Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-024-02622-0

BRIEF REPORT

Which "working memory" are we talking about? Complex span tasks versus *N*-back

Alexander P. Burgoyne[1,2] · **David J. Frank**[3] · **Brooke N. Macnamara**[4,5]

Accepted: 18 November 2024

Abstract

Psychologists and neuroscientists often use complex span tasks or the *n*-back to measure working memory capacity. At first glance, both tasks require many cognitive processes attributed to the construct, including the maintenance of information amidst interference. Nevertheless, evidence for their convergent validity is mixed. This poses consequences for the interpretation of working memory performance in cognitive neuroscience, developmental psychology, applied psychology, and executive functioning research. We recruited a large and diverse sample using a multisite approach ($N = 1{,}272$; community and university participants) and had them complete multiple working memory capacity, updating, and fluid intelligence tests. We found strong evidence for a dissociation between complex span and *n*-back tests, and more broadly, between working memory capacity and updating factors. Observed correlations between complex span and *n*-back performance were modest ($\bar{r} = .25$), and at the latent level, the two factors only shared 20% of their variance. Each explained unique variance in fluid intelligence, and each was more strongly related to fluid intelligence than to each other, with updating measures demonstrating stronger relations to fluid intelligence. These results were interpreted via the *disengagement hypothesis*. What distinguishes updating measures from working memory capacity measures is their relative emphasis on disengagement from outdated information; disengagement drives their strong relation with fluid intelligence because problem-solving requires generating hypotheses but also discarding those discovered to be false. We suggest that researchers who want to measure and draw conclusions about working memory capacity or updating should not use complex span tasks and the *n*-back interchangeably.

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

Working memory refers to the cognitive system used to maintain information in service of goal-directed behavior (Baddeley, 1992; Baddeley & Hitch, 1974). Two types of tasks are frequently used to measure individual differences in working memory capacity—*complex span tasks* and the *n-back*—but it is unclear whether they measure the same ability, and if not, how they differ.

Complex span tasks challenge subjects to remember a sequence of items while completing an interleaved secondary task. They were developed following the discovery that short-term memory tests, which lack a secondary processing task, did not predict reading comprehension (see Daneman & Merikle, 1996). Daneman and Carpenter (1980) hypothesized that, rather than passive storage, the ability to maintain information *while processing* other information was crucial for comprehension. This ability mapped closely onto Baddeley and Hitch's (1974) concept of working memory, and complex span tasks such as reading span were developed to measure this construct (Daneman & Carpenter, 1980). In the decades that followed, complex span tasks have been frequently used to measure working memory capacity in cognitive psychology (Engle et al., 1999; Kyllonen & Christal, 1990; Miyake et al., 2001).

In neuroscientific research, by contrast, the *n*-back has been the most prevalent measure of working memory capacity (Owen et al., 2005; Wager & Smith, 2003). The task is

✉ Alexander P. Burgoyne
burgoyn4@gmail.com

[1] Human Resources Research Organization (HumRRO), Alexandria, VA 22314, USA

[2] Georgia Institute of Technology, Atlanta, GA 30332, USA

[3] Youngstown State University, Youngstown, OH 44555, USA

[4] Case Western Reserve University, Cleveland, OH 44106, USA

[5] Purdue University, West Lafayette, IN 47907, USA

Published online: 04 December 2024

particularly well-suited for fMRI studies because it affords control over the timing of stimuli. In the *n*-back, subjects are presented a continuous sequence of items and must identify whether the current item matches the item that was presented *n* items ago (e.g., three items ago). Accuracy rates decrease as *n* increases (Mackworth, 1959). One challenge is that, in addition to having to maintain *n* items, subjects must also rapidly update the contents of working memory to disengage from outdated information (Oberauer, 2009; Shipstead et al., 2016; Szmalec et al., 2011), otherwise the memory load quickly becomes unfeasible.

Kane et al. (2007) examined performance on the *n*-back and its relation to one complex span test (operation span) and one fluid intelligence test (Raven's matrices) in a sample of 129 undergraduates. Two-back performance did not correlate significantly with either measure, whereas three-back performance correlated descriptively more strongly with Raven's matrices than with operation span. Indeed, correlations between the *n*-back and operation span ranged from nonsignificant to weak (all $r\text{s} \leq .22$). Furthermore, operation span and the *n*-back each accounted for unique variance in fluid intelligence, suggesting a dissociation between the two measures. Kane et al. (2007) noted that the *n*-back demands speeded recognition, whereas complex span tests demand serial recall, and that these abilities may reflect different aspects of the working memory system.

Another distinction might be the relative contributions of *maintenance* versus *disengagement* (Burgoyne & Engle, 2020; Shipstead et al., 2016). In complex span tasks, the challenge is for subjects to maintain access to memory items while completing secondary distractor tasks. According to the executive attention view (e.g., Engle, 2002, 2018), the "capacity" that is measured by such tasks reflects the interplay between attention and short-term memory, but it is primarily the attentional component that drives the predictive validity of working memory capacity measures (e.g., Daneman & Carpenter, 1980; Engle et al., 1999). This attentional component maps closely onto Baddeley's concept of the central executive (e.g., Baddeley & Hitch, 1974; Baddeley & Logie, 1999). Within the context of complex span tasks, attention supports the maintenance of information amidst distraction and interference, and this is a major source of individual differences in performance.

By contrast, in the *n*-back, subjects must continuously update the contents of working memory to disengage from outdated information. An inability to disengage makes one susceptible to *far lures*—for instance, calling an item from six-back a "target" during a three-back task. For example, Shipstead et al. (2016) found that the negative correlation between fluid intelligence and *n*-back false alarms strengthened as the lure position increased, indicating that individuals with lower fluid intelligence were more likely to mistake far lures as targets. Attention is also hypothesized to support performance on updating tasks; however, its primary function is to remove no-longer relevant memory items from focus. Unlike complex span tasks, the memory set size used in *n*-back tasks is typically small (e.g., two or three items); maintenance is necessary, but the source of individual differences in performance is hypothesized to stem from differences in the ability to disengage from those outdated items that make one susceptible to far lures.

Shipstead et al.'s (2016) results fit into a framework we refer to as the "disengagement hypothesis." This framework views maintenance and disengagement as two critical factors linking executive functioning to fluid intelligence (Burgoyne & Engle, 2020; Burgoyne et al., 2019; Hambrick & Altmann, 2015; Shipstead et al., 2016). When subjects attempt to solve novel problems in fluid intelligence tests, they must generate hypotheses and discard incorrect hypotheses. An inability to disengage from outdated hypotheses leads to perseveration and stymies problem-solving. Both maintenance and disengagement are important for problem-solving (e.g., Carpenter et al., 1990), but the relative contribution of each could be driving the different relationships of complex span tests and the *n*-back with fluid intelligence, as well as their less-than-perfect correlation with each other (e.g., Kane et al., 2007).

If the disengagement hypothesis is correct, then not only should *n*-back tasks explain unique variance in fluid intelligence but so should other measures of updating. Furthermore, updating tasks, including those that do not share method-specific variance, should load together on a latent factor that is separable from one derived from tasks requiring maintenance more than disengagement (i.e., complex span and related tasks). For comparison, if the disengagement hypothesis is incorrect, then we would not expect updating tasks to capture unique variance in fluid intelligence above and beyond working memory capacity tasks. Furthermore, we would predict that updating and working memory capacity measures would not be distinguishable at the latent level; if this were the case, it would provide strong disconfirmatory evidence against the disengagement hypothesis.

Importantly, not all studies have shown strong dissociations between complex span tests and the *n*-back. Schmiedek et al. (2009) administered multiple complex span tests and updating tasks (including two *n*-back tests) that varied in stimulus content to a sample of 96 undergraduates and community participants. Using latent variable modeling, they found a correlation of $r = .96$ between factors representing complex span and updating performance—this correlation was not significantly different from 1.0. The authors concluded that "updating tasks measure [working memory] equally well as [complex span tests]" (p. 1095).

