Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-025-02721-6

COMMENTARY

Complex span and the *n*-back lack convergent validity as measures of working memory: Reply to Wilhelm et al. (2025)

Alexander P. Burgoyne[1] · David J. Frank[2] · Brooke N. Macnamara[3]

Accepted: 22 May 2025

Abstract
In our target article, "Which Working Memory Are We Talking About? *N*-Back vs. Complex Span Tests," we analyzed data from 1,272 participants and demonstrated that complex span and *n*-back tasks lack convergent validity as measures of working memory. Evidence for their dissociation included 1) exploratory factor analyses revealing two distinct factors with near-zero cross-loadings, 2) confirmatory factor analyses showing these factors share one-fifth of their reliable variance, and 3) both factors correlating more strongly with fluid intelligence than with each other. Structural equation modeling demonstrated that *n*-back and complex span factors each explained significant unique variance in fluid intelligence (24% and 14% respectively), beyond their jointly explained variance (30%). These findings align with previous meta-analytic results and support a theoretical framework where complex span tasks emphasize information maintenance while *n*-back tasks require rapid disengagement from outdated information. Our analyses extended beyond method-specific effects by replicating these results at the broader construct level with additional measures of updating and working memory capacity. In their commentary, Wilhelm et al.'s alternative single-factor model suggests a near-perfect association between working memory and fluid intelligence ($\beta = .97$). Their model relies on inconsistently applied correlated error terms selected through a data-driven approach. Notably, modification indices suggest improvements to their model that would bring it closer to our two-factor structure, consisting of clusters of measures representing working memory capacity on one hand and updating on the other. Recognizing these distinctions advances our understanding of cognitive abilities and helps avoid the jingle fallacy.

Keywords Working memory capacity · Updating · Fluid intelligence · Maintenance · Disengagement

In our target article, "Which Working Memory Are We Talking About? *N*-Back vs. Complex Span Tests" (Burgoyne et al., 2024), we found strong evidence for a dissociation between *n*-back and complex span tasks, and more generally, between latent variables reflecting updating and working memory capacity. We concluded that the two constructs are separable, and that researchers should not use *n*-back and complex span measures interchangeably. Our conclusions were based on data from a diverse sample of 1,272 participants recruited from local communities and several universities that varied in selectivity. We review the evidence below.

Evidence for the dissociation between *n*-back and complex span measures

We conducted multiple analyses to investigate the convergent validity (or lack thereof) between *n*-back and complex span measures. The results are summarized in Fig. 1. First, an exploratory factor analysis (EFA) on the *n*-back and complex span tests revealed two distinct factors, as did an EFA on a broader set of updating and working memory capacity measures. In both cases, a two-factor solution was preferred over a one-factor solution based on the results of parallel analyses and eigenvalues exceeding 1.0. In the first case, *n*-back and complex span measures loaded highly on separate factors with cross-loadings that approached zero (ranging from −0.02 to 0.03). After including additional updating

Related article can also be found at https://doi.org/10.3758/s13423-024-02622-0.

✉ Alexander P. Burgoyne
aburgoyne@humrro.org; burgoyn4@gmail.com

[1] Human Resources Research Organization (HumRRO), 66 Canal Center Plaza, Suite 700, Alexandria, VA 22314, USA

