SCIEPublish

Article

Are Memory Updating Tasks Valid Working Memory Measures? A Meta-Analysis

Chenxi Wang [1,2,3,†], Jun Wang [1,2,4,†], Xuan Zeng [1,2,†], Nancy Xiaonan Yu [3] and Tianyong Chen [1,2,*]

[1] Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China; cwang552@cityu.edu.hk (C.W.); wangjun@psych.ac.cn (J.W.); zengx@psych.ac.cn (X.Z.)
[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Department of Social and Behavioral Sciences, City University of Hong Kong, Hong Kong, China; nancy.yu@cityu.edu.hk (N.X.Y.)
[4] Department of Military Training, China Coast Guard Academy, Ningbo 315801, China

* Corresponding author. E-mail: chenty@psych.ac.cn (T.C.)
† These authors contributed equally to this work.

Received: 9 October 2024; Revised: 1 November 2024; Accepted: 25 January 2025; Available online: 18 February 2025

ABSTRACT: The memory updating (MU) process is a core component of working memory (WM). To systematically examine the validity of two commonly used MU tasks as WM measures, the present meta-analysis (76 studies, total N = 16,184) synthesized results on the correlation between the two MU tasks and two criterion tasks (working memory capacity (WMC) and fluid intelligence (*Gf*)). Results indicated a moderate correlation between running memory (RM) and WMC (r = 0.42, 95% CI = [0.37, 0.48]), a weak correlation between n-back and WMC (r = 0.23, 95% CI = [0.19, 0.28]), and moderate correlations between both RM (r = 0.40, 95% CI = [0.35, 0.46]) and n-back (r = 0.34, 95% CI = [0.32, 0.37]) and *Gf*. Subgroup analyses showed that memory load moderated the correlation between RM and WMC, and stimulus-onset asynchrony moderated the correlation between n-back and both WMC and *Gf*. The recollection and recognition nature of RM and n-back contributed to their different correlation with WMC, and the involvement of controlled attention in both tasks accounted for their association with *Gf*. The present meta-analysis indicated that RM is a more valid WM measure in behavioral studies on individual differences.

Keywords: Memory updating; Working memory capacity; Fluid intelligence; Meta-analysis

1. Introduction

Working memory (WM) refers to a cognitive system that "provides temporary storage and manipulation of the information necessary for complex cognitive tasks" [1]. A core feature of WM is its dynamic nature, as representations in WM are constantly being updated, with older and no longer relevant information replaced by newer and more relevant ones [2]. Measures of WM capacity (WMC), such as the most widely used complex span tasks (also referred to as WM span tasks [3]), have captured this feature of dynamism. For example, in an operation span task, participants need to solve a math operation and then remember a word before moving to the next operationword pair. At the end of the trial, they are asked to recall the words in order [4]. In order to complete this task, participants have to update the retained list of words every time they see a new word. Given its importance, the memory updating (MU) process has been extensively addressed when investigating WM [5], but systematic evaluation of the validity of MU tasks in measuring WM is rare.

A commonly used MU task is the running memory (RM) task [6], which requires the participants to watch or listen to a series of items and then recall a fixed number of the most recent items in sequence (fixed partial recall). The length of the item series is usually unknown, so the participants have to constantly update their representations of the most recent items from the beginning to the end of each trial [3]. Multiple operations are involved in the RM task, including monitoring, encoding, maintaining, and updating representations of the stimuli throughout the trial. Moreover, the non-stop presentation of stimuli within a trial demands all these processes to be executed in a very short period of time,

https://doi.org/10.70322/ldmh.2025.10003

especially simultaneously registering the new stimulus and maintaining the older ones. These characteristics align with the core elements of WM, and RM has indeed been found to be closely linked with WM. Researchers identified a stable correlation (moderate to high) between RM and both WMC and fluid intelligence (*Gf*, the ability to reason and to solve problems [7]) tasks, supporting RM as a WM task that predicts higher-order cognitive functions [8,9]. Adding to this notion, neural substrates of RM greatly overlap with that of WM. Neuroimaging studies found that RM mainly relies on the prefrontal and parietal lobes [10,11], and poor performance on RM was observed in stroke patients with frontal damage [12].

Another frequently used MU task is the n-back task, which requires the participants to monitor a series of stimulus displayed in sequence, determine whether or not the current item matched the one presented n items ago, and make a response by key pressing [13]. As in the case of RM, n-back involves multiple cognitive operations, including encoding, monitoring and maintaining representations of a stimulus, as well as comparing it with a previously maintained one [14]. Also, frontoparietal activations were found to be associated with n-back tasks (see [15,16] for meta-analyses), and people with dorsolateral prefrontal dysfunctions, such as schizophrenic had poorer performance on n-back tasks (see [17] for review). Besides, performance on n-back has also been found to be closely related to *Gf* [13,18].

Although the face validity of n-back as a WM measure seems to have been endorsed by the above evidence, doubts on its criterion validity still exist. In contrast with the case of RM, weak or even no correlation was found between n-back and WMC [13,18–21]. Other researchers argued that these studies had methodological issues such as task selection, and their studies found wide-ranging correlations (0.17 to 0.51) between n-back and WMC [22].

It is worth noting that n-back and RM bear substantial disparities, although both tasks are popular MU tasks involving serial representation of stimuli at the center of the screen. During an n-back task, participants are instructed to compare whether a stimulus matches another, rendering the task familiarity- and recognition-based; during an RM task, on the other hand, participants are instructed to recall a number of items, making the task recollection-based and therefore, more similar to complex span tasks. Such differences in the two tasks may result in their discrepant correlation with WMC. However, there has been no study systematically synthesizing and comparing existing results on the correlation between these two MU tasks and the criterion tasks.