Several aspects of Schmiedek et al.'s (2009) results call their conclusion into question. Notably, two of the three complex span measures had weak, nonsignificant correlations with *n*-back performance, yet at the latent level, the constructs were perfectly correlated. At the observed level, the only complex span test that correlated significantly with *n*-back performance was rotation span, and both tasks used visuospatial memoranda. Commenting on Schmiedek et al.'s (2009) analyses, Redick and Lindsey (2013) observed that rotation span had a much higher loading on the complex span factor (.70) than did the other two complex span tasks (.34 and .37). Thus, rotation span was the primary "driver" of the complex span factor, which might help explain the near-perfect correlation Schmiedek et al. observed between complex span and updating factors.

In light of these conflicting results, Redick and Lindsey (2013) conducted a meta-analysis to clarify the relation between complex span and *n*-back performance. Across 20 studies (total $N=2{,}178$), they found that the strength of the correlation between complex span and *n*-back performance was stronger for visuospatial memory items than for verbal memory items. Overall, however, they observed only a relatively weak meta-analytic correlation, $\bar{r}=.20$. They concluded that complex span and *n*-back tasks should not be used interchangeably.

The present study

In this multisite study, we provide the largest primary test of the disengagement hypothesis and assessment of the relation between complex span, *n*-back, and fluid intelligence to date. We use multiple varied indicator measures, latent variable analyses, and a much larger sample ($N=1{,}272$) than past primary studies. Furthermore, we broaden the scope of the analysis to include additional indicators of working memory capacity and updating, permitting inferences that go beyond task-level observations (i.e., complex span vs. *n*-back) and get closer to the theoretical abilities underpinning performance. We test whether (1) complex span and *n*-back performance is highly correlated at the observed and latent level, (2) whether *n*-back performance is more strongly related to fluid intelligence than complex span performance, and (3) whether each set of measures accounts for unique variance in fluid intelligence. We repeat these analyses at the broader construct level (i.e., working memory capacity vs. updating) to resolve questions regarding the latent structure of cognitive abilities related to executive functioning.

Method

Participants

This study was part of a multisite research endeavor conducted at Case Western Reserve University, Georgia Institute of Technology, Michigan State University, Texas A&M University—Commerce, and Youngstown State University. We recruited undergraduate participants from each university. The universities varied in size, region, and admission selectivity. Further, we sampled participants from the greater Cleveland, Ohio, community. All participants were required to be 18–39 years of age. This study was approved by each university's Institutional Review Board. In total, 1,272 participants were included in the analytical sample.

Procedure

Data were collected as part of a larger project that consisted of two sessions lasting up to 2.5 h each. All cognitive tests considered in this manuscript were administered during Session 1 of the study. All tasks were programmed using E-Prime (Version 3.0). Further information regarding the scope of data collection can be found at the following link: (https://osf.io/c39y6). Participants received either research participation credits or financial compensation for their participation.

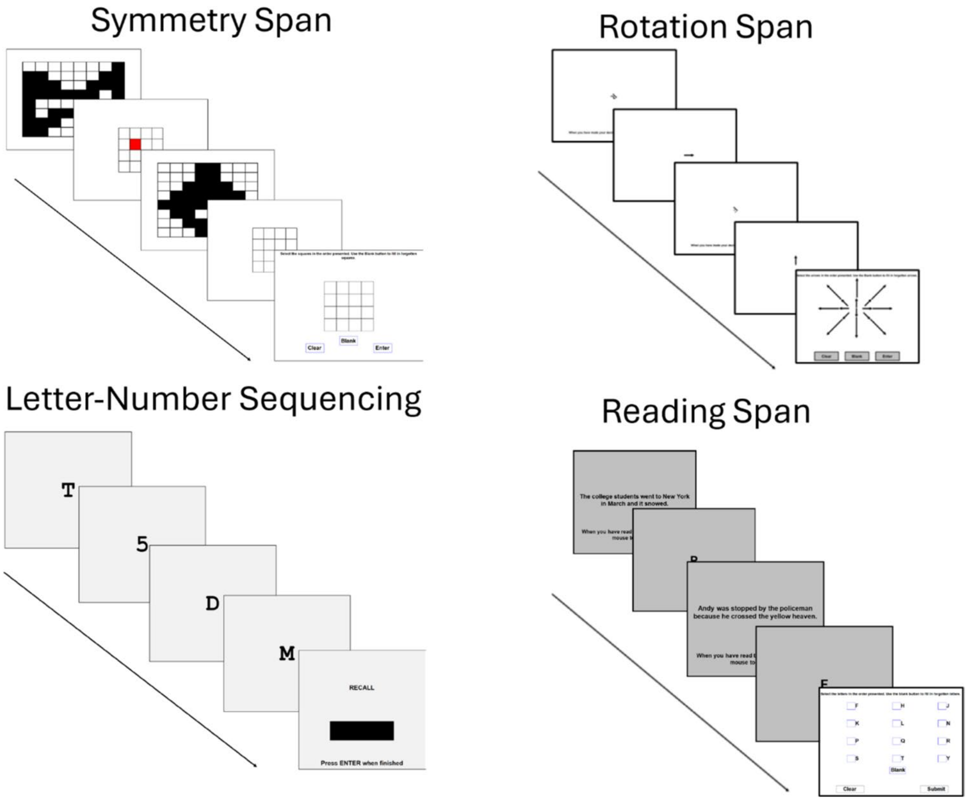
During data collection, participants were tested in small groups with a research assistant assigned to proctor each session. The research assistant ensured the participants understood the instructions. Research assistants took extensive notes on participant conduct, which were used to make decisions about data exclusions described below.

Working memory capacity

Figure 1 presents a schematic illustration of the four working memory capacity tasks (i.e., symmetry span, rotation span, reading span, and letter–number sequencing).

Symmetry span (Kane et al., 2004; Unsworth et al., 2005) The aim of symmetry span is to measure visuospatial memory while performing a secondary visuospatial processing task. On each trial, the participant is shown a grid and asked to determine whether or not it is symmetrical. Next, they are shown a 4 × 4 grid of squares, one of which is red. Their goal is to memorize the location of the red square. This symmetry–square interleaving pattern continues two to five times (i.e., the set sizes used in the task). Afterward, the participant reports the location where the red squares appeared in order. We gave participants 12 trials: three of each set size. We used the partial scoring method (Conway et al., 2005) as the measure of performance; that is, rather than using "all-or-nothing" scoring, participants received credit for the number of red squares that they recalled in the correct serial position.

Rotation span (Kane et al., 2004; Unsworth et al., 2005) The aim of rotation span is to measure visuospatial memory



The college students went to New York in March and it snowed.

Andy was stopped by the policeman because he crossed the yellow heaven.

Fig. 1 Schematic of the four working memory capacity tasks

while performing a secondary visuospatial (rotation) processing task. On each trial, the participant is shown a letter they mentally rotate to determine its orientation (mirror-imaged or normal). Next, they are shown an arrow pointed in one of eight directions. This letter–arrow interleaving pattern continues two to five times. Afterward, the participant is asked to report the arrows in the order they appeared. We gave participants 12 trials: three of each set size. We used the partial scoring method (Conway et al., 2005) as the measure of performance; participants received credit for the number of arrows that they recalled in the correct serial position.

Reading span (Daneman & Carpenter, 1980) The aim of reading span is to measure verbal memory while performing a secondary verbal processing task. On each trial, the participant is shown a sentence and asked to determine whether it makes grammatical sense. Next, they are shown a single letter to be remembered. This letter–sentence interleaving pattern continues three to seven times. Afterward, the participant is asked to report the letters in the order they appeared. We gave participants six trials, and the set sizes of those trials were determined randomly due to a coding error. Thus, typical scoring approaches were not suitable for this administration of the task. We ameliorated this issue by computing average performance for each set size for each participant, and then using these as indicators of a "reading span" latent factor. The model fit the data well, $\chi^2(5) = 11.70$, $p = .039$, CFI = .986, TLI = .973, RMSEA = .029. Scores on the reading span factor were saved using the regression approach provided by the *umx* package (Bates et al., 2019) in R, using full information maximum likelihood estimation to handle missing data.