[2] Youngstown State University, Youngstown, OH, USA

[3] Purdue University, West Lafayette, IN, USA

Published online: 25 June 2025

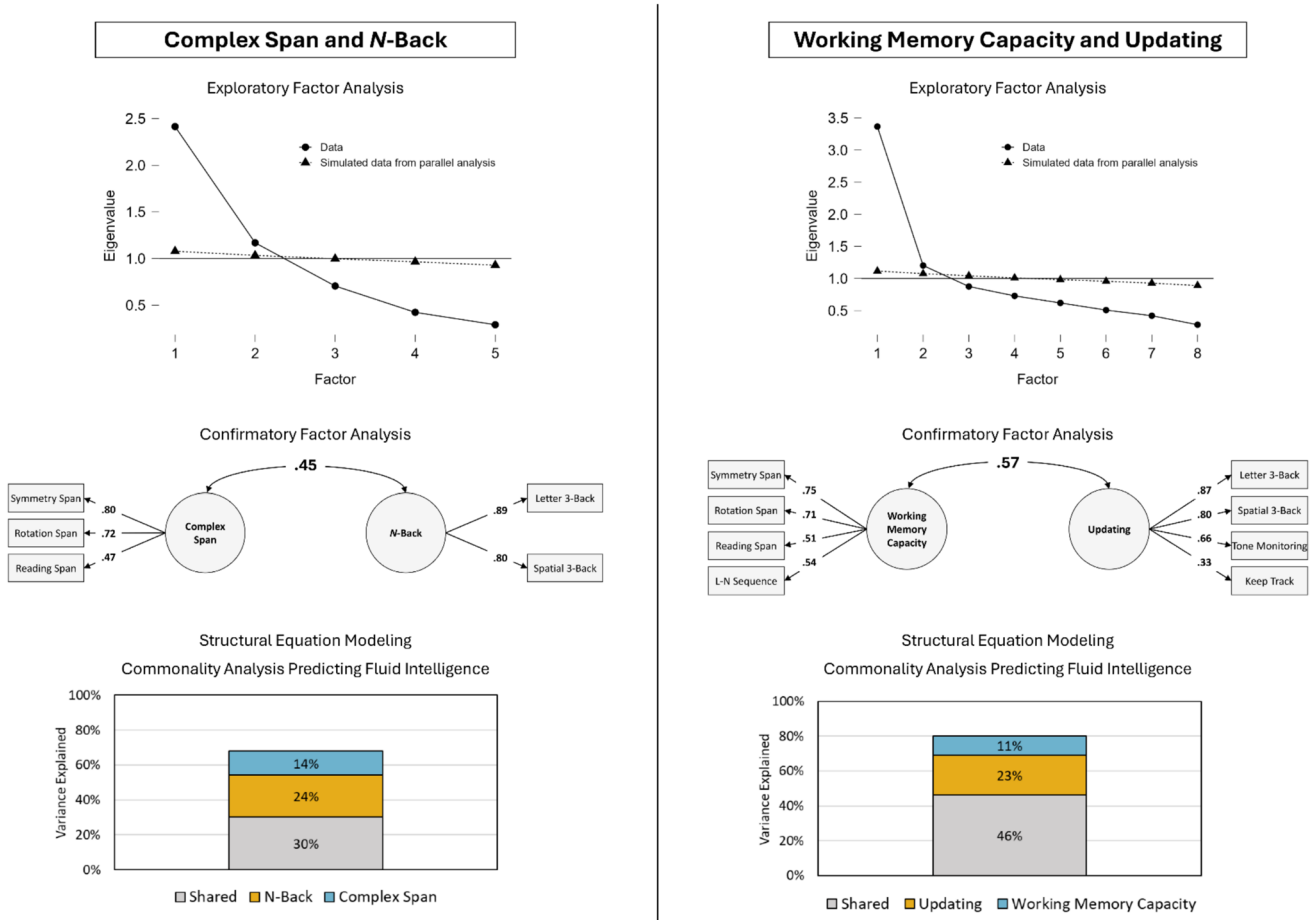

Fig. 1 Summary of results from Burgoyne et al. (2024)

and working memory capacity measures, again a two-factor structure emerged, and this time the *n*-back and complex span measures' cross-loadings became slightly more negative (ranging from −0.15 to −0.08). This indicates a trend towards greater dissociability when the measures are considered at the broader construct level.

Second, confirmatory factor analysis (CFA) on the *n*-back and complex span tests revealed the two factors correlated at $r = .45$ ($R^2 = 20\%$). In other words, after partialing out measurement error, they share one-fifth of their reliable variance. Not surprisingly, one cannot constrain this correlation to 1.0 without significant loss in model fit. This pattern was replicated at the construct level with additional measures of updating and working memory capacity; the latent correlation of $r = .57$ ($R^2 = 32\%$) indicates that updating and working memory capacity share one-third of their reliable variance.

Third, when we included a latent factor with four measures of fluid intelligence (i.e., novel problem-solving ability) in the CFA, we found that both complex span and *n*-back factors had numerically stronger correlations with fluid intelligence than they did with each other (compare rs = .66 and .74 to $r = .45$; or at the broader construct level, compare rs = .76 and .84 to $r = .61$).

