Citation for the published version:

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*, 1341–1357. https://doi.org/10.1037/xge0001141

Individual Differences in Updating are not related to Reasoning Ability and Working Memory Capacity

Gidon T. Frischkorn[1], Claudia C. von Bastian[2], Alessandra S. Souza[1,3] & Klaus Oberauer[1]

[1]University of Zurich, Switzerland

[2]University of Sheffield, United Kingdom

[3]University of Porto, Portugal

Word Count:

Abstract: 181; Manuscript: 9962

No. of Figures & Tables: 7 & 3

Author Note:

Correspondence regarding this article should be addressed to Gidon T. Frischkorn, University of Zurich, Department of Psychology, Binzmühlestrasse 14/22, CH-8004 Zurich, Switzerland, Mail: gidon.frischkorn@psychologie.uzh.ch, Phone: +41 44 635 74 54; Scripts for data preparation and all analyses can be found at: osf.io/zkd4c.

Abstract

Previous research assumes that executive functions such as inhibition, shifting and updating explain individual differences in cognitive abilities. Of these three executive functions, updating was previously found to relate most strongly to fluid intelligence. However, this relationship could be a methodological artifact: Measures of inhibition and shifting usually isolate the contribution of this executive function to performance by contrasting conditions with high and low demands on these processes, whereas updating is measured by overall accuracy in working memory tasks involving updating. This updating measure conflates updating-specific individual differences (e.g., removal of outdated information) with variance in working memory maintenance. Re-analyzing data (N = 111) from von Bastian et al. (2016), we separated updating-specific variance from working memory maintenance variance. Updating contributed only 15% to individual differences in performance in the updating tasks, and it correlated neither with fluid intelligence nor with independent working memory measures reflecting storage and processing or relational integration. In contrast, the working memory maintenance component of the updating task correlated with both abilities. These findings challenge the view that updating contributes to variance in higher cognitive abilities.

Keywords: Updating; Executive Functions; Working Memory; Reasoning.

Individual differences in updating are not related to reasoning ability and working memory capacity

Executive functions (EF) are often defined as supervisory mechanisms that control information processing during goal-directed cognition (Miller & Cohen, 2001; Miyake et al., 2000)[1]. Factor-analytic research on individual differences has yielded the distinction of three EFs: inhibition, shifting, and updating (Karr et al., 2018; Miyake et al., 2000). Inhibition refers to focusing attention on relevant information while suppressing information that is irrelevant for the current task. Shifting refers to flexibly switching between different tasks. Updating refers to replacing outdated information in working memory (WM) by new, more relevant information. EFs have been shown to be related to a broad range of behaviors, such as clinical disorders (Snyder et al., 2015), eating behavior (Allom & Mullan, 2014), multi-tasking (Himi et al., 2019), or memory (Hedden & Yoon, 2006). Critical to the present study, EFs have been argued to play a central role in explaining individual differences in complex cognition (Barbey et al., 2012; Engle, 2002; Kovacs & Conway, 2016). In particular, some theorists (Conway et al., 2002; Shipstead et al., 2016) have proposed that EFs underlie the strong relationship between WM capacity, that is, the ability to retain access to a limited amount of information needed for complex cognition in the

[1] Some researcher use the term *executive functions* broadly, subsuming any goal-directed cognition, including fluid intelligence and working memory (Diamond, 2013). This conceptualization is not suitable if we are interested in identifying the cognitive processes underlying individual differences in fluid intelligence or working memory capacity, because we would explain a broad construct such as *fluid intelligence* by itself under another name. In this definition, the term *executive functions* is often used interchangeably with other denominators such as attentional control, executive attention, executive control, or cognitive control.

present moment (e.g., Oberauer, 2009), and fluid intelligence (Gf), that is, the ability to reason with novel information (e.g., Cattell, 1963).

Of the three EFs, updating ability has been shown to be most strongly related to Gf (Friedman et al., 2006; Wongupparaj et al., 2015). However, as we will lay out in detail below, previous studies have conflated two factors contributing to updating performance: executive processes specific to updating (i.e. the substitution of outdated information) and memory maintenance. The goal of the present study was to disentangle these two factors and investigate the extent to which they explain the relationship between updating on the one hand, and Gf and WM capacity (WMC) on the other. To foreshadow our results, we found strong evidence that maintenance, not executive control, underlies the relationship between updating and complex cognition.

How Is Updating Related to fluid intelligence and working memory capacity?