Letter–number sequencing (modified from Wechsler, 1997) The aim of the letter–number sequencing task is to measure the ability to maintain and manipulate information in memory. In each trial, participants view a continuous sequence of numbers and letters, 1,000 ms per item, in the center of the screen. Each trial includes three to seven items. At the end of each trial, participants are asked to type the numbers recalled in ascending order, followed by the letters recalled in alphabetical order. For example, if a participant observed the sequence "Q, 7, 2, H," the correct response would be "2, 7, H, Q." After three practice trials of three items, we gave participants 15 trials, three of each length. The measure of performance was the total number of correctly recalled items.

Updating

Figure 2 presents a schematic illustration of the four updating tasks (i.e., letter three-back, spatial three-back, keep track, and tone monitoring).

Letter three-back (modified from Kirchner, 1958) The aim of the letter three-back is to measure participants' ability to remember new, incoming verbal information and to discard outdated verbal information. In the letter *n*-back, subjects are shown a continuous sequence of letters and are asked to indicate whether the current letter matches the one that was presented three trials ago. For each letter, the subject used the "/" key to indicate that it was a target or the "z" key to indicate that it was a nontarget. Each letter appeared for 500 ms, then was masked for 2,000 ms during which the participant was to respond. Auditory feedback followed incorrect responses. Between each trial was a 250-ms blank display. There were 147 trials presented in a fixed order: 48 were target trials, 16 were two-back lure trials, 16 were four-back lure trials, 16 were five-back lure trials, and 51 were nonspecific nontarget trials. The measure of performance was the proportion of correct responses.

Spatial three-back (modified from Kirchner, 1958) The aim of the spatial three-back is to measure participants' ability to remember new, incoming spatial information and to discard outdated spatial information. In the spatial *n*-back, subjects are shown a continuous sequence of red squares that appear within a 4×4 grid and are asked to indicate whether the location of the current red square matches the one that was presented three trials ago. For each red square, the subject used the "/" key to indicate that it was a target or the "z" key to indicate that it was a nontarget. Each red square appeared for 500 ms, then was masked for 2,000 ms, during which the participant was to respond. Auditory feedback followed incorrect responses. Between each trial was a 250-ms blank display. There were 147 trials presented in a fixed order: 48 were target trials, 16 were two-back lure trials, 16 were four-back lure trials, 16 were five-back lure trials, and 51 were nonspecific nontarget trials. The measure of performance was the proportion of correct responses.

Tone monitoring (Miyake et al., 2000) The aim of the tone monitoring task is to measure participants' ability to keep track of incoming auditory information and to discard outdated auditory information. Participants hear a series

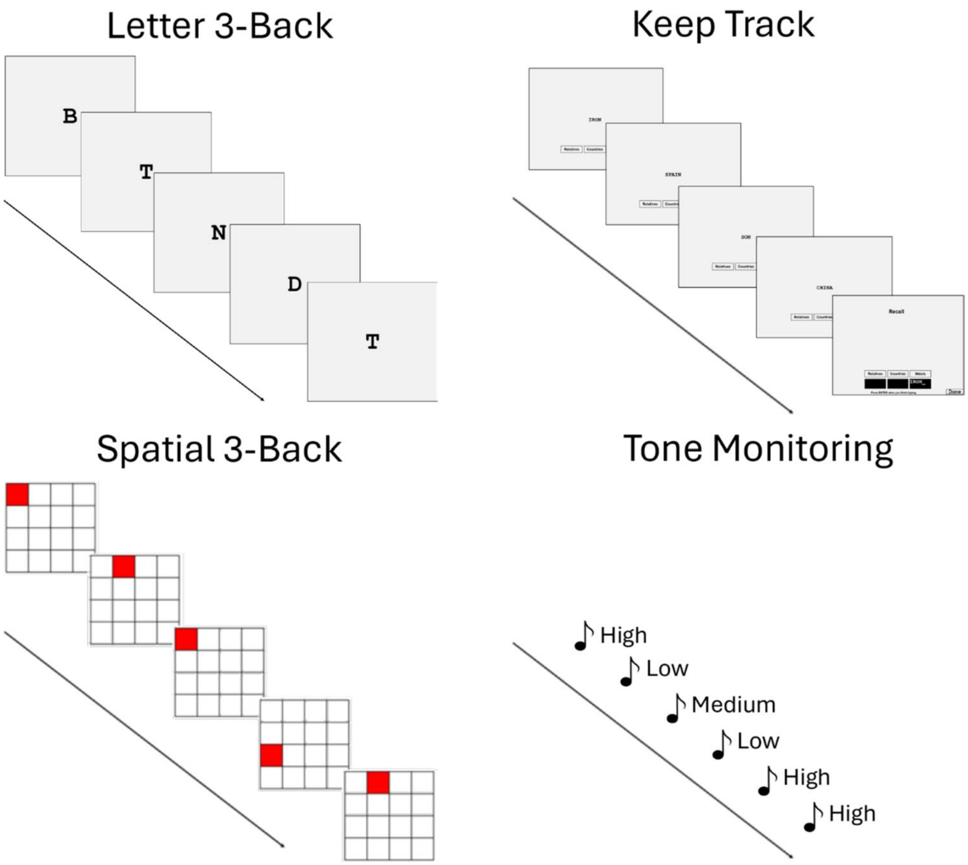


Fig. 2 Schematic of the four updating tasks

of high-pitched tones (880 Hz), medium-pitched tones (440 Hz), and low-pitched tones (220 Hz) for 500 ms each, with an interstimulus interval of 2,500 ms. Participants are tasked with pressing the space bar after hearing the fourth low tone, the fourth medium tone, and the fourth high tone in each block. Participants completed four blocks of 25 tones (eight high, eight medium, eight low, and one additional random tone) in a mixed order. If participants incorrectly pressed the space bar, the tone count reset for that pitch. Following practice trials, participants completed four blocks of trials. The measure of performance was the proportion of correct responses.

Keep track (adapted from Friedman et al., 2008) The aim of the keep track task is to measure participants' ability to monitor incoming words and to discard outdated words from memory. In each trial, participants viewed 15 words one at a time in a random order on the screen for 1,500 ms each that belonged to one of six categories (relatives, metals, animals, colors, countries, and distances). Target categories for each trial were presented at the bottom of the screen. Participants are tasked with typing the last exemplar of each target category at the end of the trial. For example, if the target categories were relatives, colors, and distances, the participant would attempt to remember the last exemplar of each of those categories and update this representation in memory when a new exemplar of each category was presented. After familiarization of the categories and a practice trial, participants completed three trials of four target categories and three trials of five target categories. Trial length (four vs. five) was randomized. The measure of performance was the proportion of words correctly recalled.

Fluid intelligence

Figure 3 presents a schematic illustration of the three fluid intelligence tasks (i.e., Raven's Advanced Progressive Matrices, letter sets, number series, and Cattell's test).

Raven's Advanced Progressive Matrices (Raven & Court, 1998) The aim of Raven's Advanced Progressive Matrices is to measure inductive reasoning in above-average intelligence samples. Participants are shown a 3×3 grid of patterns, with the pattern in the bottom right corner missing. The participant's task is to discern the rule governing the visuospatial set and then to select from eight response options the one that best completes the set. We gave participants 10 min to complete the 18 odd-numbered items, which increase in difficulty; the measure of performance was the proportion of correct responses.

Letter sets (Ekstrom et al., 1976) The aim of letter sets is to measure participants' ability to discern a rule governing sets of verbal stimuli. Participants are shown five sets of four letters and are challenged to identify which of the five sets of letters does not adhere to the same pattern as the others. We gave participants 7 min to complete 15 items; the measure of performance was the proportion of correct responses.

Number series (Thurstone, 1938) The aim of number series is to measure participants' ability to discern patterns of numerical stimuli. Participants are shown a set of numbers that follow a pattern and are challenged to identify which of four possible response options best completes the pattern. We gave participants 5 min for 15 items; the measure of performance was the proportion of correct responses.