Several variables within the two MU tasks (n-back and RM) can potentially moderate the MU tasks' association with WMC and *Gf*. First, by altering the memory load in n-back and RM, the updating demands of both MU tasks can be manipulated. Concerning the fact that the measurement of individual differences in WMC is largely based on variances in WM load, if the two MU tasks are indeed valid WM measures, their association with WMC is supposed to vary with memory load. This means that in an MU task, the higher the memory load is, the larger WMC should be required and the stronger its correlation with WMC measures should be. Another possible moderating variable is stimulus-onset asynchrony (SOA), which refers to the stimulus presentation rate, and participants tend to use different strategies during slow- and fast-presentation MU tasks [23]. For example, longer SOAs allow more time for the participants to rehearse the information, and this difference in strategy utilization can potentially affect the variance shared by the MU tasks and WMC. Furthermore, the content of stimuli (verbal *vs.* non-verbal) of a MU task is associated with domain-specific cognitive processes [24] and, therefore, could contribute to different variances shared by the MU task and the criterion task. A final candidate moderator is the response method (particularly in n-back tasks), which influences both the cognitive demands [13] and the updating and decision-making processes [25,26], potentially affecting the task's relationship with WMC and Gf. It is, therefore, intriguing whether these within-task variables influence the validity of n-back and RM as WM measurements.

To sum up, the present meta-analysis focuses on the criterion validity of the two representative MU tasks, namely the recognition-based n-back and the recollection-based RM. Complex span, the most established measurement of individual difference in WMC [3,27], was selected as a criterion task. Since an important feature of either WM or MU is its ability to predict individual differences in higher-order cognition [28–30], *Gf* was selected as another criterion. Specifically, the present study aimed to examine: (1) the correlations between the two MU tasks (RM and n-back) and the two criterion tasks (WMC and *Gf*); and (2) the moderating effects of two groups of within task variables, such as task difficulty, on these correlations.

2. Materials and Methods

The present study followed the PRISMA guideline for reporting systematic reviews, including meta-analysis [31,32].

2.1. Search Strategies and Study Selection

Three online databases (APA PsycInfo, APA PsycArticles and PubMed) and Google Scholar were searched for relevant articles in June 2023, and a supplementary search was performed in November 2024. Search terms were determined separately for each combination of the MU tasks (RM and n-back) and the criterion tasks (WMC and Gf). For example, search terms for articles that used both n-back and WMC tasks (complex span tasks) were: ("back" OR "2-back" OR "3-back") AND ("working memory span" OR "operation span" OR "reading span" OR "computation span" OR "counting span" OR "symmetry span" OR "rotation span"). Four searches were conducted for the four combinations, and no limit was set on the year of publication (see Table S1 for the full search strategy).

To be included in the present study, an article needed to meet all inclusion criteria and no exclusion criteria. Inclusion criteria required that: (1) the article was peer-reviewed and published journal article; (2) the study included at least one of the two MU tasks (RM or n-back) and one criterion task (complex span or *Gf* test); (3) participants were healthy adults (above 18 years old); and (4) the article reported Pearson correlation coefficient (*r*) between accuracy on target probes of the MU task and the criterion task, or provided other data that could be used to calculate *r*. If an article did not provide such data, the authors were contacted to obtain the necessary information. The article was included if the required data were obtained through this process. Exclusion criteria were: (1) the n-back task used did not include either 2- or 3-back. 1-back tasks were excluded since they did not adequately engage MU [33]; while 4-back tasks and above were excluded since higher loads are relatively less common and may suffer from floor effects and low reliability [18]. and (2) variants of RM (see [34,35] for example) with major adaptations of the task's core elements (e.g., SOA > 500 ms and presenting only one stimulus at the center of the screen at one time).

Retrieved records were imported into Endnote X9, and two authors independently screened titles and abstracts of the records. Relevant studies were then included in the full-text screening process, during which needed information was extracted to decide the study's eligibility for inclusion. The two authors discussed the record together when there was a disagreement, and a third author would join the discussion and provide an opinion if no consensus was reached.

2.2. Coding Process

The author, publication year, and number of experiments were extracted for each article. Further, for each experiment, we extracted the sample size, memory load of the MU task (n in n-back and number of to be remembered items), modality of the MU task (type of stimulus), response method of n-back (pressing one or two key), SOA and Pearson correlation coefficient (*r*) between the MU task and the criterion task. Memory load of n-back was coded into three types: two, three and combined (of two and three), and that of RM was coded into two types: fixed (three or four) and combined. Stimuli of n-back was coded into verbal if all stimuli were verbal (for example, word, letter or digit) and non-verbal if non-verbal stimuli (for example, face, spatial or shape) were included in the experiment. SOA of n-back was coded into two types: equal to or below 2500 ms and above 2500 ms. This cut-off value was decided based on the Baddeley's working memory model, which suggested that the phonological loop can maintain information for about 2–3 s without rehearsal. Therefore, an SOA within this range is considered to reflect "pure" WM [1]. SOA of RM was also coded into two types: equal to or below 1000 ms and above 1000 ms. This cut-off value was chosen since previous studies comparing slow- and fast-presentation RM typically used an SOA of 1000 ms or above for the slow-presentation condition (e.g., [8,23]). Number of response keys in n-back was coded into two types: one key or two keys. Some studies reported more than one correlation, for example, correlations between 2-back and *Gf* and 3-back and *Gf* were reported for a single experiment. A weighted average r would be calculated for this experiment in such cases.