Fourth, structural equation modeling (SEM) demonstrated that *n*-back and complex span factors each explained significant unique variance in fluid intelligence. Using commonality analysis (Nimon et al., 2008), we disentangled the unique contribution of each predictor to fluid intelligence from the shared variance that was jointly explained by them. While the two factors jointly explained 30% of the variance in fluid intelligence, the *n*-back factor uniquely explained an additional 24% and the complex span factor uniquely explained an additional 14%. Stated differently, not only did the *n*-back and complex span factors each explain significant unique variance in fluid intelligence, but the two factors' unique variance was greater than their shared explanatory variance (i.e., 38% shared vs. 30% unique, for a total R^2 of 68%).

We took these results as evidence for a dissociation between *n*-back and complex span measures, and more broadly, for a dissociation between updating and working memory capacity. These results are difficult to reconcile with claims of convergent validity because the evidence

convincingly suggests the opposite: EFAs identified separable factors with minimal cross-loadings; CFAs indicated the factors share one-fifth to one-third of their reliable variance and share more variance with fluid intelligence than with each other; SEMs revealed that the two constructs account for substantial unique variance in fluid intelligence above and beyond each other, demonstrating incremental validity.

These results align with findings from Redick and Lindsey's (2013) meta-analysis and a recent large-scale latent variable analysis conducted by Robison et al. (2024). Redick and Lindsey (2013) meta-analyzed 20 studies reporting a correlation between *n*-back and complex span performance ($N = 2{,}178$) and found a meta-analytic average correlation of $r = .20$, 95% CI [.16, .24]. Robison et al. (2024) administered several *n*-back and complex span tasks to more than 700 participants and found a latent correlation of $r = .43$: essentially the same result as the $r = .45$ in our target article. Our conclusions are strengthened by the fact that Redick and Lindsey (2013) and Robison et al. (2024) used different sets of tasks and analyses yet converged on the same conclusion.

Interpreting the results through the disengagement hypothesis

N-back and complex span tasks have conceptual and empirical differences. Optimistically, these differences can be leveraged to advance the state of the science regarding our understanding of cognitive performance in humans. First and foremost, a clearer understanding of the differences between measures can lead to better insights regarding past, present, and future research on cognitive abilities, helping to avoid the jingle fallacy. Second, differences between measures can serve as a useful testbed for theory development.

In our target article, we noted that the results were consistent with the *disengagement hypothesis*, which is an emerging framework for understanding relations between working memory, attention control, and fluid intelligence in terms of maintenance and disengagement (Burgoyne & Engle, 2020; Shipstead & Engle, 2013; for a review of evidence, see Shipstead et al., 2016). The basic idea underlying the theory is that some tasks emphasize maintaining the focus of attention and remembering task-relevant information, while other tasks emphasize disengaging from outdated information and preventing prepotent responses. According to the framework, attention control supports maintenance and disengagement; complex span tests demand maintaining information amidst interference from a secondary task, whereas fluid intelligence tests demand disengaging from ruled-out hypotheses to explore novel solutions (Burgoyne et al., 2019; Shipstead et al., 2016). The *n*-back presents an interesting foil for this framework because it requires remembering "*n*" memory items but also rapidly disengaging as each new item replaces a previous one. For most people, remembering three items is well within the limits of working memory capacity (Cowan, 2010), so what makes the *n*-back challenging is having to rapidly update and discard memory items that were once relevant. That is why the strong relationship we observed between the *n*-back (or updating) and fluid intelligence is particularly fascinating: it suggests that disengagement may be particularly important to novel problem-solving performance, although both maintenance and disengagement clearly play a role.

Another example of a task that requires disengagement is visual arrays/change detection. This task is often conceptualized as a test of visual working memory capacity (e.g., "remember these colored shapes and detect the change"). Shipstead and Engle (2013) showed that increasing temporal discriminability across trials using an intertrial interval led to higher capacity estimates by reducing proactive interference. In other words, disengaging from prior trials improved performance on a task that ostensibly emphasizes information maintenance. Crucially, temporal discriminability selectively benefited higher-fluid intelligence participants, not lower-fluid intelligence participants. It appears that higher-fluid intelligence performers were able to disengage from outdated memory items when given additional time to do so, establishing a link between problem-solving ability and the ability to disengage from outdated information.