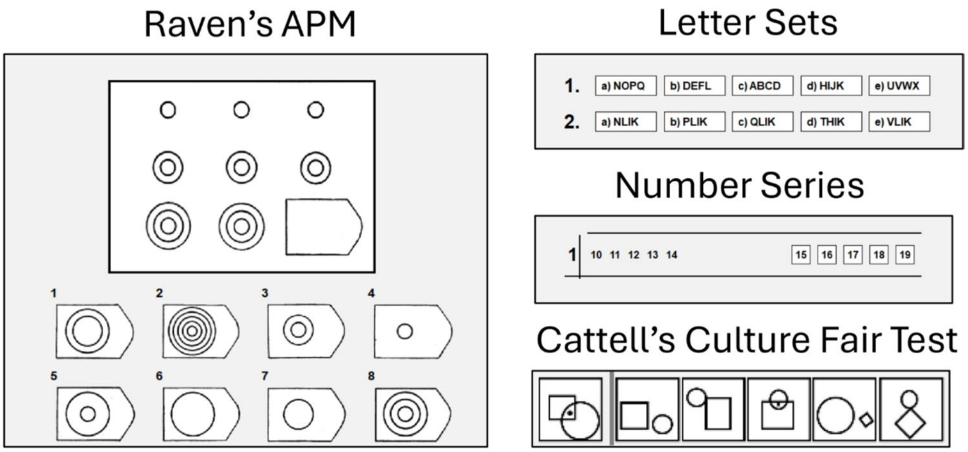

Raven's APM
1
2
3
4
5
6
7
8
Letter Sets
1. a) NOPQ b) DEFL c) ABCD d) HIJK e) UVWX
2. a) NLIK b) PLIK c) QLIK d) THIK e) VLIK
Number Series
1 10 11 12 13 14
15 16 17 18 19
Cattell's Culture Fair Test

Fig. 3 Schematic of the four fluid intelligence tasks

Cattell's Culture Fair Intelligence Test: Conditions subtest (Cattell, 1949) The aim of *Cattell's Culture Fair Test* is to measure inductive reasoning with visuo-spatial information. Participants are shown a box on the left side of the screen that contains a dot in it, as well as shapes and/or lines. The dot is in a particular location with respect to the shapes and/or lines (e.g., a dot inside a circle and a square). On the right side of the screen, subjects are shown five boxes that served as response options. These boxes also have shapes and/or lines in them, but no dot. Participant were asked to select the response option that would allow them to place a dot in an analogous location to the one that was provided in the box on the left. We gave participants 2.5 min for 10 items; the measure of performance was the proportion of correct responses.

Transparency and openness

This study was part of a larger project that was preregistered (https://osf.io/c39y6). This study was not preregistered, but we followed the same data preparation and cleaning procedures as outlined in the preregistration. Data, R code, and a codebook providing explanations of the files are available on the Open Science Framework (https://osf.io/q4pcw/?view_only=61e976d9d4a64b2ebef91cfa848c74ce).

Data preparation

Our initial dataset included 1,407 participants. We removed participants' data if they were noncompliant during the study (e.g., using their phone while completing a task; $n_{\text{subjects}} = 123$). Next, we searched for outlying scores, which were defined as scores that exceeded three standard deviations from the sample mean for that measure (i.e., z scores $> \pm 3$). Scores that were three standard deviations better than the mean were Winsorized to $z = 3$ ($n_{\text{obs.}} = 0$). Scores that were three standard deviations worse than the mean were removed ($n_{\text{obs.}} = 190$). Subjects who were missing three or more data points due to outlier exclusion were removed ($n_{\text{subjects}} = 4$). Finally, we computed Mahalanobis' distance to detect multivariate outliers; we removed subjects' data if the associated p value was $< .001$ ($n_{\text{subjects}} = 8$). We used the z-scored variables for all analyses in the manuscript. As a sensitivity check, we also analyzed scores prior to data cleaning and found the same pattern of significant results. The analyses reported below are based on the analytical sample of 1,272 participants who were retained throughout the data cleaning procedure.

Modeling approach and fit statistics

For all confirmatory factor analyses and structural equation models, we used maximum likelihood estimation with robust standard errors and full information maximum likelihood estimation for missing data. Variables were standardized prior to estimation. Confirmatory factor analyses and structural equation models were estimated using JASP (JASP Team, 2024).

We report multiple fit statistics: The χ^2 is an absolute fit index comparing the fit of the specified model to that of the observed covariance matrix. A significant χ^2 can indicate lack of fit but is heavily influenced by sample size. In large samples, such as the one used in the present study, even a slight deviation between the data and the model can lead to a significant χ^2 statistic. Therefore, we also report the comparative fit index (CFI) and Tucker–Lewis index (TLI), which compare the fit of the model to a null model in which the covariation between measures is set to zero, while adding penalties for additional parameters. For CFI and TLI, large values indicate better fit (i.e., $> .90$ or ideally, $> .95$). For the root mean square error of approximation (RMSEA) fit statistic, values less than .05 are considered great, while values less than .08 are considered adequate. For the standardized root mean square residual (SRMR), which computes the standardized difference between the observed and predicted correlations, a value of less than .08 indicates adequate fit, with lower values indicating better fit (Hu & Bentler, 1999).

Results

Descriptive statistics are presented in Table 1. Correlations are presented in Table 2. The three complex span tasks had intercorrelations ranging from $r = .33$ to $r = .57$ ($\bar{r} = .43$), whereas the two n-back tasks correlated $r = .71$ with each another.

Observed correlations between complex span and n-back performance ranged from $r = .17$ to $r = .32$ ($\bar{r} = .25$). Of the three complex span tasks, reading span had the weakest correlation with the two n-back tasks (rs of .17 and .18). Reading span uses verbal memory items; as such, these results align with Redick and Lindsey's (2013) meta-analytic finding of relatively weak associations between n-back tasks and complex span that used verbal memoranda. Reading span also had lower reliability than the other two complex span tests, which could have attenuated its correlations. Given that construct- and method-specific variance influence these relationships, as well as psychometric issues such as unreliability, our next analyses used latent variable techniques to shed more light on our research questions.

We first conducted an exploratory factor analysis on the complex span and n-back measures, using principal axis factoring and an oblique (promax) rotation. We extracted two factors with eigenvalues greater than 1 (Fig. 4). The two n-back tasks loaded highly on the first factor (loadings of .84 and .84) and the three complex span tasks loaded highly

Table 1 Descriptive statistics

Measure	*N*	Mean	*SD*	Skew	Kurtosis	Reliability
Symmetry span	1,232	28.67	8.01	−0.61	−0.18	α = .74
Rotation span	1,233	17.70	5.44	−0.35	−0.32	α = .76
Reading span	1,250	0.05	0.70	−0.52	−0.27	ω = .60
Letter–number sequencing	1,207	55.93	10.44	−0.61	0.19	α = .80
Letter *n*-back	1,221	0.61	0.16	−0.46	0.58	α = .95
Spatial *n*-back	1,215	0.53	0.16	−0.54	0.44	α = .95
Tone monitoring	1,223	0.82	0.10	−0.71	−0.08	α = .92
Keep track	1,248	0.45	0.11	−0.61	0.27	α = .59
Raven's matrices	1,221	0.55	0.19	−0.40	−0.09	α = .76
Letter sets	1,225	0.69	0.18	−0.76	0.20	α = .72
Number series	1,234	0.63	0.20	−0.29	−0.44	α = .74
Cattell's test	1,254	0.52	0.17	−0.03	−0.29	α = .50

All variables were standardized prior to data cleaning and analysis (see Data Preparation section); variables were rescaled to their original units to report descriptive statistics. α = Cronbach's alpha; ω = coefficient omega; Listwise *n* = 1,127

Table 2 Correlation matrix

Measure	1	2	3	4	5	6	7	8	9	10	11
1. Symmetry span	–										
2. Rotation span	.57	–									
3. Reading span	.39	.33	–								
4. Letter–number sequencing	.34	.34	.35	–							
5. Letter *n*-back	.32	.30	.17	.34	–						
6. Spatial *n*-back	.27	.27	.18	.33	.71	–					
7. Tone monitoring	.38	.39	.26	.41	.55	.47	–				
8. Keep track	.20	.18	.18	.26	.25	.25	.28	–			
9. Raven's matrices	.40	.37	.25	.39	.48	.45	.54	.29	–		
10. Letter sets	.31	.31	.23	.45	.46	.40	.51	.28	.54	–	
11. Number series	.39	.36	.29	.44	.42	.36	.47	.27	.47	.47	–
12. Cattell's test	.20	.21	.13	.25	.32	.34	.37	.16	.38	.33	.29

Pairwise *n* ranges from 1,181 to 1,248. All correlations are significant at $p < .001$

on the second factor (loadings of .82, .69, and .47 [reading span]). Cross-loadings were negligible (range: – .02 to .03). The factors correlated $r = .44$.

