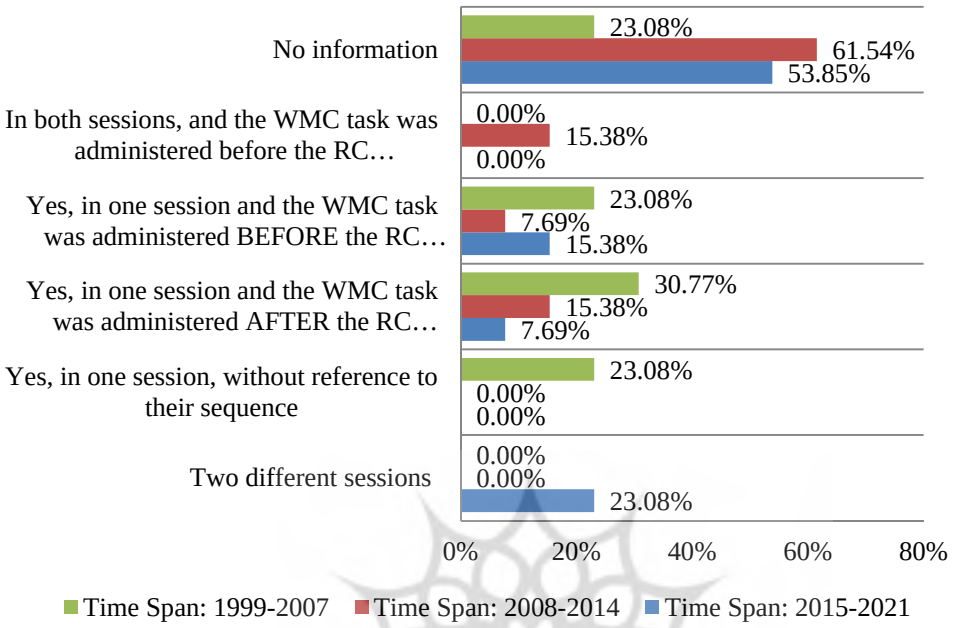

Figure 24. *Is the WMC Measure Administered in the Same Session as the RC Attachment Task?*

Figure 25 portrays a trend towards the employment of more individual-based test administration as time passed, with the period of 2015-2021 having the most individual-based test administration (k=11, 84.62%) of experiments.

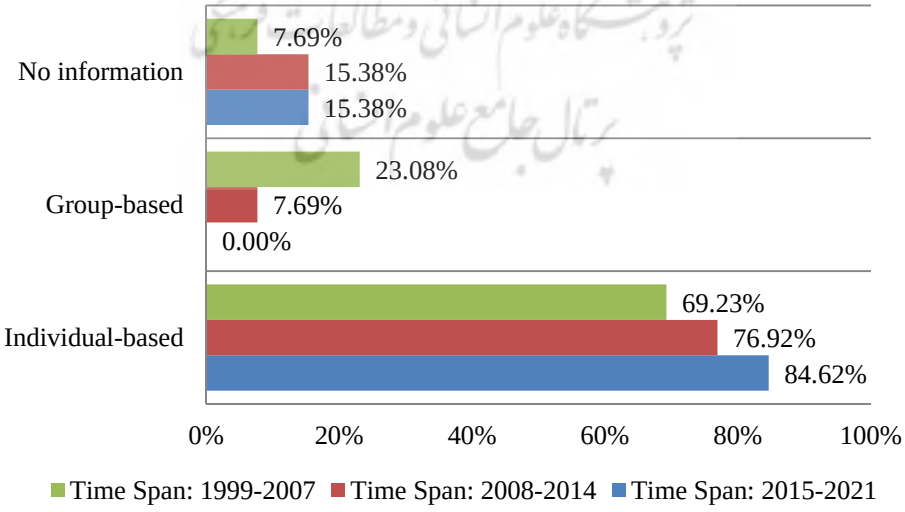

Figure 25. *Individual-Based Vs. Group-Based Administration*

Concerning reliability investigation, as portrayed in Figure 26, in all the three-time spans, insufficient or, no attention was committed to investigating reliability; however, the time span of 2015-2021 held the lead (k=4, 30.77%) in reporting reliability as compared with the other two time spans. A note is in order: 'reliability not reported' refers to occasions when reliability is discussed in the studies but not reported, but 'no information' refers to when reliability information is neither discussed nor reported.

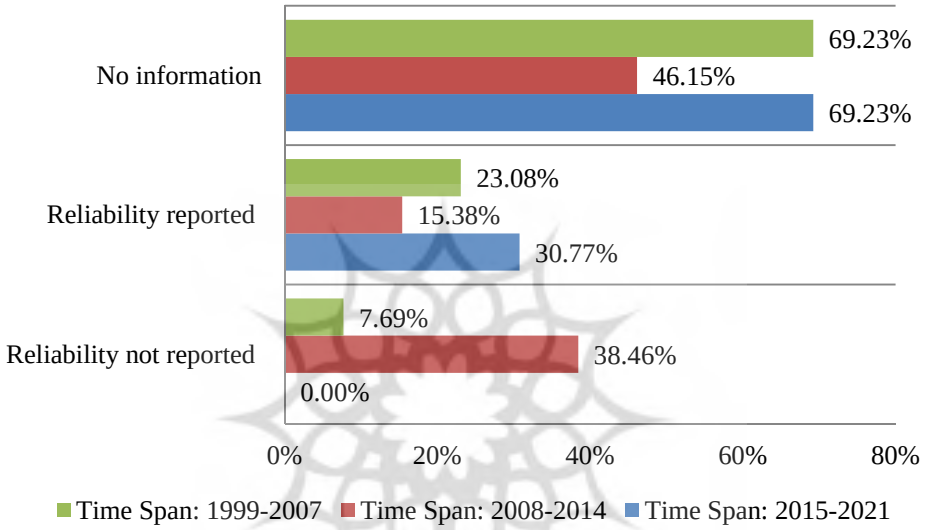

Figure 26. *Reliability Reported*

Figure 27 displays detailed information on the scoring system used. However, due to the complexity of the comparison, we provide further figures on every distinct feature below (Leeser & Sunderman, 2016).