As we noted in our target article, other theoretical frameworks are also consistent with the results we obtained. Our study was primarily focused on the convergent validity (or lack thereof) of updating and working memory capacity measures, and was not designed to provide disconfirmatory evidence for the disengagement hypothesis or alternative theoretical frameworks. Nevertheless, the results compelled a theoretical explanation and are consistent with predictions made by the disengagement hypothesis and other frameworks.

Reply to Wilhelm, Thelen, and Schmiedek (2025)

In their commentary, Wilhelm, Thelen, and Schmiedek (WTS; 2025) take an alternative approach to analyzing our data and arrive at a model that reflects a near-perfect association between working memory and fluid intelligence. Their model contained a single working memory factor with a standardized predictive path of $\beta = 0.97$ to fluid intelligence ($R^2 = 94\%$). Although we appreciate the response our work has generated, there are several aspects of WTS's commentary, model, and conclusions that warrant further consideration.

In their narrative review, they begin by stating, "A first set of prominent studies (Kyllonen & Christal, 1990) found

the relation between WM and gf to range between .80 and .90. Improved analysis (Oberauer et al., 2005) of a popular meta-analysis (Ackerman et al., 2005) also found the relation between WM and gf to reach a magnitude of around .85. Many newer studies replicate this estimate (for instance Monteiro et al., 2025)" (p.1)

This review of the literature selects stronger working memory-fluid intelligence relationships than are typically found. Kyllonen and Christal's (1990) conclusion that working memory capacity and fluid intelligence were isomorphic came with a question mark and an exclamation point (?!). This claim has been tempered in the past 35 years, as most studies find a moderate-to-strong relation between fluid intelligence and working memory capacity, but one that is significantly and substantively weaker than 1.0. For example, Ackerman et al.'s (2005) meta-analysis reported an average observed correlation of $r = .48$. Kane et al.'s (2005) reanalysis of latent variable studies yielded a median correlation of $r = .72$. As another example, using random effects modeling, Oberauer et al. (2005) reclassified tasks from Ackerman et al. (2005) and found a standardized predictive path of $\beta = .85$ (an R^2 of 72%). More recently, Burgoyne et al. (2023) found a latent correlation of $r = .63$ in an online study and $r = .53$ in an in-lab study between complex span measures and fluid intelligence. And the large-scale study by Robison et al. (2024) found a latent correlation of $r = .66$ between complex span and fluid intelligence factors. In short, the suggestion that working memory capacity and fluid intelligence are nearly isomorphic is at odds with the available empirical evidence; most studies find latent correlations less than .80 and observed correlations much lower than that.

In their reanalysis, WTS created a single working memory factor with loadings on all the complex span measures, *n*-back measures, and additional measures tapping working memory capacity and updating. WTS also specified two critical correlated error terms, one between the *n*-back tasks, and another between two of the three complex span tests (Symmetry Span and Rotation Span). The third complex span test, Reading Span, was not allowed any correlated error terms, despite sharing common method variance with the other complex span tests. Therefore, the complex span residual correlation in WTS's model appears to account for visuospatial content, not common-method variance. By contrast, the correlated residuals of the Spatial *N*-Back and Letter *N*-Back measures serve to partial out common-method variance related to the *n*-back procedure. Thus, the rationale for the two sets of correlated error terms appears inconsistent: one captures content-specific variance while the other captures method-specific variance. These correlated error terms were apparently selected using a data-driven approach: "We specified a general factor model with one WM factor and two correlated errors for the two exceptionally high correlations mentioned above" (p.2).

From this data-driven approach, WTS emphasized that their model has better fit statistics than the models we presented in our target article. We explored whether other data-driven models might fit even better. We specified their model and examined the modification indices, which provided suggestions that would lead to further improvements in model fit. Interestingly, these suggestions make their model more similar to our original SEM.

Below, we share the table of modification indices for WTS's model (Table 1). The first suggestion is to correlate the error terms between the two complex span tasks Symmetry Span and Reading Span. The second suggestion is to correlate the error terms between the complex span task Reading Span and the additional working memory capacity test, Letter–Number Sequencing. Taken together, these correlated error terms among the working memory capacity measures lead to a model that comes close to replicating the working memory capacity factor we originally specified.