We repeated this analysis after including the additional measures of working memory capacity (letter–number sequencing) and updating (tone monitoring and keep track). Two factors had eigenvalues > 1. Tone monitoring primarily loaded on the first *n*-back/updating factor (loadings of .49 vs .27). Letter–number sequencing primarily loaded on the second complex span/working memory capacity factor (loadings of .21 vs .41). Keep track had similar loadings on both factors (loadings of .22 vs .18). In the following analyses, keep track served as an indicator of updating because it requires rapid disengagement, but we note that the conclusions of the manuscript are robust to its exclusion.

We investigated the relation between complex span and *n*-back performance at the latent level by comparing the fit of two confirmatory factor analysis models. In the first, the complex span measures loaded on one factor and the *n*-back measures loaded on another factor. As shown in Fig. 5, the correlation between the two factors was $r = .45$ [.39, .51], indicating that they shared 20.25% of their variance. In the second model, we tested whether this correlation could be constrained to 1 without loss in model fit. Constraining the correlation significantly worsened model fit, $\Delta\chi^2(1) = 514.61$, $p < .001$, indicating that complex span and *n*-back measures differ at the latent level.

We repeated this analysis after including the broader working memory capacity (letter–number sequencing) and updating measures (tone monitoring and keep track). We found that the two factors correlated $r = .57$ [.51, .62]

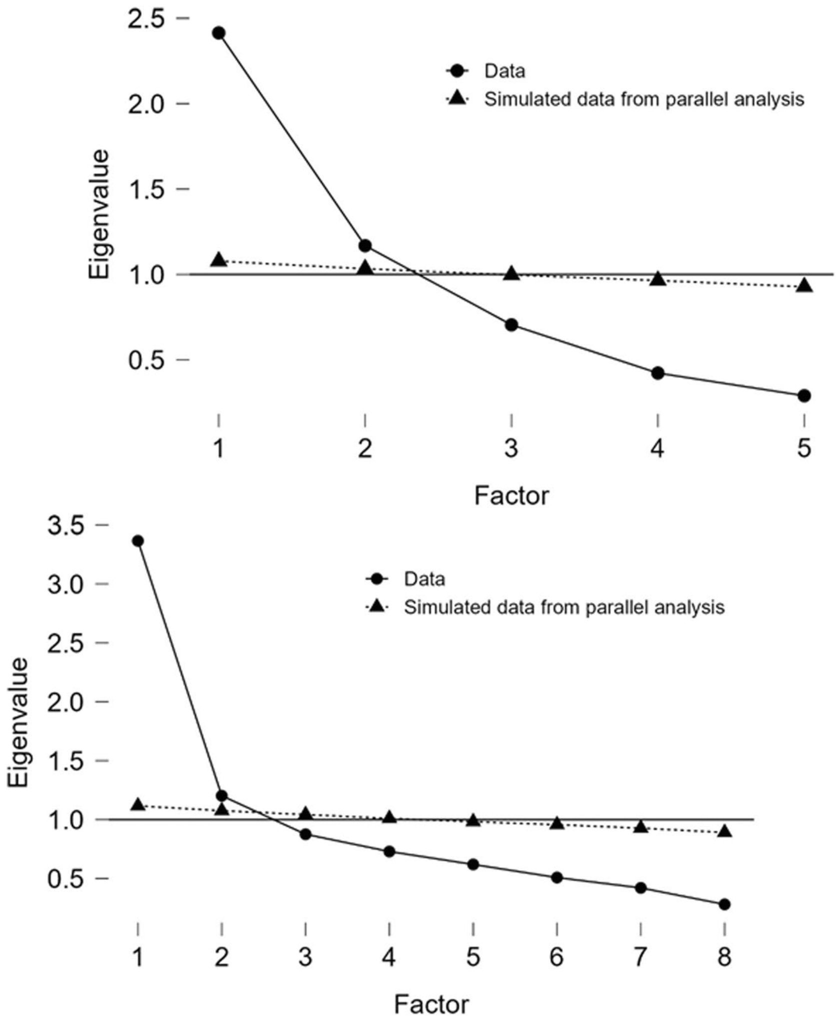


Fig. 4 Scree plots from exploratory factor analyses. *Note.* Top panel: Scree plot from exploratory factor analysis with complex span and *n*-back measures. The model fit the data well, $\chi^2(1)=2.54$, $p=.111$; CFI = .999, TLI = .991, RMSEA = .035, 90% CI [.000, .091], SRMR = .005. Bottom panel: Scree plot from exploratory factor analysis including additional measures of working memory capacity (letter–number sequencing) and updating (tone monitoring and keep track). Model fit was adequate, $\chi^2(13)=89.88$, $p<.001$; CFI = .974, TLI = .944, RMSEA = .068, 90% CI [.055, .082], SRMR = .027

($R^2=32.5\%$) and could not be set equal to each another without loss in fit, $\Delta\chi^2(1)=412.52$, $p<.001$.

Next, we tested whether the complex span or the *n*-back factor was more strongly related to fluid intelligence. The correlation between complex span and fluid intelligence was $r=.66$ [.61, .72], whereas the correlation between *n*-back and fluid intelligence was $r=.74$ [.70, .78] (Fig. 6). Constraining these correlations to be equal significantly worsened model fit, $\Delta\chi^2(1)=5.22$, $p=.022$, indicating that fluid intelligence was significantly more strongly related to the *n*-back factor than the complex span factor. Furthermore, both the *n*-back and complex span factors were significantly more strongly related to fluid intelligence than to each other, $\Delta\chi^2(1)=92.52$, $p<.001$, and $\Delta\chi^2(1)=51.60$, $p<.001$, respectively.

We repeated this analysis for the broader working memory capacity and updating factors and found the same pattern of results. Fluid intelligence correlated significantly more strongly with updating ($r=.84$ [.81, .87]) than with working memory capacity, $r=.76$ [.71, .81], $\Delta\chi^2(1)=7.36$, $p=.007$.

Finally, we tested whether the complex span and *n*-back factors accounted for *unique* variance in fluid intelligence above and beyond one another. Using structural equation modeling, we specified the complex span and *n*-back factors as correlated predictors of fluid intelligence (Fig. 7). Both regression paths were statistically significant (complex span $\beta=.42$; *n*-back $\beta=.55$), indicating that both factors explained unique variance. The *n*-back factor explained significantly more unique variance in fluid intelligence than the