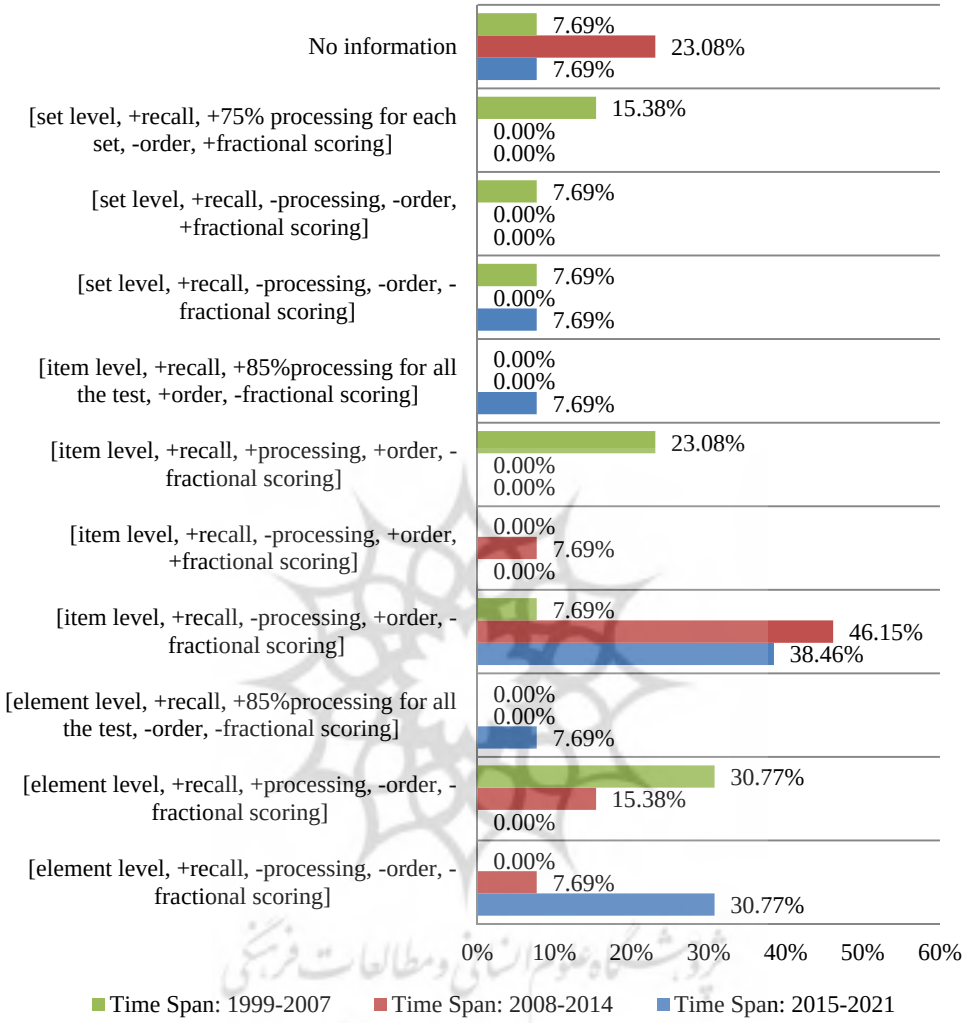

Figure 27. *Scoring System*

Regarding the scoring level (i.e., whether scores are credited to elements, items, or sets), noticeable diversity was observed in the three-time spans (Figure 28). In the 1999-2007 span, equal regard was afforded to set, item and element level scoring (k=4, 30.77%). But this changed perceivably in the time span of 2008-2014, in which period set level scoring was not used at all, and instead, more attention was devoted to 'item level scoring' (k=7, 53.85%), and a comparatively smaller amount of interest was shown in 'element level scoring' (k=3, 23.08%). This latter pattern of attention did not change drastically for the period of 2015-2021. Interestingly, set level scoring was employed in

only one experiment (Dai, 2015). However, more regard was afforded to employing 'item level scoring' (*k*=6, 46.15%) and 'element level scoring' (*k*=5, 38.46%).

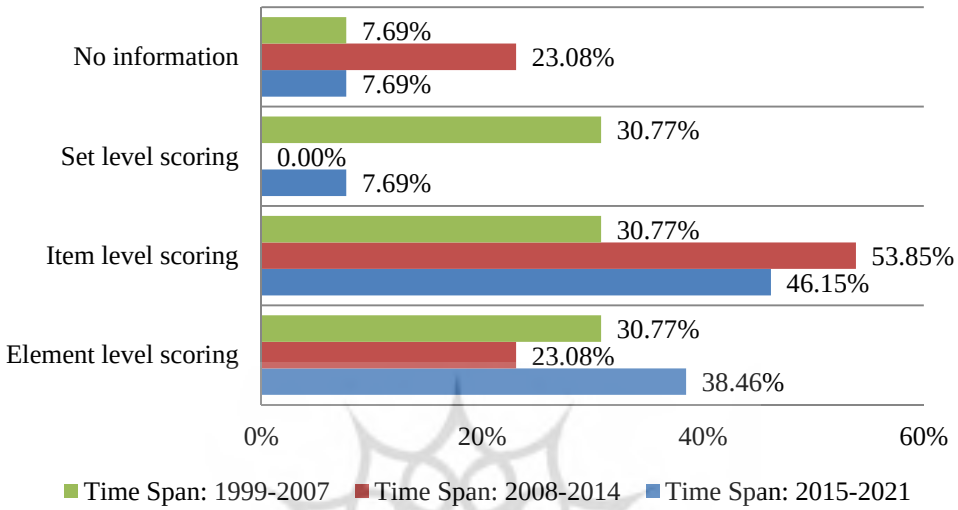

Figure 28. *Scoring Level*

Besides a recall score granted to performance in WMC tests, sometimes 'processing accuracy' was heeded for crediting a score to an individual's WMC. As can be seen in Figure 29, a trend toward disregarding such a scoring system was observed. In fact, while in the 1999-2007 span, 53.85% (*k*=7) of experiments considered crediting scores only when a set, item or element was processed accurately, this feature of the scoring system was attended to only in 15.38% (*k*=2) of experiments in 2008-2014 period and totally disregarded in 2015-2021 span. However, some experiments considered a threshold level of processing accuracy for either the whole test or a set. For instance, in the 1999-2007 period, 15.38% (*k*=2) of experiments considered 75% processing accuracy for each set. However, this feature was not regarded in 2008-2014 but was considered in two experiments (15.38%) in the 2015-2021 period, with a higher threshold level (i.e., 85%).

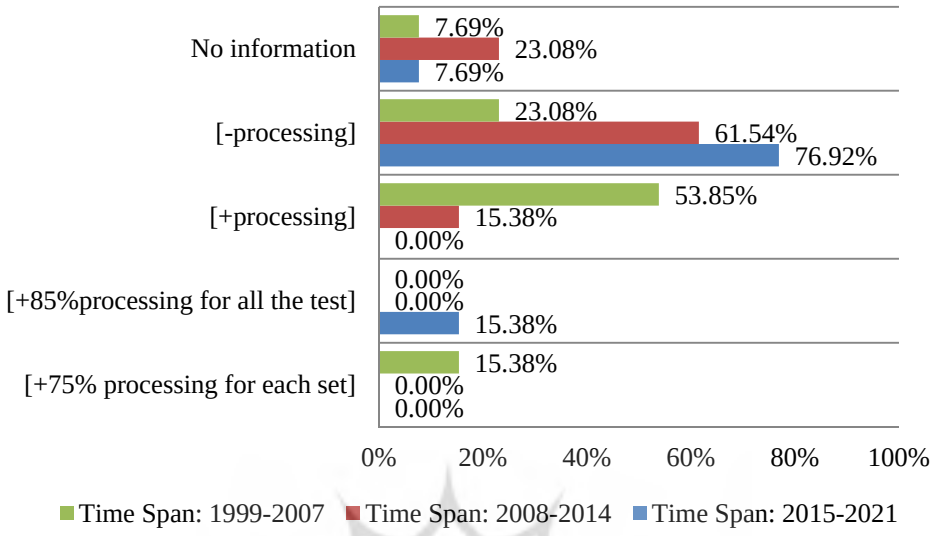

Figure 29. *Scoring of Processing Accuracy*

Another feature that received scores was the order of recall. Some researchers allocated scores to the to-be-recalled information only when such information was recalled in the correct serial order, which might have influenced the way the participants were categorized as high spans and low spans. As seen in Figure 30, in 1999-2007, 30.77% (*k*=4) of the thirteen experiments considered recall order for scoring. This feature was attended more rigorously in 2008-2014 (k=7, 53.85%), and in 2015-2021 (*k*=6, 46.15%).