Whereas it is well-established that WMC and Gf are strongly related (e.g. Kyllonen & Christal, 1990; Süß et al., 2002), it is still a matter of ongoing theoretical debate as to *why* they are related. Some researchers argue that WMC and Gf are related because they both rely on executive control ability (Conway et al., 2002; Shipstead et al., 2016). More specifically, Shipstead et al. (2016) proposed that executive control is deployed through two different mechanisms that contribute to performance in both WM and Gf tasks to different degrees: maintenance and disengagement. This view builds on the conceptualization of WMC as executive control ability, where the maintenance of information in WM requires focusing attention on the to be remembered information and, additionally, disengaging from potentially distracting information (Engle, 2002; Kane & Engle, 2002). However, according to Shipstead et al. (2016), traditional WMC measures, such as complex span tasks, tap mainly maintenance and rely on disengagement only when it comes to avoiding distraction from secondary task demands. In contrast, Shipstead and colleagues argue, solving reasoning problems, as used to measure Gf, involves mainly disengaging from no longer relevant information (e.g., incorrectly deducted, or induced rules) and, only to a lesser degree, focusing and maintaining relevant information.

The relationship between updating ability and Gf constitutes a special case because, different to complex span tasks, updating tasks capture both mechanisms more equally: they require maintaining information while also disengaging from outdated information in WM (Ecker et al., 2010). Therefore, according to Shipstead et al.'s (2016) theoretical perspective, there should be a strong correlation between updating ability and measures of Gf and of WMC, including WMC tasks without updating demand.

In contrast, other researchers conceptualize WM more generically as an ensemble of cognitive components that holds information temporarily active for ongoing information processing (Cowan, 2017). The generic WM definition separates cognitive processes (or components) that are responsible for WM maintenance from executive-control processes in general, and from updating in particular. Therefore, from the perspective of theories building on the generic definition of WM (Cowan et al., 2005; Martínez et al., 2011; Oberauer, 2009), there is no reason to expect a close relationship between WMC and updating ability. In this view, the strong correlation between WMC and Gf does not reflect shared variance of executive control, so no relation between updating and Gf is predicted either.

Measuring Updating-Specific Processes

To adequately address the different conceptualizations of WMC and their predictions about relationships of individual differences in updating task with Gf or other WMC measures, the individual differences related to WM maintenance need to be separated from EF demands specific to updating. *WM maintenance* refers to the ability to hold several distinct items – sometimes referred to as chunks – available for processing over a few seconds. It is the main limiting factor of performance in WM tasks, as shown by the fact that when maintenance demands are reduced, memory performance is nearly perfect: Everyone can remember 1 or 2 items, but memory performance decreases when memory load surpasses 4-5 items (Cowan, 2001; Luck & Vogel, 1997). Therefore, the main source of variance shared by WMC measures is WM maintenance.

Updating tasks (e.g., n-back, keep-track, or arithmetic updating tasks) share with WMC measures (e.g., complex span, spatial short-term memory, or binding tasks) that people have to maintain information over a few seconds. Therefore, part of the variance of accuracy in updating tasks reflects WM maintenance. This is the reason why many Updating tasks are valid measures of WM capacity (Oberauer et al., 2000; Schmiedek et al., 2009; Wilhelm et al., 2013).

Despite the similarities between updating tasks and common measures of WMC, updating tasks require processes beyond WM maintenance. Specifically, updating tasks involve a combination of retrieving, transforming, and substituting or removing information stored in WM (Ecker et al., 2010). For instance, in an arithmetic updating task (e.g., Oberauer et al., 2000), each updating step involves retrieving one of the digits held in WM, transforming it according to a given arithmetic operation (e.g., "+2"), and substituting the old digit by the result. Other common tasks to assess updating – for instance the *N*-back (Kirchner, 1958), keep-track (Miyake et al., 2000), or running span tasks (Friedman et al., 2006) – require retrieval and substitution of information in WM but no transformation. Specifically, these tasks require selectively accessing some information in WM and substituting it by new information. To conclude, the selective replacement of outdated information is the characteristic feature of WM updating (Ecker, Lewandowsky, et al., 2014).