The third suggestion is to correlate the error terms between Letter *N*-Back and one of the two additional updating tests, Tone Monitoring. Correlating error terms among these updating tasks comes close to replicating the updating factor that we originally specified. All three of these modifications result in statistically significant improvements in model fit above and beyond WTS's proposed model. In short, if a data-driven approach is to be taken as diagnostic, the data suggest two correlated clusters with working memory capacity tests on one hand and updating tests on the other for a better model fit.

WTS state that working memory capacity and fluid intelligence are not exactly but are almost the same. For example, despite the very strong association they observed between their single working memory factor and fluid intelligence (i.e., $\beta = .97$), it cannot be set to 1.0 without

Table 1 Modification indices for the model in Wilhelm et al.'s (2025) reanalysis

Measure 1	Relation	Measure 2	Modification Index	EPC
Symmetry Span	Correlation	Reading Span	40.839	0.135
Reading Span	Correlation	Letter–Number Sequencing	29.033	0.123
Letter *N*-Back	Correlation	Tone Monitoring	20.825	0.068

Note: EPC = expected parameter change (standardized)

loss in model fit. They go on to note several differences between working memory and fluid intelligence. However, most of these differences were not applicable to our dataset.

First, they state that working memory tests are always somewhat speeded, implying that fluid intelligence tests may not be. In our study, all the fluid intelligence tests had a time limit (e.g., 2 min for Cattell's test, 5 min for number series, 7 min for letter sets, and 10 min for Raven's matrices). Thus, speeded performance is a characteristic that was shared across the working memory and fluid intelligence tests.

Second, the authors state that working memory tests are always administered via computer, whereas fluid intelligence tests may or may not be. In our study, all the tests we administered were programmed in E-Prime and delivered via computer. Thus, computerized assessment is a characteristic that was shared across the working memory and fluid intelligence tests. Additionally, this claim is historically inaccurate: the original complex span test of working memory was developed for analog administration, not computerized administration (see Daneman & Carpenter, 1980, for the original version of the Reading Span test).

Third, WTS state that knowledge helps performance in fluid intelligence tasks but not in working memory tasks. We note that one of the defining characteristics of fluid intelligence is that it represents *novel* problem-solving ability, and efforts are made to remove the influence of knowledge from test content. Though sometimes using stimuli that represent activated long-term knowledge (e.g., numbers, letters) or requiring relatively simple acquired quantitative abilities (e.g., number series task), the goal of fluid intelligence measures is to assess individuals' problem-solving performance on a decontextualized, novel playing field (Schneider & McGrew, 2018). In contrast, working memory tests frequently use stimuli (e.g., letters or numbers) that represent activated long-term knowledge, and studies have shown that working memory performance is reduced when stimuli are unfamiliar and have no associated knowledge (e.g., using unfamiliar Klingon or Chinese characters; see Hicks et al., 2016; Zimmer & Fischer, 2020). Additionally, Reading Span requires considerable knowledge to make judgments about whether each sentence makes sense. Thus, in our study, 1) stimuli that represent activated knowledge were present in both the fluid intelligence and working memory measures, and 2) if anything, knowledge likely played more of a role in working memory, not fluid intelligence performance.

Finally, the authors state that fluid intelligence tasks might stress concept formation, which they claim is "irrelevant" in working memory tasks. We note that elaboration is one strategy test-takers use to aid performance on working memory tests, and the act of elaboration can be seen as an example of concept formation (Gobet, 1998).

To summarize, Wilhelm et al. (2025) offer several differences between working memory and fluid intelligence, but most of them are not applicable to the current work or contain inaccuracies. Moreover, in our view, focusing on these details overlooks the primary distinguishing characteristics of the constructs. From our standpoint, fluid intelligence is about novel problem solving, working memory capacity is about maintenance of information amidst interference, and updating is about disengaging from outdated information in service of a goal.

Conclusion

The evidence suggests a dissociation between *n*-back and complex span measures, and between updating and working memory capacity factors more broadly. The two sets of measures, and the broader constructs they represent, are not interchangeable. We think this distinction will be useful to psychologists and neuroscientists who might otherwise use the measures interchangeably. The results also suggest a compelling link between updating ability and fluid intelligence which can be used to inform theoretical developments in the field.