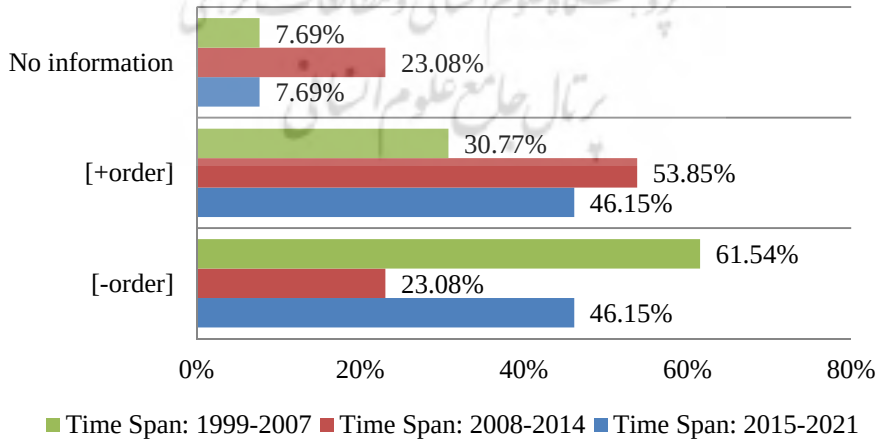

Figure 30. *Scoring Order of Recall*

As for the fractional scoring of partially completed sets, items or elements, as portrayed in Figure 31, in the 1999-2007 span, researchers allocated fractional (i.e., partial) scores in 23.08% (*k*=3) of experiments. This, however, declined to one experiment (7.69%) in the span time of 2008-2014 and no experiment in the span time of 2015-2021.

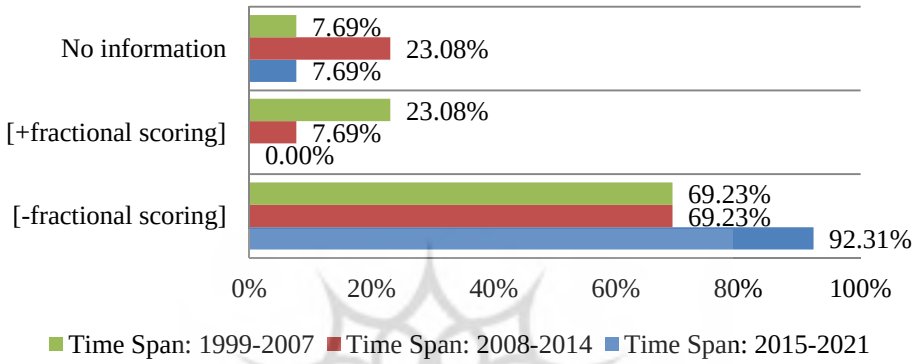

Figure 31. *Fractional Scoring*

As for 'measurement scale of WMC, illustrated in Figure 32, throughout time spans 1999-2007 and 2008-2014, WMC was treated almost equally (53.85% vs. 46.15%) as nominal or continuous. But as time passed, this variable was much less (*k*=4, 30.77%) treated as a nominal variable.

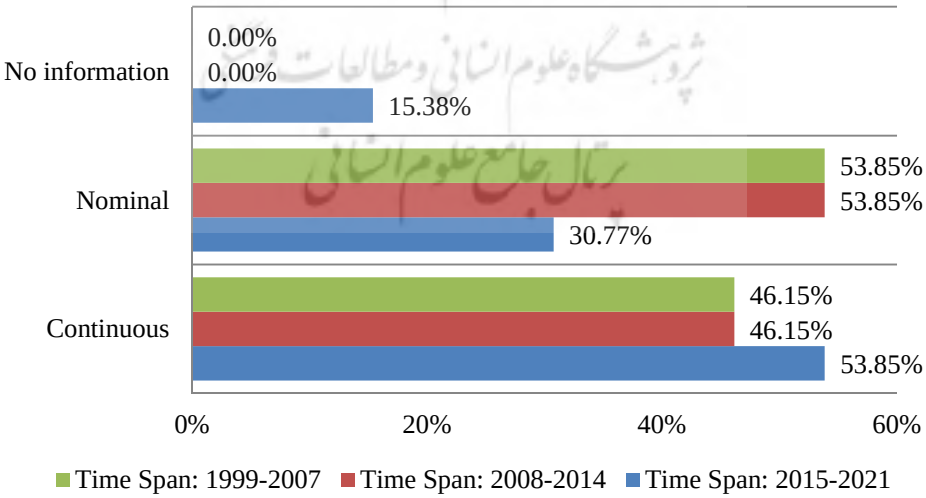

Figure 32. *Measurement Scale of WMC*

Obviously, when we decide on a cut-off score, we treat the variable of interest as a nominal one. So, this feature is not applicable when we consider WMC as a continuous variable. As shown in Figure 33, in the 2015-2021 span, WMC was considered a continuous variable, with no exception. But as for WMC as a nominal variable, the cut-off score was specified in three different ways, with 'mean as the cut-off point' being the most-widely used strategy (*k*=4, 30.77%).

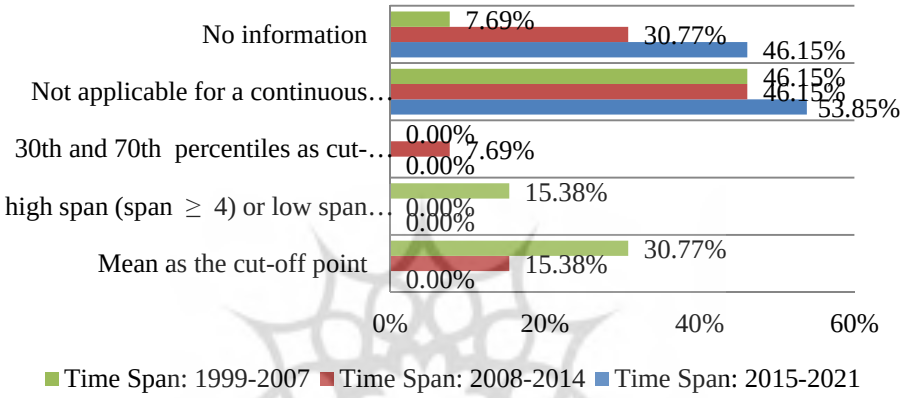

Figure 33. *Cut-Off Score for Categorizing Participants into WMC Levels*

Discussion

A burgeoning growth of concern in methodological quality and transparency can be conspicuously seen in the expanding number of published methodological syntheses (e.g., Crowther et al., 2021; Derrick, 2016; Liu & Brown, 2015; Marsden, Thompson et al., 2018; Plonsky et al., 2020; Sok et al., 2019). Such research endeavors are "lively testimony to the fact that methodologies no longer have ancillary status in our work" (Byrnes, 2013, p. 825) and indicative of the fact that scholarly awareness in the significance of methodological issues of quality and transparency in research is escalating.