2.3. Publication Bias Estimation

Publication bias was estimated through a combination of fail-safe numbers (N_{fs}), Egger's test, and. N_{fs} refers to the number of studies needed to render the meta-results non-significant [36], and a smaller N_{fs} indicates a high possibility of publication bias. When an N_{fs} is smaller than $5k + 10$ (k being the number of studies included in the analysis), there is an elevated risk that the positive findings have resulted from publication bias instead of actual effects [37]. Concerning the Egger's test, an intercept with a large absolute value that reached statistical significance ($p < 0.05$) indicates a high risk of publication bias [38] and a need for further exploration. Funnel plots with trim and fill analysis examine the symmetry of study distribution and estimates adjusted effect sizes by correcting for potential asymmetry in the funnel plot [39].

2.4. Statistical Analysis

The analyses were conducted using CMA 3.3. First, a Fisher's *Z* score was converted from the original *r* for each experiment. With T inverse-variance weighting, these *Z* scores and standard errors (SE) were used to calculate a summary Fisher's *Z*, which would then be converted back into a summary *r*. This summary *r* was used as an indicator of effect size, with |*r*| within (0.5, 1.0] representing strong, (0.3 to 0.5] moderate, and [0.0, 0.3] weak correlation [40]. Heterogeneity of the studies was examined through a combination of I^2 and the *p*-value for Cochran's *Q*. According to the Cochrane Handbook, fixed effects models should only be adopted when there is small heterogeneity between the included studies ($I^2 < 40\%$). Therefore, if $p \leq 0.05$ or $I^2 \geq 40\%$, a random effect model would be adopted [41–43]. To test moderating effects, 95% CIs of the subgroup correlations between the MU and criterion tasks were compared. Non-overlapping CIs indicate a significant difference between the subgroup correlations and moderating effect [43].

3. Results

A total of 1616 records were retrieved. After removing 366 duplicates, 1250 articles were included in the title and abstract screening, where 885 were excluded due to irrelevant research object, 86 due to characteristics of the participants, 1 due to publications being retracted, 7 due to being review study, and 1 being media report. Within the 270 articles retained for full-text screening, 109 were excluded due to irrelevant research objects, 2 due to characteristics of the participants and 83 due to incomplete data. Finally, 76 articles with a total sample size of 16,184 were included in the present meta-analysis. Specifically, 27 articles reported the correlation between n-back and WMC, 34 between n-back and *Gf*, 22 between RM and WMC, and 21 between RM and *Gf*. See Supplementary Tables S2–S5 for basic information on the included studies and Supplementary Figures S1–S4 for the forest plots.

3.1. Publication Bias

N_{fs} for RM and WMC, n-back and WMC, RM and *Gf*, and n-back and *Gf* were 5890, 3733, 7284, and 7939, respectively, much larger than $5k + 10$ (Table 1). Also, Egger's test indicated that the intercepts for the four sets of studies were −0.10 ($p = 0.905$, 95% CI = [−1.88–1.67]), −0.18 ($p = 0.837$, 95% CI = [−1.92–1.56]), −2.05 ($p = 0.084$, 95% CI = [−4.39–0.30]) and 0.32 ($p = 0.519$, 95% CI = [−0.67–1.31]), respectively. In addition, trim-and-fill analyses showed no change in the pattern of results. Together, these results indicated low-risk of publication bias.

Table 1. Publication bias of main results.

Correlation	*k*	N_{fs}	Egger's Test				Trim-and-Fill Analyses		
			Intercept	*SE*	95% CI	*p*	k_t	Adjusted *r*	95% CI
RM & WMC	23	5890	−0.10	0.85	[−1.88–1.67]	0.905	1	0.43	[0.37–0.48]
n-back & WMC	29	3733	−0.18	0.85	[−1.92–1.56]	0.837	3	0.22	[0.17–0.26]
RM & *Gf*	22	7284	−2.05	1.13	[−4.39–0.30]	0.084	1	0.39	[0.34–0.45]
n-back & *Gf*	40	7939	0.32	0.49	[−0.67–1.31]	0.519	7	0.32	[0.29–0.35]

Note. *k* = the number of studies; *SE* = standard error; CI = confidence interval; k_t = the number of studies trimmed; WMC = working memory capacity; RM = running memory; *Gf* = general fluid intelligence.

3.2. Summary Effects

The results of all four heterogeneity tests were statistically significant ($p < 0.001$, Table 2), and therefore, random effect models were adopted for all main analyses. Regarding association with WMC, a moderate correlation was found between RM and WMC ($r = 0.42$, 95% CI = [0.37, 0.48]), and a weak correlation was found between n-back and WMC ($r = 0.23$, 95% CI = [0.19, 0.28]). Results indicated that the *r* for RM was significantly larger than that for n-back, since there was no overlap between the two 95% CIs. Regarding *Gf*, a moderate correlation was found between both RM and *Gf* ($r = 0.40$, 95% CI = [0.35, 0.46]), and n-back and *Gf* ($r = 0.34$, 95% CI = [0.32, 0.37]). The overlap in the two 95% CIs indicated that the difference between these two correlations with *Gf* was not statistically significant.

Table 2. Meta-analysis of main results and Heterogeneity.

Correlation	*k*	Sample Size	Effect Size		Heterogeneity	
			r	95% CI for *r*	*Q*	I^2 (%)
RM & WMC	23	3239	0.42	[0.37, 0.48]	96.88 ***	77.29
n-back & WMC	29	6456	0.23	[0.19, 0.28]	149.09 ***	81.22
RM & *Gf*	22	4960	0.40	[0.35, 0.46]	121.59 ***	82.73
n-back & *Gf*	40	7906	0.34	[0.32, 0.37]	93.85 ***	58.45

Note. *k* = the number of studies; CI = confidence interval; WMC = working memory capacity; RM = running memory; *Gf* = general fluid intelligence. *** $p < 0.001$.