As these unique processes are what theoretically constitutes updating, assessing updating ability should neither be reduced to nor conflated with measuring WM maintenance. Yet, the studies that indicated stronger relationship of updating with Gf and WMC than for inhibition and shifting measured updating as the average accuracy in WM updating tasks (Friedman et al., 2006; Wongupparaj et al., 2015). This average performance score conflates updating-specific variation – the ability to replace outdated WM contents by new ones – with individual differences in WM maintenance, as measured by all short-term and working-memory tasks. In contrast, inhibition and shifting have been measured by difference scores between an experimental condition demanding the EF to be measured, and a control condition demanding it less. For instance, in the Stroop task (Stroop, 1935), inhibition is measured by the performance difference between congruent and incongruent trials, with a smaller difference reflecting more successful inhibition of the misleading word meaning. These difference scores isolate the variance due to EF by controlling for confounding processes (e.g., the efficiency of stimulus encoding, processing, and motor response).

Like measures used for inhibition and shifting, the EF demands in updating tasks can be isolated by subtracting performance in a control condition not involving updating from performance in an experimental condition requiring updating. The resulting difference represents the ability to efficiently update information without compromising memory performance. Thus, individuals with high updating abilities should show smaller performance losses between the two conditions than individuals with low updating abilities. Critically, previous studies did not isolate this updating-specific variance, and thus might have overestimated the strength of the relationship of updating with Gf.

The few studies that distinguished individual differences specific to updating processes and related them to other standard WMC measures or Gf found inconsistent results. For example, Ecker et al. (2010) adapted the arithmetic updating tasks described above for verbal material, and introduced updating steps that could include every possible combination of retrieval, transformation, and substitution demands. They found that only the accuracy of retrieval (r = .55) and of transformation (r = .49) were positively correlated with other common measures of WMC (a composite score of an operation span, a sentence span, and a spatial short-term memory task), but substitution accuracy was not. Individual differences in the speed of updating processes were unrelated to WMC. Similarly, Ecker et al. (2014) observed no correlation between the efficiency of removing old information from WM (i.e., the speed with which participants finished updating information in a self-paced updating task) and WMC. However, in a more recent study, Singh et al. (2018) found that removal efficiency was related to WMC ($-.23 < r < -.30$). They also found that Gf was related to removal efficiency (r = -.21), but this relationship was fully mediated by WMC, speaking against the suggestion that disengagement underlies the correlation between updating and Gf. In sum, findings are inconsistent regarding the relationship of cognitive processes specific to updating with Gf and WMC. Moreover, two out of the three described studies (Ecker, Lewandowsky, et al., 2014; Singh et al., 2018) focused on the efficiency of removal processes, thereby neglecting individual differences in the ability to accurately substitute information in working memory.

Present Study

In the present study, we investigated the relationship of updating to Gf and WMC by re-analyzing data published by von Bastian et al. (2016). The updating tasks in this dataset resemble commonly used keep-track tasks but, critically, contain trials with and without updating demands. Thus, these tasks allow for addressing two key limitations of the previous literature: conflating updating with maintenance (Friedman et al., 2006; Wongupparaj et al., 2015), and lack of accuracy-based paradigms (Singh et al., 2018). By contrasting the updating condition with a control condition requiring no updating at all – as is the standard procedure for inhibition and shifting measures – we isolated updating-specific variance associated with disengagement from variance related to WM maintenance.

Difference scores often suffer from poor reliability (Hedge et al., 2018). We circumvent this problem using Bayesian structural-equation models that isolate only reliable individual differences in the updating effect, and Bayesian generalized mixed models that additionally separated out trial-noise from the true-effect of updating (Rouder & Haaf, 2019). By isolating updating as an executive control process separate from WM maintenance, the present study provides a more valid assessment of the predictive power of updating for Gf and WMC than previous studies.

Furthermore, we examined two further aspects of WMC: storage and processing (WM SP), and relational integration (WM RI). WM SP refers to maintaining the representations of several memory items while processing distractors, and this is usually measured with complex span or Brown-Peterson tasks – which are also the paradigms used in this study. WM RI refers to building new relations between elements to create structural representations (Oberauer et al., 2000, 2003). WM RI is usually measured with tasks in which participants have to monitor ensembles of stimuli that change regularly and react when they form a specific constellation (e.g., a square, a rhyme, or some match between several elements). The inclusion of measures of WM SP and WM RI allowed for exploring whether updating is related differently to these two aspects of WMC.