Having taken cognizance of this significance, we ran the current study to shed light on the methodological features and methodological transparency of WMC measures employed in the literature of RC ambiguity resolution, as well as the changes or improvements that have been made across the time spans of 1999-2007, 2008-2014, and 2015-2021. To these ends, by adopting a retrospective approach, this study described the design, administration and scoring features of the WMC measures, depicted how transparency was addressed in reporting the associated methodological issues and

revealed in what ways these measures have changed throughout time. The findings revealed cross-study fluctuations in the design, administration and scoring of WMC measures. Such methodological variations seem to affect the substantive results and accompanying interpretations (Boegle et al., 2021), which in turn may lead the researchers to support or refute a specific theory of sentence processing. Moreover, such methodological variations confirm that WMC has been operationalized and measured differently across studies, which may eventuate in variation in the effect of WMC on the substantive results of RC ambiguity resolution studies. Therefore, a much-needed standardized WMC measure is called for to scrutinize whether, in the presence of such standardization, substantive results still remain incongruent.

Results also revealed a number of improvements in the tests, including the use of (a) comprehension questions, instead of reading aloud alone, to tap the processing component, (b) non-counterbalanced items to eliminate the effect of individual differences other than WMC, (c) individually-calibrated timing, for more accurate measurement of individual differences in WMC, (d) non-final nouns to simultaneously tax the processing and storage components of WM, (e) particular, selected structures to control for sentence length and complexity, and (f) (roughly) equal number of True/False responses for comprehension questions to avoid the positive response tendency. Results also showed relatively high transparency in reporting the 'design' features (M = 79.12%), moderate transparency in reporting the 'administration' features (M = 51.48%), and a moderate-high transparency level in reporting the 'scoring' features (M = 73.87%) of WMC tests.

Furthermore, to achieve comparability of results and replicability of experiments (Marsden, 2020), WMC materials and procedures should be as transparently reported as possible. Regarding comparability and replicability, a study by Makel and Plucker (2014) evidenced that in research replications with at least one author overlap, 88.7% of replications supported previous findings, while in replications with no author overlap, only 70.6% of previous findings was supported. This discrepancy in the findings of replications with and without author overlap, is surmised to have occurred due to the lack of transparency of methodological issues and reporting practices in initial research (Marsden, 2020). Similar results were obtained by Marsden, Morgan-Short et al. (2018). They found a positive, significant relationship between replication studies with 'author overlap' with the initial study (90% support) as compared to occasions in which no author overlap existed in replication studies (59% support). Moreover, when materials,

methodological procedures, and practices are not transparently elaborated, the replicability of research and, in turn, the "comparability of findings" (Baumert et al., 2020) may be less probable. Consequently, authors are recommended to follow open access, and open science practice in their reporting of methodological issues (Marsden et al., 2017, Marsden, Plonsky et al., 2018), rendering replication studies and reproducibility and comparability of results more feasible (Bolibaugh & Marsden, 2021).

Furthermore, with the objective of adding a prospective approach, we make a number of recommendations for future primary and secondary research. First, we hope that knowledge gained from this descriptive methodological synthesis directs the attention of scholars towards the standardization of the design, administration, and scoring features of WMC tests so that our understanding of this domain of research could be enriched and more perfect replications and, in turn, more comparability of the substantive results could be achievable. Moreover, although the current synthesis solely focuses on methodological features of WMC tests, it can afford a recommendation for future substantive research: More primary replication studies are called for to compare how and whether differences in methodological design, administration and scoring features of WMC tests would yield differences in results.

Regarding the reporting practices associated with transparency, we make two recommendations. First, when reported, many of the features of WMC tests varied substantially across studies. Hence, researchers are recommended to provide more transparent reporting of their practices to make replication studies easier and more accurate. Doing so can serve to promote a greater understanding of and confidence in the methods and findings. Second, since methodological transparency is an indicator of methodological rigor (Derrick, 2016), most researchers address it when doing methodological syntheses. Therefore, scarce attention is directed toward conducting independent syntheses on methodological transparency, which is much called for.

Furthermore, since this study was part of a larger project and since the limits of the larger project did not allow us to investigate the employment of WMC measures in other domains and subdomains of psycholinguistic studies, we did not go beyond synthesizing the methodological features of WMC measures in RC ambiguity resolution. Therefore, we recommend a future, more inclusive methodological synthesis of WMC measures beyond the boundaries of the current study.

Conclusion

Since their introduction in 1980 by Daneman and Carpenter, WMC measures have advanced in variant forms, with each form building upon the shoulders of previous giants such that the more recent forms seem to provide more valid measures of WMC, that is, they seem to be less prone to the two main threats to the internal validity of measures (i.e., construct under-representation and construct irrelevance, Messick, 1989). Fundamentally, these variant forms are not very different from the original ones; nevertheless, some major changes have been introduced to propound more valid WMC measures. Yet, further improvements and standardization in the design, administration and scoring features of these measures seem necessary so that they can provide more valid results and replicability and comparability of the results can be rendered more likely.

Acknowledgments

We would like to thank the editorial team of TESL Quarterly for granting us the opportunity to submit and publish the current synthesis. We would also like to express our appreciation to the anonymous reviewers for their careful, detailed reading of our manuscript and their many insightful comments and suggestions.

Declaration of conflicting interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for this article's research, authorship, and/or publication.

References

The asterisked references are included in the methodological synthesis.

Abdelghany, H., & Fodor, J. D. (1999, September). *Low attachment of relative clauses in Arabic*. Poster presented at AMLaP (Architectures and Mechanisms of Language Processing) (Vol. 99, pp. 23-25). Edinburgh, UK.

Arabmofrad, A., & Marefat, H. (2008). Relative clause attachment ambiguity resolution in Persian. *Iranian Journal of Applied Linguistics*, *11*(1), 29-49. https://ijal.khu.ac.ir/article-1-71-fa.html

Ariji, K., Omaki, A., & Tatsuta, N. (2003). Working memory restricts the use of semantic information in ambiguity resolution. In P. Slezak (Ed.), *Proceedings of the 4th International*

Conference on Cognitive Science. Sydney, Australia: University of New South Wales, pp. 19-25.

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2019). *Introduction to research in education* (10th ed.). Cengage Learning.