3.3. Moderator Analysis

Random effect models were adopted for all moderator analyses based on the results of the heterogeneity tests (Table 3). For RM, memory load (3 & 4 or combined) and SOA (equal to or below 1000 ms and above 1000 ms) were tested for moderating effects. Memory load but not SOA was found to be an effective moderator of the correlation between RM and WMC. Specifically, the correlation was higher in the combined condition ($r = 0.48$, 95% CI = [0.42, 0.55]) than in the 3 & 4 condition ($r = 0.34$, 95% CI = [0.28, 0.41]). Regarding the correlation between RM and *Gf*, no statistically significant difference was found between the subgroup correlations for either variable. For n-back, memory load (2, 3 or combined), stimuli (verbal or non-verbal), SOA (equal to or below 2500 ms and above 2500 ms) and response method (1 or 2 keys) were identified as potential moderators. Only SOA was found to be an effective moderator of the correlation between n-back and WMC. Specifically, the correlation was higher when SOA was equal to or below 2500 ms ($r = 0.29$, 95% CI = [0.25, 0.34]) than above 2500 ms ($r = 0.13$, 95% CI = [0.04, 0.21]). For each other moderator, there was an overlap in the subgroup 95% CIs for *r*, indicating that there was no statistically significant difference between the subgroup correlations and that none of these other within-task variables moderated the correlation between n-back and the criterion tasks.

Table 3. Moderator analyses.

Moderator	*k*	Effect size		*Z*	*Q*	I^2 (%)
		r	95% CI for *r*			
load						
RM & WMC						
Fixed (3 or 4)	11	0.34	[0.28, 0.41]	10.00 ***	29.01 **	65.53
Mixed	11	0.48	[0.42, 0.55]	15.32 ***	24.48 **	59.14
n-back & WMC						
2	11	0.19	[0.11, 0.26]	4.86 ***	38.19 ***	73.82
3	12	0.22	[0.16, 0.29]	6.84 ***	54.33 ***	79.75
Mixed	12	0.25	[0.17, 0.34]	5.67 ***	39.48 ***	72.14
RM & *Gf*						
Fixed (3 or 4)	10	0.38	[0.28, 0.49]	7.13 ***	89.84 ***	89.98
Mixed	11	0.42	[0.36, 0.48]	14.32 ***	29.58 **	66.20
n-back & *Gf*						
2	18	0.33	[0.30, 0.37]	16.88 ***	38.82 **	56.21
3	17	0.33	[0.28, 0.38]	13.63 ***	46.17 ***	65.35
Mixed	14	0.36	[0.31, 0.40]	16.13 ***	21.86	40.54
SOA						
RM & WMC						
≤1000	7	0.41	[0.33, 0.49]	10.33 ***	16.05 *	62.61
>1000	11	0.44	[0.35, 0.53]	9.48 ***	68.71 ***	85.45
n-back & WMC						
≤2500	17	0.29	[0.25, 0.34]	12.95 ***	37.74 **	57.61
>2500	11	0.13	[0.04, 0.21]	3.01 **	48.28 ***	79.29
RM & *Gf*						
≤1000	8	0.43	[0.31, 0.54]	7.31 ***	58.23 ***	87.98
>1000	16	0.39	[0.33, 0.45]	12.78 ***	60.13 ***	75.05
n-back & *Gf*						
≤2500	25	0.36	[0.33, 0.39]	23.57 ***	46.57 **	48.46
>2500	12	0.29	[0.23, 0.34]	10.29 ***	32.52 **	66.17
Stimuli						
n-back & WMC						
verbal	18	0.23	[0.17, 0.28]	8.28 ***	68.62 ***	75.23
non-verbal	17	0.24	[0.17, 0.31]	6.96 ***	82.08 ***	80.51
n-back & *Gf*						

verbal	21	0.33	[0.30, 0.37]	18.19 ***	59.05 ***	66.13
non-verbal	24	0.35	[0.32, 0.38]	20.39 ***	54.17 ***	57.54
Number of response key						
n-back & WMC						
One	11	0.20	[0.12, 0.29]	4.57 ***	83.94 ***	88.09
Two	15	0.25	[0.19, 0.31]	8.22 ***	46.75 ***	70.05
n-back & *Gf*						
One	21	0.34	[0.31, 0.37]	23.46 ***	40.08 *	50.10
Two	17	0.35	[0.30, 0.40]	13.00 ***	53.54 ***	70.12

Note. k = the number of studies; CI = confidence interval; WMC = working memory capacity; RM = running memory; *Gf* = general fluid intelligence. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4. Discussion

To add evidence to the debate on whether MU tasks are valid WM measures, the present meta-analysis examined the criterion validity of the two most widely used MU tasks, namely the n-back task and the RM task. Results indicated that n-back tasks were weakly correlated with complex span tasks, replicating a previous meta-analysis [21], and moderately correlated with *Gf*. In contrast, RM tasks were moderately correlated with both complex span and *Gf*. It is worth noting that although the word WMC was used in the results and the following discussions, the complex span task was selected as the indicator of this theoretical construct throughout the present study. Therefore, when interpreting the results, it should be borne in mind that our findings cannot necessarily be generalized to other WMC indicators.

4.1. MU and WMC

Although being an established measurement of WMC, the complex span has been criticized for its task impurity due to the inherent storage-processing trade-off, resulting in difficulty in interpreting individual differences in performance. Additionally, there have been debates on whether the scoring scheme was reasonable, taking performance on only the memory task but not the processing task into account [44]. These limitations have motivated researchers to use MU tasks as alternative indicators of WM, making it intriguing and crucial to examine the relationship between these MU tasks and complex spans as well as other higher-order cognitive abilities.