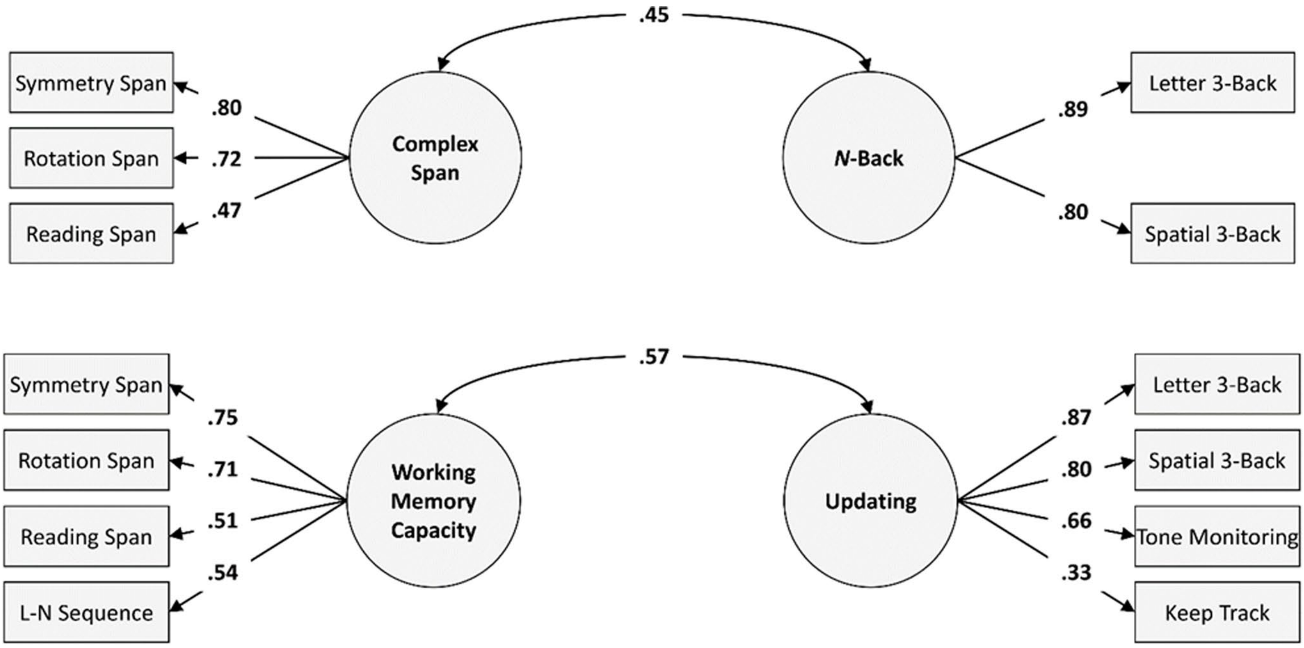


Fig. 5 Confirmatory factor analyses of complex span and *n*-back and of working memory capacity and updating. *Note.* Top panel: Confirmatory factor analysis with one factor for complex span measures and another factor for *n*-back measures. Model fit was excellent: $\chi^2(4)=1.93$, $p=.748$; CFI = 1.00, TLI = 1.00, RMSEA = .000, 90% CI [.000, .030], SRMR = .006. Bottom panel: Confirmatory factor analysis including additional measures of working memory capacity and updating. Model fit was less than adequate: $\chi^2(19)=248.09$, $p<.001$; CFI = .919, TLI = .880, RMSEA = .097, 90% CI [.087, .108], SRMR = .064; there was a large residual variance for keep track (.89), but it was retained on conceptual grounds as a measure of updating. $N=1{,}272$

complex span factor did, $\Delta\chi^2(1)=5.22$, $p=.022$. Together, the predictors explained 67.9% of the variance.

We repeated this analysis for the broader working memory capacity and updating factors and found the same pattern of results. Updating explained significantly more unique variance in fluid intelligence (β = .60) than working memory capacity did (β = .40), $\Delta\chi^2(1)=7.36$, $p=.007$. Together, the predictors explained 80.2% of the variance in fluid intelligence.

Discussion

In this multisite study ($N=1{,}272$), we found strong evidence for a dissociation between complex span and *n*-back tasks, and more generally, between working memory capacity and updating measures. On this basis, we suggest that researchers who want to measure working memory capacity should not use complex span and *n*-back tasks interchangeably.

What are the sources of variance that distinguish these measures? Kane et al. (2007) identified *recall versus recognition* as one candidate. In the *n*-back, subjects make continuous recognition judgments, whereas in complex span tasks, subjects attempt to recall memory items at the end of each trial. Thus, performance on the *n*-back can be supported by recognition or a familiarity function rather than explicit recall (Oberauer, 2009; Szmalec et al., 2011), and these abilities may reflect different aspects of the working memory system (Kane et al., 2007). Although recognition could support performance in some *n*-back tasks, our *n*-back tasks had as many lure trials as target trials; relying on mere familiarity would not be an effective strategy as it would result in a high false alarm rate.

The evidence presented here suggests other distinguishing features. In structural equation models, complex span and *n*-back performance each explained unique variance in fluid intelligence above and beyond the other. This same pattern was obtained when analyzing broader factors representing working memory capacity and updating. The key distinguishing features of these measures and constructs may be the role of maintenance relative to disengagement.

Attention control supports both maintenance and disengagement (Burgoyne & Engle, 2020; Shipstead et al., 2016). Working memory capacity measures (e.g., complex span tasks) emphasize maintenance; the role of attention control is to protect memory items from the consequences of interference from the secondary task. This attention control process is likely what explains the shared variance between working memory capacity and fluid intelligence. When problem-solving, one's ability to generate, test, and keep track of hypotheses depends on the ability to maintain this information despite interference.

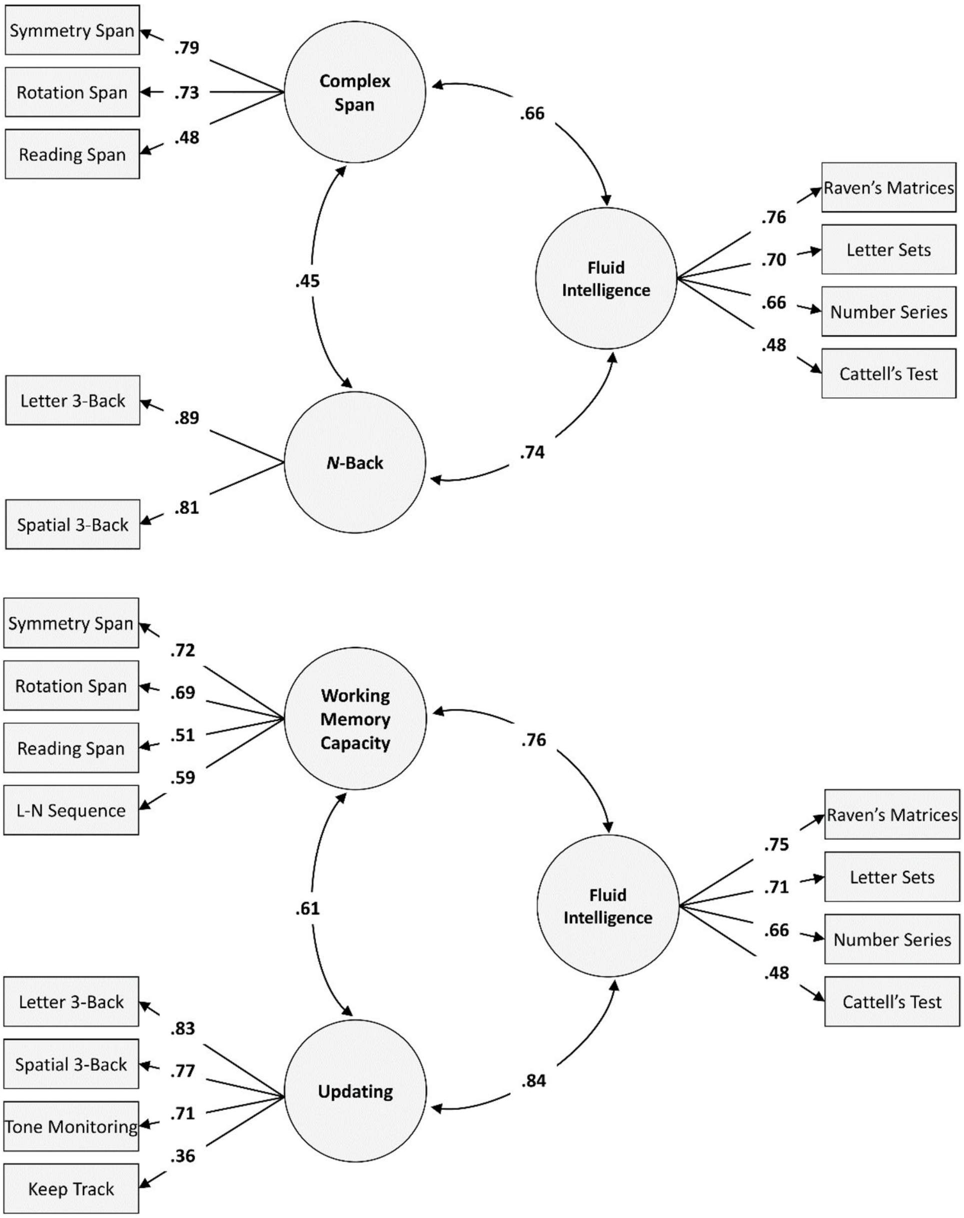


Fig. 6 Confirmatory factor analyses examining relationships with fluid intelligence. *Note*. Top panel: Confirmatory factor analysis examining the relation between fluid intelligence, complex span, and *n*-back performance at the latent level. Model fit was excellent: $\chi^2(24) = 62.07$, $p < .001$; CFI = .989, TLI = .984, RMSEA = .035, 90% CI [.025, .046], SRMR = .022. Bottom panel: Confirmatory factor analysis including additional measures of working memory capacity and updating. Model fit was adequate: $\chi^2(51) = 450.56$, $p < .001$; CFI = .919, TLI = .896, RMSEA = .078, 90% CI [.072, .085], SRMR = .053. $N = 1{,}272$