*Başer, Z. (2018). *Syntactic priming of relative clause attachment in monolingual Turkish speakers and Turkish learners of English.* Unpublished doctoral dissertation. Middle East Technical University, Turkey. https://open.metu.edu.tr/bitstream/handle/11511/27252/index.pdf

Baumert, A., Buchholz, N., Zinkernagel, A., Clarke, P., MacLeod, C., Osinsky, R., & Schmitt, M. (2020). Causal underpinnings of working memory and Stroop interference control: testing the effects of anodal and cathodal tDCS over the left DLPFC. *Cognitive, Affective, & Behavioral Neuroscience*, *20*(1), 34-48. https://doi.org/10.3758/s13415-019-00726-y

Bigby, M. (2009). The hierarchy of evidence. In H. Williams, M. Bigby, T. Diepgen, A. Herxheimer, L. Naldi & B. Rzany (Eds.) *Evidence-based dermatology* (pp. 34-37). John Wiley & Sons.

Boegle, R., Gerb, J., Kierig, E., Becker-Bense, S., Ertl-Wagner, B., Dieterich, M., & Kirsch, V. (2021). Intravenous delayed gadolinium-enhanced MR imaging of the endolymphatic space: a methodological comparative study. *Frontiers in Neurology*, *12*, 360. https://doi.org/10.3389/fneur.2021.647296

Bolibaugh, C., & Marsden, E. (2021). *Reproducibility and research integrity in applied linguistics*. MetaArXiv. https://doi.org/10.31222/osf.io/3ucbv

Brown, H. D. (2014). *Principles of language learning and teaching* (6th ed.). Longman.

Brysbaert, M., & Mitchell, D. C. (1996). Modifier attachment in sentence parsing: Evidence from Dutch. *The Quarterly Journal of Experimental Psychology Section A*, *49*(3), 664-695. https://doi.org/10.1080/713755636

Byrnes, H. (2013). Notes from the editor. *The Modern Language Journal*, *97*(4), 825–827. https://doi.org/10.1111/j.1540-4781.2013.12051.x

Carreiras, M. (1992). Estrategias de analisis sintactico en el procesamiento de frases: cierre temprano versus cierre ultimo [Strategies for syntactic analysis in sentence processing: early closure vs. late closure]. *Cognitiva* 4, 3–27. https://dialnet.unirioja.es/servlet/articulo?codigo=122589

Carreiras, M., & Clifton, C. (1999). Another word on parsing relative clauses: Eyetracking evidence from Spanish and English. *Memory & Cognition*, *27*(5), 826-833. https://doi.org/10.3758/BF03198535

*Cheng, Y., Rothman, J., & Cunnings, I. (2021). Parsing preferences and individual differences in nonnative sentence processing: Evidence from eye movements. *Applied Psycholinguistics*, *42*(1), 129-151. https://doi.org/10.1017/S014271642000065X.

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

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine, (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 3-18). Russell Sage Foundation.

Crowther, D., Kim, S., Lee, J., Lim, J., & Loewen, S. (2021). Methodological synthesis of cluster analysis in second language research. *Language Learning*, *71*(1), 99-130. https://doi.org/10.1111/lang.12428.

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition*, *30*(1), 73-105. https://doi.org/10.1016/0010-0277(88)90004-2.

*Dai, Y. (2015). Working memory in L2 sentence processing: The case with relative clause attachment. In Z. Wen, M. B. Mota, & A. McNeill (Eds.) *Working memory in second language acquisition and processing* (pp. 105-124). Multilingual Matters.

De Vincenzi, M., & Job, R. (1993). Some observations on the universality of the late-closure strategy. *Journal of Psycholinguistic Research*, *22*(2), 189-206. https://doi.org/10.1007/BF01067830

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly*, *50*(1), 132-153. https://doi.org/10.1002/tesq.217.

Dussias, P. E., & Sagarra, N. (2007). The effect of exposure on syntactic parsing in Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *10*(1), 101-116. https://doi.org/10.1017/S1366728906002847

Ehrlich, K., Fernandez, E. M., Fodor, J. D., Stenshoel, E., & Vinereanu, M. (1999). *Low attachment of relative clauses new data from Swedish, Norwegian, and Romanian.* Poster presented at the 12th Annual CUNY Conference on Human Sentence Processing, (pp. 18-20). New York, NY. https://drive.google.com/file/d/1yej4wKh8BO06Ws3_BwsmJ5JksyrdJKlW/view

Elliott, M. N., Haviland, A. M., Kanouse, D. E., Hambarsoomian, K., & Hays, R. D. (2009). Adjusting for subgroup differences in extreme response tendency in ratings of health care: impact on disparity estimates. *Health Services Research*, *44*(2p1), 542-561. https://doi.org/10.1111/j.1475-6773.2008.00922.x.

Farsani, M. A., & Babaii, E. (2020). Applied linguistics research in three decades: a methodological synthesis of graduate theses in an EFL context. *Quality & Quantity*, *54*(4), 1257-1283. https://doi.org/10.1007/s11135-020-00984-w.

Farsani, M. A., Jamali, H. R., Beikmohammadi, M., Ghorbani, B. D., & Soleimani, L. (2021). Methodological orientations, academic citations, and scientific collaboration in applied linguistics: What do research synthesis and bibliometrics indicate? *System*, *100*, 102547. https://doi.org/10.1016/j.system.2021.102547

Fedorova, O., & Yanovich, I. (2005). Early preferences in RC-attachment in Russian: The effect of Working Memory differences. In J. Lavine, S. Franks, M. Tasseva-Kurktchieva, & H. Filip (Eds.), *Proceedings of FASL 14*. Ann Arbor, MI: Michigan Slavic Publications, 113–128.

Felser, C., Marinis, T., & Clahsen, H. (2003). Children's processing of ambiguous sentences: A study of relative clause attachment. *Language Acquisition*, *11*(3), 127–163.

Fernández, E. M. (2003). *Bilingual sentence processing: Relative clause attachment in English and Spanish* (Vol. 29). Amsterdam: John Benjamins Publishing.

Fernández, E. M., & Sekerina, I. A. (2015). The Interplay of visual and prosodic information in the attachment preferences of semantically shallow relative clauses. In L. Frazier & E. Gibson (Eds.), *Explicit and Implicit Prosody in Sentence Processing* (pp. 241-261). Springer. https://doi.org/10.1007/978-3-319-12961-7_13

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies.* Unpublished doctoral dissertation. University of Connecticut, Connecticut, USA. https://www.researchgate.net/publication/27401728_On_Comprehending_Sentences_Syntactic_Parsing_Strategies

Frenck-Mestre, C., & Pynte, J. (1997). Syntactic ambiguity resolution while reading in second and native languages. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *50A*(1), 119–148. https://doi.org/10.1080/027249897392251

Hemforth, B., Konieczny, L. & Scheepers C. (2000) Modifier attachment: Relative clauses and coordinations. In B. Hemforth, & L. Konieczny (Eds.), *German sentence processing*. *Studies in theoretical psycholinguistics*, Vol. 24, (pp. 161-186). Springer, Dordrecht. https://doi.org/10.1007/978-94-015-9618-3_6

*Hocking, I. (2003). *Resources and parsing.* Unpublished doctoral dissertation. The University of Exeter. https://ianhocking.com/thesis.pdf

*Hopp, H. (2014). Working memory effects in the L2 processing of ambiguous relative clauses. *Language Acquisition*, *21*(3), 250-278. https://doi.org/10.1080/10489223.2014.892943

Hou, Z., & Aryadoust, V. (2021). A review of the methodological quality of quantitative mobile-assisted language learning research. *System*, *100*, 102568. https://doi.org/10.1016/j.system.2021.102568

*James, A. N., Fraundorf, S. H., Lee, E. K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language*, *102*, 155-181. https://doi.org/10.1016/j.jml.2018.05.006.