MU has long been identified as an important executive process closely correlated with WMC at the level of individual differences [22,45]. Regarding why this correlation is different for n-back and RM, the types of retrieval involved in the two tasks can give us a clue. While RM is a recollection-based task requiring serial recall of retained items, similar to complex span tasks, n-back is recognition-based and requires a serial judgment of matching between items. In line with this argument, a previous study has explicitly examined how WMC was linked with recognition and recollection-induced variance [20]. They found that the performance of people with high and low WMC differed on recollection-based but not recognition/familiarity-based measures. A further Structural Equation Model showed a correlation between WMC and the recollection but not the familiarity latent factor in tasks involving both. Studies on a recall-variant of n-back supported the above point of view. This variant asks the participants to recall the n-back item after displaying a list of stimuli [46,47], turning the task into a recollection-based one that shares more variance with WMC. We conducted a supplemental analysis and found a moderate correlation (r = 0.44, 95% CI = [0.35, 0.53]) between this recall-variant of n-back and WMC.

The Embedded-Processes Model of WM [48] provided a useful theoretical framework for understanding the mechanisms underlying recognition- and recollection-based tasks. The model proposed a hierarchical structure within WM, comprising an activated subset of long-term memory and a capacity-limited focus of attention (FoA), where a limited storage of representations can be accessed and retrieved quickly. According to this model, the limit of WMC reflects the number of items that can be stored in FoA [49]. While both types of storage can afford the demands of recognition tasks, FoA meets the needs of recall tasks [20]. Therefore, during RM and complex span tasks, in order to successfully recollect the to-be-remembered items, participants need to actively update the contents in FoA, replacing its storage with the newest relevant items. On the other hand, as performance on n-back tasks is driven mainly by familiarity-based discrimination [18], representations of the stimuli were more likely to be passively stored [50] and would soon fade from FoA, leaving traces only in the activated subset long-term memory. Here, items are at a lower level of activation than in FoA [48] but can still meet the need for recognition tasks. In a word, RM and n-back examine different layers of WM, with RM focusing on FoA and n-back measuring the activated long-term memory.

Moderator analysis provided further support for the above explanation. The moderation tests on the type of stimuli and the number of response keys were both non-significant. The type of stimuli is associated with cognitive operations

specific to certain domains [24], while the number of response keys is associated with response bias and decision thresholds [25,26]. Therefore, the lack of moderation effects indicated that the observed main results were domain-general and not subjected to the factors related to decision-making or response process.

Concerning the moderation effects of memory load, the correlation between fixed-load (3 or 4) RM and WMC was lower than between mixed-load RM and WMC. As the upper limit of memory load in the mixed condition (ranging from 6 to 8) is much larger than in the fixed condition (see Supplementary Table S3), mixed-load RM casts higher demand on individual's ability to update information in FoA of WM and thus, share more variance with individual difference in WMC. It is worth noting that even the fixed-load RM, where memory load is similar to that in commonly used n-back, still had a moderate correlation with WMC, higher than that of n-back. That is to say, the stronger correlation between RM and WMC cannot be explained by the larger memory load of RM than n-back, supporting a distinction in the WM components that the two MU tasks address. For n-back, on the other hand, memory load did not moderate its correlation with WMC. Since n-back mainly taps the activated subset of long-term memory, which does not have a strict capacity limit as FoA [49], it is reasonable that n-back is less sensitive to the manipulation of memory load than RM. Furthermore, as the capacity of FoA is the major source of variation in WMC, even if the individual difference in the efficiency of activated long-term memory is better manifested when "*n*" is larger, the shared variance between n-back and WMC will remain unchanged. Together, the moderating effects of memory load strengthened the main result that the recollection-based RM is a more qualified WM measure than the recognition-based n-back.

Concerning the moderator for the association between n-back and WMC, when SOA was longer than 2500 ms, its correlation with WMC (and also with *Gf*) was significantly weaker than when SOA was shorter, which may have resulted from the different components n-back tasks with different SOA emphasize. During MU tasks, updating changed items and maintaining remembered items have been identified as two separate operations [5,51]. Response time is required for the updating operation and the switching between updating and maintenance [51]. During a slow-presentation n-back task, there is more time left for the maintenance of stored items after participants have made a response for the present trial and updated needed information. Therefore, we argue that n-back with a long SOA, to a great extent, assesses the maintenance operation rather than focusing on updating efficiency. That means slow-presentation n-back is more suitable for investigating the representation of maintained information instead of studying the individual difference in MU.

Beyond theoretical considerations, psychometric factors could contribute to the stronger correlation between RM and complex span compared to n-back. First, researchers found inconsistent reliability on n-back, and the unsatisfactory reliability, especially in 3-back, could attenuate the validity of the n-back task [18,52]. Additionally, a higher task specificity or weaker discrimination power of n-back can be an alternative explanation for its limited shared variances with other WMC tasks, including complex spans. If supported by future empirical evidence, these characteristics would suggest that although n-back is suboptimal to measure individual differences in domain-general WMC, it can still serve as a valid measurement when a specific domain of WM needs to be addressed.

4.2. MU and Gf

Our results have provided evidence for the close association between MU tasks (RM and n-back) and a widely used construct of higher-order cognition, *Gf*. This association can be accounted for by controlled attention or cognitive inhibition in both tasks. Controlled attention refers to a cognitive resource for the active maintenance of useful information and inhibits distracting information in the face of concurrent processing and/or distractors [53]. Similarly, cognitive inhibition denotes the ability to resist unwanted mental representations [54]. During both RM and n-back, as the to-be-remembered items change, controlled attention is essential for inhibiting old and no longer needed items and, at the same time, maintaining the new relevant ones. Previous studies have revealed that this ability to inhibit irrelevant information is important in the correlation between MU and *Gf*. In a variant of n-back, lure trials (probes matching a more recent item than the *n*th-to last one) were added to the task to measure the efficiency of inhibition, and compared with accuracy on the non-lure trials, accuracy on the lure trials was found to more closely associated with *Gf*. Furthermore, neural activation specific to the lure trials (mainly in dorsolateral prefrontal and parietal cortex) explained the correlation between lure-trial accuracy and *Gf* [7,55]. In the original n-back tasks that the presented study investigated, although lures are not explicitly labeled, probes matching an item near but not the nth-to last stimuli still exist, tasking on the ability to inhibit the temptation to endorse these familiar probes.