In contrast, updating measures (e.g., *n*-back) emphasize disengagement; the role of attention control is to remove outdated items from focus (Shipstead et al., 2016). This attention control process is likely what explains the covariance between updating and fluid intelligence. When problem-solving, one's ability to discard hypotheses discovered to be false depends on one's ability to disengage. Our structural equation models demonstrated that updating correlated with fluid intelligence more strongly than working memory capacity, suggesting that disengagement may be especially crucial for problem-solving.

Nevertheless, there are alternative theoretical accounts that are also consistent with our results. For example, the "binding hypothesis" (e.g., Wilhelm et al., 2013) views the working memory system as reflecting the ability to rapidly build, maintain, and update bindings or relations between memory items. Within this framework, maintenance of active bindings and disengagement from

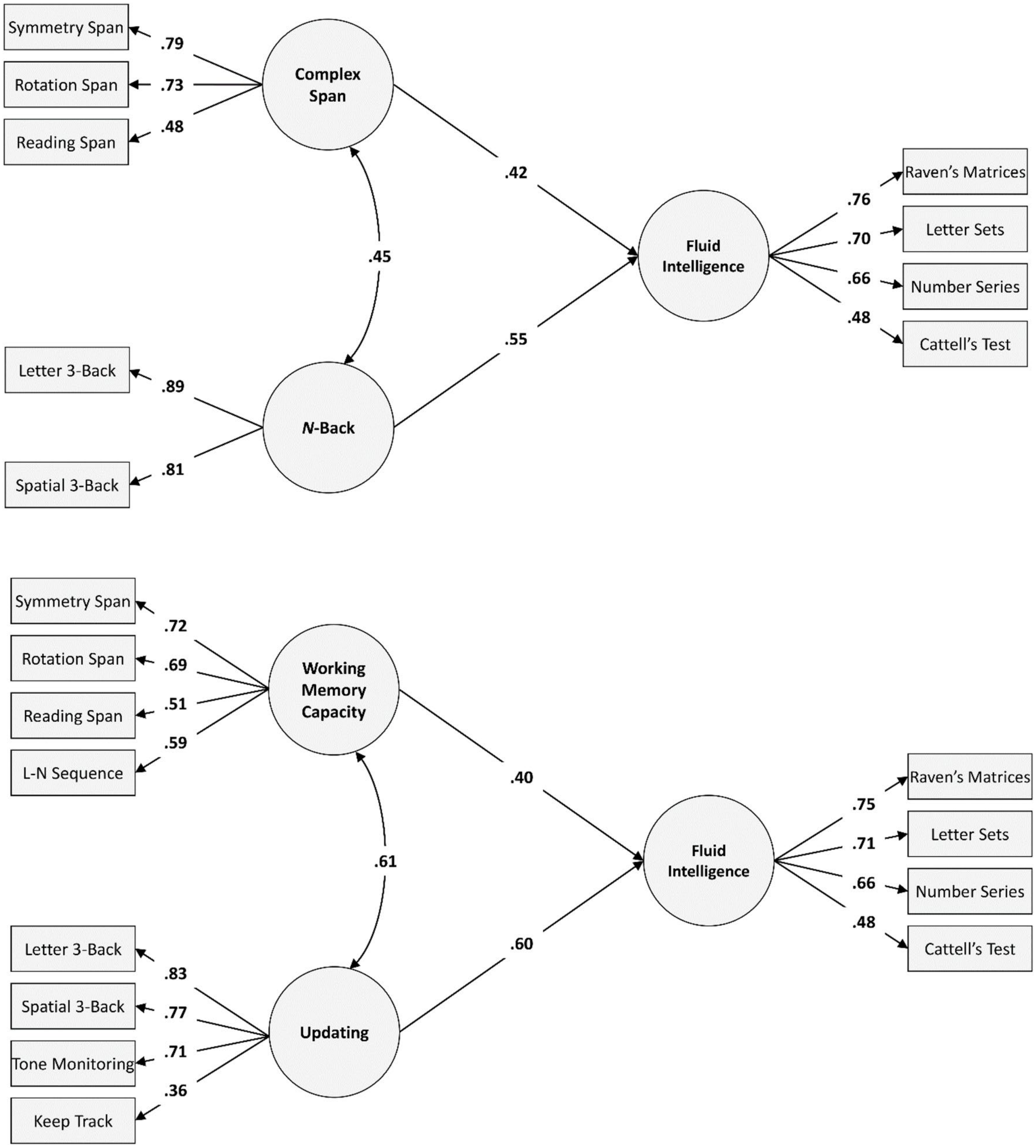


Fig. 7 Structural equation models predicting fluid intelligence. *Note*. Top panel: Structural equation model with complex span and *n*-back factors specified as correlated predictors of fluid intelligence. Model fit was excellent: $\chi^2(24) = 62.07$, $p < .001$; CFI = .989, TLI = .984, RMSEA = .035, 90% CI [.025, .046], SRMR = .022. Bottom panel: Structural equation model including additional measures of working memory capacity and updating. Model fit was adequate: $\chi^2(51) = 450.56$, $p < .001$; CFI = .919, TLI = .896, RMSEA = .078, 90% CI [.072, .085], SRMR = .053. $N = 1{,}272$

no-longer-relevant bindings can explain the dissociation between categories of working memory tasks such as complex span and the *n*-back. That is, the general principles of maintenance and disengagement can be incorporated into other working memory frameworks, including the binding hypothesis as well as Cowan's embedded-process model (Cowan et al., 2020). The goal of our study was not to adjudicate between these different frameworks but to examine relationships among cognitive constructs and discuss the cognitive process of disengagement as a potential explanation for the observed relationships.

Our results from 1,272 university and community participants revealed that updating was strongly related to fluid intelligence ($r = .84$) and, to a lesser extent, working memory capacity ($r = .61$). These findings stand in contrast to the results of Frischkorn et al. (2022), who analyzed a sample of 111 young adults and concluded that updating-specific variance was not related to fluid intelligence or working memory capacity. How can this discrepancy be reconciled? Several issues with the measures and models from the study preclude strong conclusions. For example, the updating-specific factor in Frischkorn et al.'s study did not effectively capture updating-specific variance. The measures of updating and "nonupdating" consisted of three versions of a short "keep track" task (15 trials requiring updating and five trials that did not). Across several models, the loadings for the updating-specific indicators were not credibly different from zero, and the best-fitting models did not include updating-specific variance. Further, the "nonupdating" indicators were only weakly correlated with each other (*r*s of .16, .20, and .27), two suggested ceiling effects (accuracy of 95% and 91%), and these indicators had imbalanced and generally low loadings (e.g., .6, .3, .3) on a common factor. Oddly, the "updating" and "nonupdating" latent factors were highly correlated with each other (e.g., a standardized path of .88 in one model in which it was freely estimated). Although Frischkorn et al. claimed null relationships between updating and other constructs of interest, they also stated that "strictly speaking, these results preclude any further investigation of relationships of updating-specific variance with the other covariates" (p. 1349). In comparison, our large sample with a range of abilities, robust factor loadings, and model fit provide compelling evidence for a strong relationship between updating and fluid intelligence.