Authors' Contributions [anonymized initials]: Conceptualization (lead), Formal Analysis (lead), Funding Acquisition (supporting), Project Administration (equal), Visualization (lead), Writing–Original Draft Preparation (lead), Writing–Review & Editing (lead). [anonymized initials]: Conceptualization (equal), Funding Acquisition (supporting), Project Administration (equal), Writing–Original Draft Preparation (supporting), Writing–Review & Editing (supporting). [anonymized initials]: Conceptualization (equal), Funding Acquisition (lead), Project Administration (equal), Writing–Original Draft Preparation (supporting), Writing–Review & Editing (supporting).

Funding This work was sponsored in part by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) and was accomplished under Grant # W911 NF-22–1-0226. The views, opinions, and/or findings contained in this paper are those of the authors and shall not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documents. Data collected prior to obtaining this grant are also included in this paper.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflicts of interest The authors declare no conflicts of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131*(1), 30–60.

Burgoyne, A. P., & Engle, R. W. (2020). Attention control: A cornerstone of higher-order cognition. *Current Directions in Psychological Science, 29*(6), 624–630.

Burgoyne, A. P., Tsukahara, J. S., Mashburn, C. A., Pak, R., & Engle, R. W. (2023). Nature and measurement of attention control. *Journal of Experimental Psychology: General, 152*(8), 2369–2402.https://doi.org/10.1037/xge0001408

Burgoyne, A. P., Frank, D. J., & Macnamara, B. N. (2024). Which "working memory" are we talking about? Complex span tasks versus *N*-back. *Psychonomic Bulletin & Review*, 1–15. Advance online publication. https://doi.org/10.3758/s13423-024-02622-0

Burgoyne, A. P., Hambrick, D. Z., & Altmann, E. M. (2019). Place-keeping ability as a component of fluid intelligence: Not just working memory capacity. *The American Journal of Psychology, 132*(4), 439–449.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*(1), 51–57.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450–466.

Gobet, F. (1998). Expert memory: A comparison of four theories. *Cognition, 66*, 115–152.

Hicks, K. L., Foster, J. L., & Engle, R. W. (2016). Measuring working memory capacity on the web with the online working memory lab (the OWL). *Journal of Applied Research in Memory and Cognition, 5*(4), 478–489.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 66–71.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?!.*Intelligence, 14*(4), 389–433. https://doi.org/10.1016/S0160-2896(05)80012-1

Monteiro, F., Nascimento, L. B., Leitão, J. A., Santos, E. J. R., Rodrigues, P., Santos, I. M., . . . & Nascimento, C. S. (2025). Optimizing working memory assessment: Development of shortened versions of complex spans, updating, and binding tasks.*Psychological Research, 89*(2), 65. https://doi.org/10.1007/s00426-025-02083-7

Nimon, K., Lewis, M., Kane, R., & Haynes, R. M. (2008). An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example. *Behavior Research Methods, 40*, 457–466.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H. M. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 61–65.

Redick, T. S., & Lindsey, D. R. (2013). Complex span and *n*-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review, 20*, 1102–1113.

Robison, M. K., Miller, A. L., Wiemers, E. A., Ellis, D. M., Unsworth, N., Redick, T. S., & Brewer, G. A. (2024). What makes working memory work? A multifaceted account of the predictive power of working memory capacity. *Journal of Experimental Psychology: General, 153*(9), 2193–2215. https://doi.org/10.1037/xge0001629

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). Guilford Press.

Shipstead, Z., & Engle, R. W. (2013). Interference within the focus of attention: Working memory tasks reflect more than temporary maintenance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(1), 277–289.

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science, 11*(6), 771–799.

Wilhelm, O., Thelen, J., & Schmiedek, F. (2025). We are talking about WM as a broad ability factor! Comment on Burgoyne, Frank, and Macnamara (2024). *Psychonomic Bulletin & Review*, 1–4.

Zimmer, H. D., & Fischer, B. (2020). Visual working memory of Chinese characters and expertise: The expert's memory advantage is based on long-term knowledge of visual word forms. *Frontiers in Psychology, 11*, 516.