*Jun, S. A., & Bishop, J. (2015). Prominence in relative clause attachment: Evidence from prosodic priming. In L Frazier, E Gibson (Eds.), *Explicit and Implicit Prosody in Sentence Processing* (pp. 217-240). Springer.

Kamide, Y., & Mitchell, D. C. (1997). Relative clause attachment: Nondeterminism in Japanese parsing. *Journal of Psycholinguistic Research*, *26*(2), 247-254. https://doi.org/10.1023/A:1025017817290

*Kaya, M. (2012). Working memory and relative clause attachment preferences in Turkish: An eye-tracking study. *Studia Uralo-Altaica*, *49*, 265-278. https://bit.ly/3BdSQOS

*Kim, J. (2009). Working memory effects on L2 relative clause processing. *담화· 인지언어학회 학술대회 발표논문집*, 59-67. https://bit.ly/2XWRiKH

*Kim, J. H., & Christianson, K. (2013). Sentence complexity and working memory effects in ambiguity resolution. *Journal of psycholinguistic research*, *42*(5), 393-411. https://doi.org/10.1007/s10936-012-9224-4.

*Kim, J. H., & Christianson, K. (2017). Working memory effects on L1 and L2 processing of ambiguous relative clauses by Korean L2 learners of English. *Second Language Research*, *33*(3), 365-388. https://doi.org/10.1177/0267658315623322.

Leeser, M., & Sunderman, G. (2016). Methodological issues of working memory tasks for L2 processing research. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 89-104). John Benjamins. https://doi.org/10.1075/bpa.3

Li, S., Ellis, R., & Zhu, Y. (2019). The associations between cognitive ability and L2 development under five different instructional conditions. *Applied Psycholinguistics*, *40*(3), 693-722. https://doi.org/10.1017/S0142716418000796

Li, S., & Wang, H. (2018). Traditional literature review and research synthesis. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 123-144). Palgrave Macmillan. https://doi.org/10.1057/978-1-137-59900-1_6

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review*, 21, 861–883. https://doi.org/10.3758/s13423-013-0565-2

Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing*, *30*, 66-81. https://doi.org/10.1016/j.jslw.2015.08.011.

Loewen, S., & Plonsky, L. (2015). *An A–Z of applied linguistics research methods*. Macmillan International Higher Education. https://doi.org/10.1007/978-1-137-40322-3_1

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General*, *130*(2), 199-207. https://doi.org/10.1037/0096-3445.130.2.199.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comments on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, 109, 35–54. http://dx.doi.org/10.1037//033-295X.109.1.35.

Makel, M., & Plucker, J. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, *43*(6), 304-316. https://doi.org/10.3102/0013189X14545513

*Marefat, H., & Farzizadeh, B. (2018). Relative clause ambiguity resolution in L1 and L2: Are processing strategies transferred?. *Iranian Journal of Applied Linguistics (IJAL)*, *21*(1), 125-161. https://ijal.khu.ac.ir/article-1-2855-fa.html

*Marefat, H., Samadi, E., & Yaseri, M. (2015). Semantic priming effect on relative clause attachment ambiguity resolution in L2. *Applied Research on English Language*, *4*(2), 78-95. https://are.ui.ac.ir/article_15504_57a77d169b7f27df38b964fad01f162c.pdf

Marsden, E. (2020). Methodological transparency and its consequences for the quality and scope of research. In J. McKinley & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 15-28). Routledge.

Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages (pp. 1-21). Routledge.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews, and recommendations for the field. *Language Learning*, *68*(2), 321–391. https://doi.org/10.1111/lang.12286.

Marsden, E., Thompson, S., & Plonsky, L. (2017). Open science in second language acquisition research: The IRIS repository of research materials and data. In *SHS Web of Conferences* (Vol. 38, p. 00013). EDP Sciences.

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*(5), 861-904. https://doi.org/10.1017/S0142716418000036

McClean, S., Bray, I., Bray, I., de Viggiani, N., Bird, E., & Pilkington, P. (2019). *Research methods for public health*. Sage.

*Mendelsohn, A., & Pearlmutter, N. J. (1999). *Individual differences in relative clause attachment ambiguities*. Poster presented at the 12th Annual CUNY Conference on Human Sentence Processing, City University of New York.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, *18*(2), 5-11. https://doi.org/10.3102/0013189X018002005

Miyamoto, E. T. (1998). *Relative clause processing in Brazilian Portuguese and Japanese*. Unpublished doctoral dissertation. Massachusetts Institute of Technology, Massachusetts, USA. https://dspace.mit.edu/bitstream/handle/1721.1/68344/42471751-MIT.pdf?sequence=2&isAllowed=y

Najjari, R., & Mohammadi, M. (2017). The Development of Reading and Operation Span Tasks in Persian as Measures of Working Memory Capacity for Iranian EFL Learners. *Journal of Teaching Language Skills*, *36*(2), 129-162. https://doi.org/10.22099/jtls.2017.24688.2215

Nakanishi, T. (2015). A meta-analysis of extensive reading research. *TESOL Quarterly*, *49*(1), 6-37. https://doi.org/10.1002/tesq.157

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417-528. https://doi.org/10.1111/0023-8333.00136

Norris, J. M., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. M. Norris, & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 3-50). John Benjamins Publishing.

*Omaki, A. (2005). *Working memory and relative clause attachment in first and second language processing*. Unpublished master's thesis. University of Hawaii. https://scholarspace.manoa.hawaii.edu/bitstream/10125/11603/uhm_ma_3245_r.pdf

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492-518. https://doi.org/10.1093/applin/24.4.492

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85-110. https://doi.org/10.1017/S0267190510000115

Papadopoulou, D., & Clahsen, H. (2003). *The role of lexical and contextual information in parsing ambiguous sentences in Greek*. Department of Language and Linguistics, University of Essex.

*Payne, B. R., Grison, S., Gao, X., Christianson, K., Morrow, D. G., & Stine-Morrow, E. A. (2014). Aging and individual differences in binding during sentence understanding: Evidence from temporary and global syntactic attachment ambiguities. *Cognition*, *130*(2), 157-173. https://doi.org/10.1016/j.cognition.2013.10.005.

Pigott, T. D. (2012). *Advances in meta-analysis*. Springer.

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *The Modern Language Journal*, *98*(1), 450-470. https://doi.org/10.1111/j.1540-4781.2014.12058.x.

Plonsky, L. (2017). Quantitative research methods. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language research* (pp. 505–521). Routledge. https://doi.org/10.4324/9781315676968

Plonsky, L., & Gass, S. (2011). Quantitative Research Methods, Study Quality, and Outcomes: The Case of Interaction Research. *Language Learning, 61*(2), 325–366. https://doi.org/10.1111/j.1467-9922.2011.00640.x

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *The Modern Language Journal*, *102*(4), 713-731. https://doi.org/10.1111/modl.12509.