In sum, our study aims to clarify the role of EF for complex cognitive abilities as reflected in WMC and Gf. Of the three psychometrically identified dimensions of EF – inhibition, shifting, and updating – only updating has shown a substantial correlation with Gf in previous studies (Friedman et al., 2006; Wongupparaj et al., 2015). Here, we test whether those findings were due to a confound between updating and maintenance, or whether a substantial correlation can also be established between specific measures of updating ability on the one hand, and WMC and Gf, on the other.

Method

Participants

Of the original sample (N = 121 young adults aged 19 to 35) collected by von Bastian et al. (2016), one participant had to be excluded due to an experimenter error. In addition, we discarded uni- and multivariate outliers identified by the Mahalanobis distance from the different measures. Specifically, data points with a Mahalanobis distance larger than $\chi^2_{p < .01}$ with df = N_{var} were discarded. Multi-variate outliers were first identified for each measure (i.e., Updating tasks, Gf, WM SP, and WM RI) separately and then across all measures. Thus, the present analyses are based on data from 111 participants (67 female, 44 male, M_{age} = 24.28, SD_{age} = 3.71) with an average of 15.88 years of education ($SD_{education}$ = 3.39) of which 94 were university students and 17 were not.

Measures

We analyzed the tasks tapping updating, WM SP, WM RI, and Gf used by von Bastian et al. (2016). Table 1 displays average performance and reliability estimates for the tasks tapping these constructs, and Table 3 displays their correlations. The correlation matrix of all variables is available in the Appendix.

Updating. The three updating tasks were similar in design to the keep-track task used by Miyake et al. (2000). Participants had to remember an initial set of items and subsequently update some of these items one by one, replacing them by new stimuli. At the end of each trial, participants were asked to recall the most recent items. Importantly, in some trials no updating occurred. In these trials, participants were prompted to recall the items directly following their encoding, hence these trials only required WM storage of the initial items. Code and scripts for running the tasks in Tatool Web (von Bastian et al., 2013) are available online at http://www.tatool-web.com/#/doc/lib-bat-uzh-ef-updating.html.

Table 1
Average Performance, Descriptive Statistics, and Reliability Estimates for the Sample (N = 111) and All Tasks and Measures Used in This Study.

Construct	Task	Updating	*M*	*SD*	Min	Max	Est. Rel.[a]
Updating	Figural	no	.70	.22	.20	1.00	.94
		yes	.59	.16	.15	.93	.94
	Numerical	no	.91	.13	.50	1.00	.92
		yes	.72	.19	.21	1.00	.95
	Verbal	no	.95	.08	.72	1.00	.84
		yes	.72	.12	.47	.97	.90
WM SP	Brown-Peterson		.80	.12	.45	1.00	.95
	Complex Span		.57	.15	.27	.88	.92
WM RI	Figural		2.64	.37	1.43	3.33	.40
	Numerical		2.85	.70	1.30	4.36	.70
	Verbal		2.75	.63	.80	4.02	.70
Gf	Diagramming relationships		.74	.14	.33	1.00	.61
	Letter Sets		.84	.14	.27	1.00	.62
	Locations		.68	.18	.20	1.00	.64
	Nonsense Syllogisms		.69	.15	.30	1.00	.41
	Raven's APM		.70	.21	.17	1.00	.66

Note. Performance was measured as proportion of correct responses, except for WM RI tasks, which used sensitivity (*d*'). WM = working memory; SP = storage and processing; RI = relational integration. APM = advanced progressive matrices; Min = minimum; Max = maximum; Est. Rel. = estimated reliability.
[a] Reliability was estimated via odd-even correlations and corrected for test length with the Spearman-Brown prophecy formula.

The updating tasks used materials from three different content domains: figural, verbal, and numerical. Panel A of Figure 1 illustrates the three tasks. In the *figural updating tasks*, participants had to remember, update, and recall the colors of five different shapes. Each updating step involved the presentation of one of the to-be-remembered shapes in a new color, and participants had to update the color of the respective shape. Using the same procedure, the *numerical updating tasks* used digits ranging from 1 to 9 in four different colors, and the *verbal* *updating tasks* used consonants (except “Y”) presented in five different locations on the screen. Thus, memory set size varied between 4 (numerical updating task) and 5 items (figural and verbal updating tasks). In addition, the number of updating steps in the three tasks varied from 7 (numerical), through 9 (verbal), to 10 (figural). All tasks comprised 20 trials with updating and 5 trials without updating, which were randomly intermixed. Although there were less trials without updating, reliability estimates (see Table 1) suggest that individual differences in performance could still be measured adequately.