Limitations

The first limitation of this work is that we administered only two *n*-back tests, one with spatial memory items and another with verbal memory items. As a result, our *n*-back latent factor may contain more method-specific variance than if we had administered three *n*-back tests. This study was part of a larger investigation, and it was not possible to include additional *n*-back tests in the larger investigation's design. To mitigate this issue, we also modeled a broader "updating" latent factor that incorporated two additional tasks: tone monitoring and keep track. The influence of common method variance was likely reduced in this factor, and yet models using this factor revealed stronger evidence for a dissociation than the models that included just complex span and *n*-back tasks. Therefore, we would predict that adding a third *n*-back task might strengthen the results presented here but would not be likely to change the overall conclusions. A similar argument could be made regarding the operationalization of the updating factor; additional measures (for an example, see Ecker et al., 2014) could be included to broaden the factor.

The second limitation of this work is the relatively lower internal consistency reliability for the complex span tasks (.74, .76, and .70) than the *n*-back tasks (.95 and .95). Although the reliability estimates for the complex span tasks were not low by conventional standards (Parsons et al., 2019), unreliability reduces the shared variance among measures and could have consequences for the latent factor attempting to capture this common variance. Reliability could be increased by adding more trials to the complex span tasks. However, we do not think our results are merely due to differing reliabilities of the indicator measures. When we added a more-reliable measure (letter–number sequencing: .80) to the working memory capacity factor and slightly less-reliable measures (keep track and tone monitoring: .92 and .59) to the updating factor, making the reliabilities of the indicator measures more balanced across factors, the dissociation between the factors increased.

Conclusion

Taking a step back, the present results raise an important question for our conceptualization of what complex span tasks measure, as well as the structure of executive functions. Though in Friedman and Miyake's (2017) model of inhibition, complex span and updating tasks are described as measuring closely related constructs, we provide strong evidence that these two types of measures are not equivalent. The complex span and *n*-back tests were only moderately related to one another at the latent level ($r = .45$), demonstrating poor convergent validity for two sets of tasks that ostensibly measure the same construct.

These tasks may differ not only in their affordance of recall versus recognition, but also in their relative demands on maintenance versus disengagement. We suggest that researchers think carefully about the ability they wish to measure before choosing tasks to administer. Accurate interpretations of measures are necessary for understanding the latent structure of executive functions and developing robust theories of cognitive constructs.

Acknowledgements The authors are grateful to David Z. Hambrick for supervision and assistance collecting data at Michigan State University.

Author contributions All authors were involved in the conceptualization, drafting, and revising of the manuscript.

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 # W911NF-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 Data are available on the Open Science Framework (https://osf.io/q4pcw/?view_only=61e976d9d4a64b2ebef91cfa848c74ce).

Code availability R code is available on the Open Science Framework (https://osf.io/q4pcw/?view_only=61e976d9d4a64b2ebef91cfa848c74ce).

Declarations

Conflicts of interest The authors declare no conflicts of interest.

Ethics approval This study was approved by each university's Institutional Review Board.

Consent to participate All participants provided informed consent.

Consent for publication Consent for publication was obtained from all relevant parties.

Open practices statement This study was part of a larger project that was pre-registered (https://osf.io/c39y6). This study was not preregistered, but we followed the same data preparation and cleaning procedures as outlined in the preregistration. Data, R code, and a codebook providing explanations of the files are available on the Open Science Framework (https://osf.io/q4pcw/?view_only=61e976d9d4a64b2ebef91cfa848c74ce).

References

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8*, 47–89. https://doi.org/10.1016/S0079-7421(08)60452-1

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

Bates, T. C., Neale, M. C., & Maes, H. H. (2019). umx: A library for structural equation and twin modelling in R. *Twin Research and Human Genetics, 22*, 27–41.

Burgoyne, A. P., & Engle, R. W. (2020). Attention control: A cornerstone of higher-order cognition. *Current Directions in Psychological Science, 29*(6), 624–630. https://doi.org/10.1177/0963721420969371

Burgoyne, A. P., Hambrick, D. Z., & Altmann, E. M. (2019). Placekeeping ability as a component of fluid intelligence: Not just working memory capacity. *The American Journal of Psychology, 132*(4), 439–449. https://doi.org/10.5406/amerjpsyc.132.4.0439

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. https://psycnet.apa.org/doi/https://doi.org/10.1037/0033-295X.97.3.404

Cattell, R. B. (1949). *Culture fair intelligence test, Scale 1, handbook.* Institute of Personality and Ability Testing.

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*, 769–786. https://doi.org/10.3758/BF03196772

Cowan, N., Morey, C. C., & Naveh-Benjamin, M. (2020). An embedded-processes approach to working memory: How is it distinct from other approaches, and to what ends? In R. H. Logie, V. Camos, & N. Cowan (Eds.), *Working memory: State of the science* (pp. 44–84). Oxford University Press. https://doi.org/10.1093/oso/9780198842286.003.0003

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*(4), 422–433. https://doi.org/10.3758/BF03214546

Ecker, U. K., Lewandowsky, S., & Oberauer, K. (2014). Removal of information from working memory: A specific updating process. *Journal of Memory and Language, 74*, 77–90.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests: 1976*. Educational Testing Service.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19–23.

Engle, R. W. (2018). Working memory and executive attention: A revisit. *Perspectives on Psychological Science, 13*(2), 190–193.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309–331. https://psycnet.apa.org/doi/https://doi.org/10.1037/0096-3445.128.3.309

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex, 86*, 186–204. https://doi.org/10.1016/j.cortex.2016.04.023

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corely, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General, 137*(2), 201–225. https://doi.org/10.1037/0096-3445.137.2.201

Frischkorn, G. T., Von Bastian, C. C., Souza, A. S., & Oberauer, K. (2022). Individual differences in updating are not related to reasoning ability and working memory capacity. *Journal of Experimental Psychology: General, 151*(6), 1341–1357. https://doi.org/10.1037/xge0001141

Hambrick, D. Z., & Altmann, E. M. (2015). The role of placekeeping ability in fluid intelligence. *Psychonomic Bulletin & Review, 22*, 1104–1110. https://doi.org/10.3758/s13423-014-0764-5

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

JASP Team. (2024). *JASP* (Version 0.18.3) [Computer software]. https://jasp-stats.org

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the *N*-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615–622. https://psycnet.apa.org/doi/https://doi.org/10.1037/0278-7393.33.3.615

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217. https://psycnet.apa.org/doi/https://doi.org/10.1037/0096-3445.133.2.189

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology, 55*(4), 352–358. https://doi.org/10.1037/h0043688

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

Mackworth, J. F. (1959). Paced memorizing in a continuous task. *Journal of Experimental Psychology, 58*(3), 206–211. https://psycnet.apa.org/doi/https://doi.org/10.1037/h0049090

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621–640. https://psycnet.apa.org/doi/https://doi.org/10.1037/0096-3445.130.4.621

Oberauer, K. (2009). Design for a working memory. *Psychology of Learning and Motivation, 51*, 45–100. https://doi.org/10.1016/S0079-7421(09)51002-X

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

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science, 2*(4), 378–395.

Raven, J. C., & Court, J. H. (1998). *Raven's progressive matrices and vocabulary scales* (Vol. 759). Oxford Psychologists Press.

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. https://doi.org/10.3758/s13423-013-0453-9

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 1089–1096. https://doi.org/10.1037/a0015730

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. https://doi.org/10.1177/1745691616650647

Szmalec, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011). Control of interference during working memory updating. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 137–151. https://psycnet.apa.org/doi/https://doi.org/10.1037/a0020365

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 1,* ix + 121.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498–505. https://doi.org/10.3758/BF03192720

Wager, T. D., & Smith, E. E. (2003). Neuroimaging studies of working memory. *Cognitive, Affective, & Behavioral Neuroscience, 3*(4), 255–274. https://doi.org/10.3758/CABN.3.4.255

Wechsler, D. (1997). *Wechsler adult intelligence scale-III (WAIS-III) manual.* The Psychological Corporation.

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, Article 433.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.