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning, 65*(S1), 9-36. https://doi.org/10.1111/lang.12111.

Plonsky, L., & Oswald, F. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). Routledge.

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, *36*(4), 583-621. https://doi.org/10.1177/0267658319828413.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638.

Shakki, F., Naeini, J., Mazandarani, O., & Derakhshan, A. (2020). Instructed second language English pragmatics in the Iranian context. *Journal of Teaching Language Skills, 39*(1), 201-252. https://doi.org/10.22099/jtls.2020.38481.2886

Samadi, E. (2014). *Relative Clause Attachment Ambiguity Resolution in L2: The Role of Semantics.* Unpublished master's thesis. The University of Tehran. http://utdlib.ut.ac.ir/PDFViewer/_PDFViewer/38962

Samavarchi, L. & Rezai, M. J. (2014). Online Processing of English Wh-Dependencies by Iranian EFL Learners. *Journal of Teaching Language Skills*, *32*(4), 63-83. https://doi.org/10.22099/jtls.2014.1856

Shariat, M. (2019). *Relative clause ambiguity resolution: A case for individual differences among second language learners.* Unpublished master's thesis. The University of Tehran. http://utdlib.ut.ac.ir/PDFViewer/_PDFViewer/109743

Sheykholmoluki, H. (2014). *Memory span and proficiency: Do they affect attachment preference of Persian l2 learners of English?* Unpublished master's thesis. The University of Tehran. http://utdlib.ut.ac.ir/PDFViewer/_PDFViewer/39235

Shin, J. (2020). A meta-analysis of the relationship between working memory and second language reading comprehension: Does task type matter? *Applied Psycholinguistics*, 41, 873-900. https://doi.org/10.1017/S0142716420000272

Sok, S., Kang, E. Y., & Han, Z. (2019). Thirty-five years of ISLA on form-focused instruction: A methodological synthesis. *Language Teaching Research*, *23*(4), 403-427. https://doi.org/10.1177/1362168818776673.

Soleimani, E. (2018*). Relative clause ambiguity resolution by second language learners of English: Impact of WMC and discourse cues.* Unpublished master's thesis. University of Tehran. http://utdlib.ut.ac.ir/PDFViewer/_PDFViewer/108119

Swets, B., Desmet, T., Hambrick, D. Z., & Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: a psychometric approach. *Journal of Experimental Psychology: General*, *136*(1), 64-81. https://doi.org/10.1037/0096-3445.136.1.64.

Traxler, M. J. (2007). Working memory contributions to relative clause attachment processing: A hierarchical linear modeling analysis. *Memory & Cognition*, *35*(5), 1107-1121. https://doi.org/10.3758/bf03193482.

Traxler, M. J. (2009). A hierarchical linear modeling analysis of working memory and implicit prosody in the resolution of adjunct attachment ambiguity. *Journal of Psycholinguistic Research*, *38*(5), 491-509. https://doi.org/10.1007/s10936-009-9102-x.

Whitney, P., & Budd, D. (1999). A separate language-interpretation resource: Premature fractionation? *Behavioral and Brain Sciences*, *22*(1), 113-113. https://doi.org/10.1017/s0140525x99421788.

Yeh, L. H. (2011). *The role of cross-language activation in syntactic ambiguity.* Unpublished doctoral dissertation. The University of Texas at El Paso. https://scholarworks.utep.edu/cgi/viewcontent.cgi?article=3614&context=open_etd

Zhang, M., & Plonsky, L. (2020). Collaborative writing in face-to-face settings: A substantive and methodological review. *Journal of Second Language Writing*, *49*, 100753. https://doi.org/10.1016/j.jslw.2020.100753

Ziegler, N. (2016). Synchronous computer-mediated communication and interaction. *Studies in Second Language Acquisition*, *38*(3), 553-586. https://doi.org/10.1017/S027226311500025X.

Appendix

List of 46 Coded Features of WMC Measures

A. Publication Information

1. *Authors* – (open response) What are the names of the authors?
2. *Year of publication* – (open response) What is the publication year of the retrieved study?
3. *Title of the study* – (open response) What is the title of the retrieved study?
4. *Publication type* – What type of publication was the retrieved study? (1 = Journal article, 2 = Conference proceedings, 3 = Thesis or dissertation, 4 = experimental book chapter).

B. Design Features

5. *WM Task* – What type of WM tasks have been employed? (1 = Complex WM task, 2 = Simple WM task).
6. *Origins of WMC Tests* – What is the origin of the employed WMC measure? (1 = Already-developed WMC tests, 2 = Researcher-developed WMC tests, 3 = Researcher-adapted WMC tests, 4 = Translated version, 5 = No information).
7. *Piloting* – Were the researcher-developed WMC measures piloted? (1 = Yes, 2 = No).
8. *Approach to WMC Measurement* – What approach to WMC measurement has been employed? A single-measure, non-psychometric approach in which only one test is used to tap the construct of WMC? Or a multiple-measure, psychometric approach in which multiple measures are employed to assess and tease apart the effects of multiple construct? (1= Single-Measure, 2 = Multiple-Measure).
9. *Type of WMC Measure* – What type of test was employed to measure the construct of WMC? (1 = Reading span test, 2 = Operation span test, 3 = Alphabet span test, 4 = Minus digit span test, 5 = Word span test, 6 = Listening span test, 7 = Spatial span test, 8 = No Information).
10. *Language of WMC Test* – What language was the WMC test written in? (1 = Operation span test [Language Neutral – since the operation span test is mathematical, it is independent of any language, and thus language neutral], 2 = English, 3 = Turkish, 4 = Korean, 5 = Chinese Russian, 6 = Japanese, 7 = Dutch).
11. *Addressing Sentence Length and Complexity* – Were sentence length and complexity controlled in the design of the WMC tests? (1 = Both sentence length and complexity, 2 = Neither sentence length nor complexity, 3 = Just sentence length, 4 = Not applicable, 5 = No information).
12. *Recall Tasks* – What type of recall task was employed to elicit what the participants recalled during taking the WMC test? (1 = Letters after sentences or equations, 2 = Final words of the sentence, 3 = Enhanced words in non-final position, 4 = Words after the sentence, 5 = Words in alphabetical order, 6 = Minus two digit task, 7 = Direction of tops of letters, 8 =

Storage-only: No. of correctly remembered words, 9 = Storage-only: Digits in the same order as heard, 10 = No information).