Controlled attention has also been found to be one of the two components explaining the shared variance between WMC tasks and *Gf*, with the other being the capacity component per se [56–58]. In addition, other lower-level cognitive

processes are also underlying contributors to the correlation between MU and *Gf*. According to the Process Overlap Theory, the commonality between higher-level cognitive tasks often stems from a range of more basic cognitive processes [59]. For example, processing speed contributes to both the efficiency of WM [60] and *Gf* [61], and long-term memory has been found to account for individual differences in WM and its association with Gf [58]. Therefore, the weak yet significant correlation between n-back and WMC tasks can also be accounted for by their common involvement of controlled attention and other lower-level cognitive abilities.

4.3. Limitations and Future Directions

A major limitation of the present meta-analysis is the unguaranteed exhaustiveness of the search strategy. A narrow review strategy has been adopted, using specific task names rather than broader terms such as "WMC" as searching terms. While this focused approach allowed for manageable and in-depth analysis, it would have led to relevant studies being missed, particularly those examining WMC using multiple indicators. Future studies with greater resources can adopt a more comprehensive search strategy to capture the additional studies. Furthermore, to maximize the quality of the studies, only peer-reviewed published articles were included, and consequently, we may have left out pertinent grey literature. However, since tests of asymmetry did not indicate significant publication bias, the exclusion of grey literature should not have altered the major findings of the present study.

Another issue concerns the moderator analyses. The moderation effects should be interpreted cautiously due to the relatively small number of included studies in some groups and, therefore, limited statistical power. Additionally, based on existing theories and literature, only a few moderators most relevant to our research questions were selected. Other factors, such as load and modality of the complex span task, could also influence its correlation with updating tasks [21]. However, it is beyond the scope of the present study to thoroughly explore all moderating factors, and future studies addressing these factors can provide a deeper understanding of the relationship between updating and WMC measurements.

5. Conclusions

To sum up, the present meta-analysis indicated that while both MU tasks could predict higher-order cognitive functions, the recollection-based RM was more closely correlated with WMC, the most commonly used WMC measure, than the recognition-based n-back. We argue that during n-back, the representative executive process of WM, namely FoA, is less involved than during RM, and RM can, therefore serve better as a valid WM measure in behavioral studies on individual differences.

Supplementary Materials

The following supporting information can be found at: https://www.sciepublish.com/article/pii/421, Figure S1: Forest plot of correlation of n-back and WMC; Figure S2: Forest plot of correlation of RM and WMC; Figure S3: Forest plot of correlation of n-back and *Gf*; Figure S4: Forest plot of correlation of RM and *Gf*; Figure S5: Funnel Plot of n-back and WMC; Figure S6: Funnel Plot of RM and WMC; Figure S7: Funnel Plot of n-back and *Gf*; Figure S8: Funnel Plot of RM and *Gf*; Table S1: Search strategies; Table S2: Basic information for studies included in the meta-analysis (n-back and WMC); Table S3: Basic information for studies included in the meta-analysis (RM and WMC); Table S4: Basic information for studies included in the meta-analysis (n-back and *Gf*); Table S5: Basic information for studies included in the meta-analysis (RM and *Gf*).

Author Contributions

Conceptualization, C.W., J.W. and T.C.; Methodology, J.W. and X.Z.; Software, J.W. and X.Z.; Validation, C.W.; Formal Analysis, J.W. and X.Z.; Investigation, C.W. and X.Z.; Writing—Original Draft Preparation, C.W. and J.W.; Writing—Review & Editing, T.C. and N.X.Y.; Visualization, C.W. and X.Z.; Supervision, T.C. and N.X.Y.; Project Administration, T.C.; Funding Acquisition, T.C.

Ethics Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

All data analyzed during the current study are available from the corresponding author upon reasonable request.