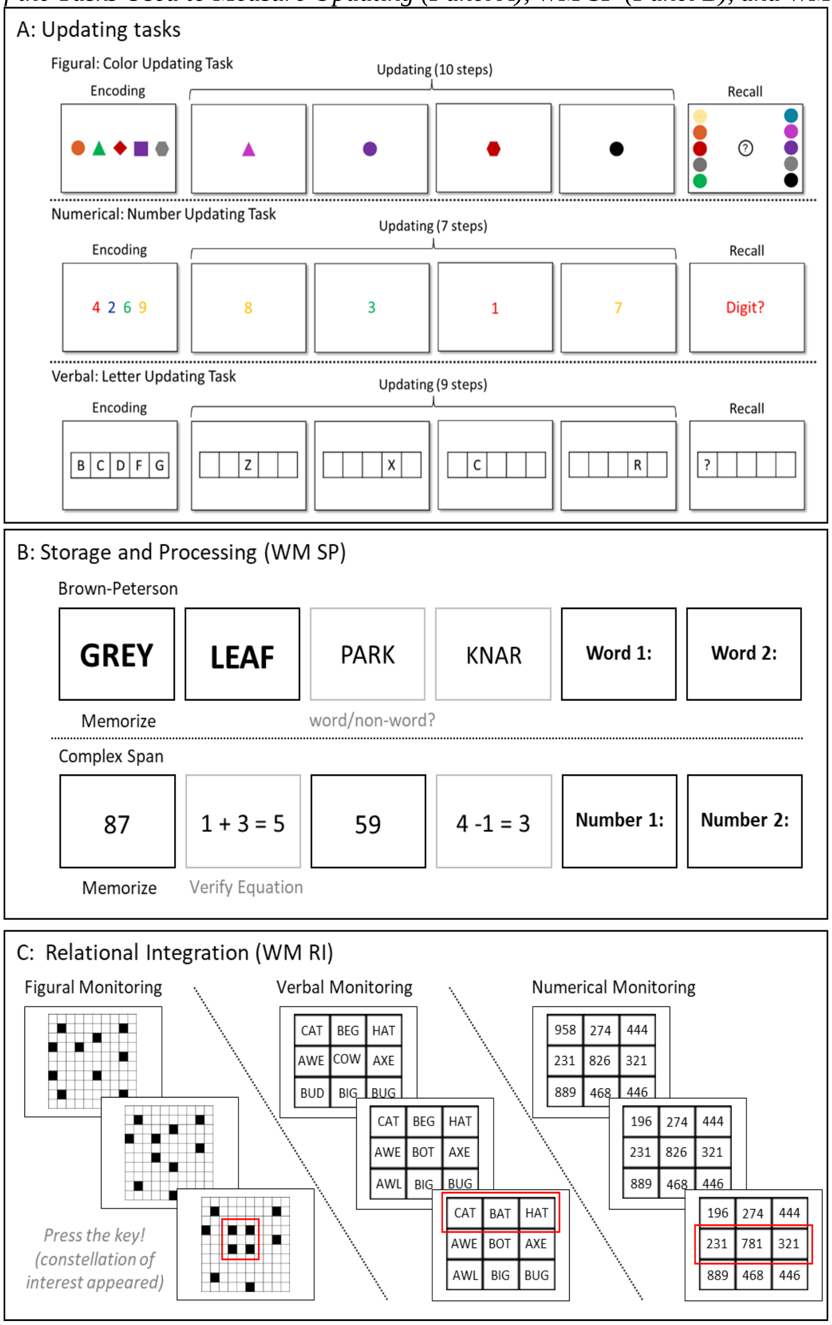
For structural equation modeling (SEM), the performance measure in the updating tasks was the proportion of correctly recalled items in trials with and without updating. For additional analyses with Bayesian hierarchical models, we used the number of correctly recalled items in each trial as performance indicator.

WM SP. Individual differences in the ability to simultaneously store and process information were measured with two tasks. In the *Brown-Peterson task* (see Figure 1, Panel B), participants first memorized 3-6 words and then performed five lexical decisions on four-character strings. At the end of each trial, participants had to recall the words in correct serial order. In the *complex span task* (see Figure 1, Panel B), participants had to remember three to six two-digit numbers while judging the correctness of a mathematical equation in between each of the memoranda. At the end of each trial, participants had to recall the memoranda in correct serial order.