13. *Engaging and Tapping the Processing Component* – What type of task was employed to engage and tap the processing component of WM? 1 = Comprehension questions, 2 = Reading aloud and comprehension questions, 3 = Reading aloud, 4 = Normal/Mirror questions, 5 = Not applicable, 6 = No information).
14. *(Non-)simultaneous engagement of WM components* – How were the recall and processing components of WM been engaged and tapped? Simultaneously or non-simultaneously? 1 = Non-simultaneous, 2 = Simultaneous, 3 = Not applicable, 4 = No information).
15. *Practice Items* – Were practice items included for the warm-up section of WMC tests? (1 = Included, 2 = Not included).
16. *Set Size* – How many set sizes did the WMC measure have? (1 = 3 set sizes, 2 = 4 set sizes, 3 = 5 set sizes, 4 = 7 set sizes, 5 = No information).
17. *Item Size* – How many item sizes did the WMC measure have? (1 = 2 item sizes, 2 = 3 item sizes, 3 = 4 item sizes, 4 = 5 item sizes, 5 = 40 item sizes, 6 = No information).
18. *Number of elements* – How many elements did the WMC measure have? (1 = 36 elements, 2 = 40 elements, 3 = 42 elements, 4 = 60 elements, 5 = 70 elements, 6 = 75 elements, 7 = 80 elements, 8 = 100 elements, 9 = 120 elements, 10 = 175 elements, 11= No information).
19. *Shortest item size* – How many elements did the shortest item size include? (1 = 2 elements, 2 = 3 elements, 3 = No information).
20. *Longest item size* – How many elements did the longest item size include? (1 = 5 elements, 2 = 6 elements, 3 = 8 elements, 4 = 7 elements, 5 = No information).
21. *Comprehension Question Type* – What type of comprehension question was employed to measure comprehension accuracy? (1 = True/False, 2 = Normal/Mirror, 3 = Not applicable, 4 = No information).
22. *Avoiding Positive Response Tendency* – What type of strategy have been used in the design of WMC measures to avoid positive response tendency? (1 = Half true, half false [i.e., the WMC test was designed such that responses to half of the items were true and responses to the other half were false], 2 = Not applicable, 3 = No information).

C. Administration Features

23. *Presentation instrument* – What type of instrument was employed to present the WMC test to the participants? (1 = Screen [i.e., laptop or computer screen], 2 = E-Prime, 3 = Microsoft PowerPoint, 4 = DMDX, 5 = Experimenter reads, 6 = No information).
24. *Presentation type* – How were the elements of WMC tests presented? (1 = Entire element is presented at once, 2 = Non-cumulative, 3 = Entire element is listened to at once, 4 = Not applicable, 5 = No information).
25. *Presentation order of items* – In what order were the elements and items of WMC tests presented? (1 = Randomized, 2 = Ascending, 3 = No information).

26. *Counterbalancing* – Were the elements and items counterbalanced for the participants? (1 = Yes, 2 = No, 3 = No information).
27. *Test discontinuation* – Were the administration of WMC tests discontinued when the participants failed to recall a certain number of target information? (1 = Yes, 2 = No, 3 = No information).
28. *Criteria for discontinuing tests* – What criteria was used for the discontinuation of WMC tests? (1 = Failing all three items in a set, 2 = Failing to recall the words from two consecutive items, 3 = Making two or more mistakes in a set, 4 = Not applicable because there is no discontinuation, 5 = No information).
29. Individual-based or Group-based administration – Were WMC tests administered individually or in group? (1= Individual-based, 2 = Group-based, 3 = No information).
30. *Timing for element presentation* – Was a time limit considered for the presentation of each element of WMC tests? (1 = Considered, 2 = Individually-calibrated timing, 3 = No information).
31. *Timing for judgment task* – Was a time limit considered for the presentation of the judgment task of WMC tests? (1 = Considered, 2 = Individually-calibrated timing, 3 = No time limit, 4 = No information).
32. *Timing for recall task* – Was a time limit considered for the presentation of the recall task of WMC tests? (1 = Considered, 2 = No time limit, 3 = No information).
33. *Recording responses to comprehension questions* – How were the responses to comprehension questions recorded? (1 = Participants wrote responses on answer sheets, 2 = Experimenter wrote responses on answer sheets, 3 = Software recorded responses, 4 = Not applicable, 5 = No information).
34. *Recording the to-be-recalled information* – How were the to-be-recalled information recorded? (1 = Participants wrote responses down on answer sheets, 2 = Participants typed responses, 3 = Experimenter wrote responses down on answer sheets, 4 = No information)
35. *Sessions and sequence of tests* – In how many sessions were WMC tests administered and in what sequence were WMC tests administered as compared with other tests? (1 = A single session, with no information about sequencing, 2 = One session: WMC test administered 'before' RC attachment task, 3 = One session: WMC test administered 'after' RC attachment task, 4 = Two sessions: WMC test administered before RC attachment task, 5 = Two sessions without reference to task sequence, 6 = No information).

D. Scoring Features

36. *Reliability Coefficients* – Were reliability coefficients for the scores of WMC tests investigated and reported? (1= Reported, 2 = Not reported, 3 = No information).
37. *WMC as continuous or nominal* – Which measurement scale was used for the analysis of the variable WMC? (1 = Continuous, 2= Nominal, 3 = No information).

38. *Cut-off score* – What type of cut-off score was employed for classifying high and low span participants? (1 = Mean as the cut-off point, 2 = High span (span $\geq$ 4) or low span (span < 4), 3 = 30th and 70th percentiles as cut-off points, 4 = Not applicable for a continuous variable, 5 = No information).
39. *Detailed features of scoring system* – What are the features of the scoring system employed for WMC measures? (1= [element level, +recall, +processing, -order, -fractional], 2 = [element level, +recall, -processing, -order, -fractional], 3 = [element level, +recall, +85%processing for all the test, -order, -fractional], 4 = [item level, +recall, -processing, +order, -fractional], 5 = [item level, +recall, +processing, +order, -fractional], 6 = [item level, +recall, -processing, +order, +fractional], 7 = [item level, +recall, +85%processing for all the test, +order, -fractional], 8 = [set level, +recall, +75% processing for each set, -order, +fractional], 9 = [set level, +recall, -processing, -order, -fractional], 10 = [set level, +recall, -processing, -order, +fractional], 11 = No information).
40. *Threshold level* – Was a threshold level employed for data trimming? Were data eliminated based on some threshold level? (1 = Not used, 2 = Used).
41. *Scoring method* – What type of scoring method was employed? (see Conway et al., 2005, for detailed explanation; 1= Inferred to be partial-credit load scoring, 2 = Inferred to be absolute scoring method, 3 = Inferred to be partial-credit unit scoring, 4 = Composite scoring of recall and processing, 5 = Partial-credit unit scoring, 6 = Composite scoring of recall, processing and order, 7 = All-or-nothing unit scoring, 8 = All-or-nothing load scoring, 9 = No information).

E. Transparency Features

42. *Sample test items in the materials section* – Were sample test items provided in the materials section of the study? (1 = Yes, 2 = No).
43. *Supplementary materials* – Was the whole test available in the study (i.e., appendix) or in supplementary material on the website of the journal etc.? (1 = Yes, 2 = No).
44. *Transparency of information for design features* – Were the design features of WMC tests reported in the study? (open response as it is the ratio of the presence of reporting a specific feature to the total presence and non-presence of that particular feature in the description of WMC measures).
45. *Transparency of information for administration features* – Were the administration features of WMC tests reported in the study? (open response as it is the ratio of the presence of reporting a specific feature to the total presence and non-presence of that particular feature in the description of WMC measures).
46. *Transparency of information for scoring features* – Were the scoring features of WMC tests reported in the study? (open response as it is the ratio of the presence of reporting a specific feature to the total presence and non-presence of that particular feature in the description of WMC measures).