Funding

This research was funded by the National Key Research and Development Program of China grant number 2020YFC2003000.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Baddeley A. Working Memory. *Science* **1992**, *255*, 556–559. doi:10.1126/science.1736359.
2. Morris N, Jones DM. Memory updating in working memory: The role of the central executive. *Brit. J. Psychol.* **1990**, *81*, 111–121. doi:10.1111/j.2044-8295.1990.tb02349.x.
3. Conway ARA, Kane MJ, Bunting MF, Hambrick DZ, Wilhelm O, Engle RW. Working memory span tasks: A methodological review and user's guide. *Psychon. B Rev.* **2005**, *12*, 769–786. doi:10.3758/BF03196772.
4. Unsworth N, Heitz RP, Schrock JC, Engle RW. An automated version of the operation span task. *Behav. Res. Methods* **2005**, *37*, 498–505. doi:10.3758/BF03192720.
5. Nyberg L, Eriksson J. Working memory: maintenance, updating, and the realization of intentions. *CSH Perspect. Biol.* **2016**, *8*, a021816.
6. Pollack I, Johnson LB, Knaff PR. Running memory span. *J. Exp. Psychol.* **1959**, *57*, 137–146. doi:10.1037/h0046137.
7. Gray JR, Chabris CF, Braver TS. Neural mechanisms of general fluid intelligence. *Nat. Neurosci.* **2003**, *6*, 316–322. doi:10.1038/nn1014.
8. Broadway JM, Engle RW. Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behav. Res. Methods* **2010**, *42*, 563–570. doi:10.3758/BRM.42.2.563.
9. Salthouse TA. Relations between running memory and fluid intelligence. *Intelligence* **2014**, *43*, 1–7. doi:10.1016/j.intell.2013.12.002.
10. Collette F, Van der Linden M. Brain imaging of the central executive component of working memory. *Neurosci. Biobehav. Rev.* **2002**, *26*, 105–125. doi:10.1016/S0149-763400063-X.
11. Collette F, Van der Linden M, Laureys S, Arigoni F, Delfiore G, Degueldre C, et al. Mapping the Updating Process: Common and Specific Brain Activations Across Different Versions of the Running Span Task. *Cortex* **2007**, *43*, 146–158. doi:10.1016/S0010-945270452-0.
12. Roussel M, Dujardin K, Hénon H, Godefroy O. Is the frontal dysexecutive syndrome due to a working memory deficit? Evidence from patients with stroke. *Brain* **2012**, *135*, 2192–2201. doi:10.1093/brain/aws132.
13. Kane MJ, Conway ARA, Miura TK, Colflesh GJH. Working memory, attention control, and the n-back task: A question of construct validity. *J. Exp. Psychol. Learn.* **2007**, *33*, 615–622. doi:10.1037/0278-7393.33.3.615.
14. Jonides J, Schumacher EH, Smith EE, Lauber EJ, Awh E, Minoshima S, et al. Verbal Working Memory Load Affects Regional Brain Activation as Measured by PET. *J. Cogn. Neurosci.* **1997**, *9*, 462–475. doi:10.1162/jocn.1997.9.4.462.
15. Mencarelli L, Neri F, Momi D, Menardi A, Rossi S, Rossi A, et al. Stimuli, presentation modality, and load-specific brain activity patterns during n-back task. *Hum. Brain Mapp.* **2019**, *40*, 3810–3831. doi:10.1002/hbm.24633.
16. Owen AM, McMillan KM, Laird AR, Bullmore E. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Hum. Brain Mapp.* **2005**, *25*, 46–59. doi:10.1002/hbm.20131.
17. Ira E, Zanoni M, Ruggeri M, Dazzan P, Tosato S. COMT, neuropsychological function and brain structure in schizophrenia: A systematic review and neurobiological interpretation. *J. Psychiatry Neurosci.* **2013**, *38*, 366–380. doi:10.1503/jpn.120178.
18. Jaeggi SM, Buschkuehl M, Perrig WJ, Meier B. The concurrent validity of the N-back task as a working memory measure. *Memory* **2010**, *18*, 394–412. doi:10.1080/09658211003702171.
19. Miller KM, Price CC, Okun MS, Montijo H, Bowers D. Is the N-Back Task a Valid Neuropsychological Measure for Assessing Working Memory? *Arch. Clin. Neuropsych.* **2009**, *24*, 711–717. doi:10.1093/arclin/acp063.
20. Oberauer K. Binding and Inhibition in Working Memory: Individual and Age Differences in Short-Term Recognition. *J. Exp. Psychol. Gen.* **2005**, *134*, 368–387. doi:10.1037/0096-3445.134.3.368.

21. Redick TS, Lindsey DRB. Complex span and n-back measures of working memory: A meta-analysis. *Psychon. B Rev.* **2013**, *20*, 1102–1113. doi:10.3758/s13423-013-0453-9.
22. Schmiedek F, Hildebrandt A, Lövdén M, Wilhelm O, Lindenberger U. Complex span versus updating tasks of working memory: The gap is not that deep. *J. Exp. Psychol. Learn.* **2009**, *35*, 1089–1096. doi:10.1037/a0015730.
23. Bunting M, Cowan N, Saults JS. How does running memory span work? *Q. J. Exp. Psychol.* **2006**, *59*, 1691–1700. doi:10.1080/17470210600848402.
24. Chen Y-N, Mitra S, Schlaghecken F. Sub-processes of working memory in the N-back task: An investigation using ERPs. *Clin. Neurophysiol.* **2008**, *119*, 1546–1559. doi:10.1016/j.clinph.2008.03.003.
25. Szmalec A, Verbruggen F, Vandierendonck A, Kemps E. Control of interference during working memory updating. *J. Exp. Psychol. Hum. Percept. Perform.* **2011**, *37*, 137–151. doi:10.1037/a0020365.
26. Towse JN, Cowan N, Hitch GJ, Horton NJ. The Recall of Information from Working Memory. *Exp. Psychol.* **2008**, *55*, 371–383. doi:10.1027/1618-3169.55.6.371.
27. Redick TS, Broadway JM, Meier ME, Kuriakose PS, Unsworth N, Kane MJ, et al. Measuring Working Memory Capacity with Automated Complex Span Tasks. *Eur. J. Psychol. Assess.* **2012**, *28*, 164–171. doi:10.1027/1015-5759/a000123.
28. Chen T, Li D. The Roles of Working Memory Updating and Processing Speed in Mediating Age-related Differences in Fluid Intelligence. *Aging Neuropsychol. Cogn.* **2007**, *14*, 631–646. doi:10.1080/13825580600987660.
29. Conway ARA, Kane MJ, Engle RW. Working memory capacity and its relation to general intelligence. *Trends Cogn. Sci.* **2003**, *7*, 547–552. doi:10.1016/j.tics.2003.10.005.
30. Friedman NP, Miyake A, Corley RP, Young SE, DeFries JC, Hewitt JK. Not All Executive Functions Are Related to Intelligence. *Psychol. Sci.* **2006**, *17*, 172–179. doi:10.1111/j.1467-9280.2006.01681.x.
31. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. doi:10.1136/bmj.n71.
32. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* **2021**, *372*, n160. doi:10.1136/bmj.n160.
33. McElree B. Working memory and focal attention. *J. Exp. Psychol. Learn. Mem. Cogn.* **2001**, *27*, 817–835.
34. Cowan N, Elliott EM, Saults JS, Morey CC, Mattox S, Hismjatullina A, et al. On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cogn. Psychol.* **2005**, *51*, 42–100. doi:10.1016/j.cogpsych.2004.12.001.
35. Rey-Mermet A, Gade M, Souza AS, von Bastian CC, Oberauer K. Is executive control related to working memory capacity and fluid intelligence? *J. Exp. Psychol. Gen.* **2019**, *148*, 1335–1372. doi:10.1037/xge0000593.
36. Fragkos KC, Tsagris M, Frangos CC. Publication Bias in Meta-Analysis: Confidence Intervals for Rosenthal's Fail-Safe Number. *Int. Sch. Res. Not.* **2014**, *2014*, 825383. doi:10.1155/2014/825383.
37. Becker BJ. Failsafe N or File-Drawer Number. In *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*; Rothstein HR, Sutton AJ, Borenstein M, Eds.; Wiley: West Sussex, UK, 2005; pp. 111–125. doi:10.1002/0470870168.ch7.
38. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* **1997**, *315*, 629–634. doi:10.1136/bmj.315.7109.629.
39. Duval S, Tweedie R. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* **2000**, *56*, 455–463.
40. Cohen J. Differences between Correlation Coefficients. In *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Routledge: New York, NY, USA, 1988; pp. 109–143.
41. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* **2003**, *327*, 557–560. doi:10.1136/bmj.327.7414.557.
42. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychol. Methods* **2006**, *11*, 193–206. doi:10.1037/1082-989X.11.2.193.
43. Deeks JJ, Higgins JPT, Altman DG, on behalf of the Cochrane Statistical Methods Group. Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions*, 2nd ed.; Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., Eds.; Wiley: Hoboken, NJ, USA, 2019; pp. 241–284. doi:10.1002/9781119536604.ch10.
44. Wilhelm O, Hildebrandt AH, Oberauer K. What is working memory capacity, and how can we measure it? *Front. Psychol.* **2013**, *4*, 433. doi:10.3389/fpsyg.2013.00433.
45. Friedman NP, Miyake A. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex* **2017**, *86*, 186–204. doi:10.1016/j.cortex.2016.04.023.
46. Shelton JT, Elliott EM, Hill BD, Calamia MR, Gouvier WD. A comparison of laboratory and clinical working memory tests and their prediction of fluid intelligence. *Intelligence* **2009**, *37*, 283–293. doi:10.1016/j.intell.2008.11.005.
47. Shelton JT, Metzger RL, Elliott EM. A group-administered lag task as a measure of working memory. *Behav. Res. Methods* **2007**, *39*, 482–493. doi:10.3758/BF03193017.
48. Cowan N. An Embedded-Processes Model of Working Memory. In *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*; Miyake A, Shah P, Eds.; Cambridge University Press: Cambridge, UK, 1999; pp. 62–101. doi:10.1017/CBO9781139174909.006.

49. Oberauer K. Access to information in working memory: Exploring the focus of attention. *J. Exp. Psychol. Learn.* **2002**, *28*, 411–421. doi:10.1037/0278-7393.28.3.411.
50. Juvina I, Taatgen NA. Modeling Control Strategies in the n-back Task. In *Proceedings of the 8th International Conference on Cognitive Modeling*; Psychology Press: New York, NY, USA, 2007.
51. Kessler Y, Oberauer K. Working memory updating latency reflects the cost of switching between maintenance and updating modes of operation. *J. Exp. Psychol. Learn.* **2014**, *40*, 738–754. doi:10.1037/a0035545.
52. Hockey A, Geffen G. The concurrent validity and test–retest reliability of a visuospatial working memory task. *Intelligence* **2004**, *32*, 591–605. doi:10.1016/j.intell.2004.07.009.
53. Engle RW, Tuholski SW, Laughlin JE, Conway ARA. Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *J. Exp. Psychol. Gen.* **1999**, *128*, 309–331. doi:10.1037/0096-3445.128.3.309.
54. Diamond A. Executive Functions. *Annu. Rev. Psychol.* **2013**, *64*, 135–168. doi:10.1146/annurev-psych-113011-143750.
55. Burgess GC, Gray JR, Conway ARA, Braver TS. Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *J. Exp. Psychol. Gen.* **2011**, *140*, 674–692. doi:10.1037/a0024695.
56. Shipstead Z, Harrison TL, Engle RW. Working memory capacity and the scope and control of attention. *Atten. Percept. Psychophys.* **2015**, *77*, 1863–1880. doi:10.3758/s13414-015-0899-0.
57. Shipstead Z, Lindsey DRB, Marshall RL, Engle RW. The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *J. Mem. Lang.* **2014**, *72*, 116–141. doi:10.1016/j.jml.2014.01.004.
58. Unsworth N, Fukuda K, Awh E, Vogel EK. Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cogn. Psychol.* **2014**, *71*, 1–26. doi:10.1016/j.cogpsych.2014.01.003.
59. Kovacs K, Conway ARA. Process Overlap Theory: A Unified Account of the General Factor of Intelligence. *Psychol. Inq.* **2016**, *27*, 151–177. doi:10.1080/1047840X.2016.1153946.
60. Ecker UK, Lewandowsky S, Oberauer K, Chee AE. The components of working memory updating: an experimental decomposition and individual differences. *JEP LMC* **2010**, *36*, 170.
61. Mackintosh NJ, Bennett ES. IT, IQ and perceptual speed. *Pers. Individ. Dif.* **2002**, *32*, 685–693. doi:10.1016/S0191-886900069-1.