The performance measure in both tasks was the proportion of correctly recalled memory items at their respective serial positions. To facilitate the use of WM SP measures in Bayesian hierarchical models, the performance measures of the two tasks were aggregated by a principal component analysis to one score.

Figure 1

Illustration of the Tasks Used to Measure Updating (Panel A), WM SP (Panel B), and WM RI (Panel C).



Note. In each of the *updating tasks* (A), participants initially encoded a memory set of 4 to 5 stimuli (colors, digits, or letters). Some trials required replacing one item at a time whenever a new stimulus was displayed for 7, 9, or 10 updating steps; in the other trials recall directly followed encoding. In the *WM SP tasks* (B), participants encoded words or two-digit numbers and had to process distractors either after encoding of all memoranda or interleaved with the encoding of memoranda. In the end, they had to recall the memoranda in forward order. In the *WM RI tasks* (C), participants had to monitor a set of stimuli, of which one or two changed unpredictably on each step. As soon as the stimuli formed a specific constellation – for example, four boxes forming a square, all words in a row or column rhyme, or all number in a row or column end on the same digit – participants had to press the space bar. This is illustrated in the figure by the red frame around the relevant constellation.

WM RI. The ability to build new relations between multiple elements and integrate them into structural representations was measured by three monitoring tasks (Oberauer et al., 2003; von Bastian & Oberauer, 2013). In these tasks (see Figure 1, Panel C), participants had to monitor an array of stimuli, some of which were replaced every 2 s, and press the space bar whenever they detected that a critical constellation between a subset of the stimuli occurred. Again, the tasks tapped into three different content domains with figural, verbal, and numerical material.

In the *figural monitoring tasks*, two of 20 dots changed their position in a 10x10 grid every 2 s, and participants had to monitor whether any four dots in the grid formed a square. In the *verbal monitoring task*, 1 of 9 words in a 3x3 grid changed every 2 s, and participants had to monitor whether three words in any direction across the grid (horizontal, vertical, or diagonal) rhymed. In the *numerical monitoring task*, 1 of 9 three-digit numbers in a 3x3 grid changed every 2 s, and participants had to monitor whether three numbers in any direction (horizontal, vertical, or diagonal) had the last digit in common.

The performance measure in the monitoring task was the sensitivity *d*' of the detection performance (i.e., *z*(Hits) – *z*(False Alarms)). For participants with a perfect hit or false alarm rate, the rates were corrected to a hit rate with ½ miss and a false alarm rate of ½ false alarm to avoid *d*' = ± Infinite. Like WM SP measures, the WM RI measures were aggregated by a principal component analysis for Bayesian hierarchical modeling.

Gf. Participants' reasoning ability was assessed with five time-restricted tests. In the short version of the *Raven's Advanced Progressive Matrices* (Arthur et al., 1999; Arthur & Day, 1994), participants had to complete a matrix pattern and choose the correct response from eight alternatives. In the *Locations Test* (Ekstrom et al., 1976), participants had to select the correct location of an "✕" by identifying the patterns of "✕" in four preceding rows of dashes. In the

Letter Sets Test (Ekstrom et al., 1976), participants had to select one letter set that deviated from a regular pattern among a set of five letter sets. In the *Nonsensical Syllogisms Test* (Ekstrom et al., 1976), participants had to decide whether conclusions drawn from two nonsensical premises were logically valid. Finally, in the *Diagramming Relationships* (Ekstrom et al., 1976), participants had to choose one out of five diagrams that best represented the set relations of three nouns. For all reasoning tasks the performance measures were the proportion of correctly solved items. Again, performance was aggregated by a principal component analysis over all tasks for Bayesian hierarchical modeling.

Statistical Analyses

In light of multiple possible analytical approaches, and to increase robustness of our results, we adopted a multiverse approach (Steegen et al., 2016) in our statistical analysis. Specifically, we used two structural equation models (SEM) for measuring latent change, as well as hierarchical Bayesian hierarchical models, to evaluate how updating-specific variance is related to reasoning and working memory capacity. Convergence of results across these different analytical choices can increase our confidence that the outcome is not limited to only one set of modeling specifications. Raw data and scripts to preprocess and analyze the data can be accessed at osf.io/zkd4c.

Data preprocessing. We preprocessed all data similar to the procedure described by von Bastian et al. (2016). For the SEMs, all variables were *z*-standardized to avoid ill-defined covariance structures due to large differences in the absolute variance of the different measures. For Bayesian hierarchical models, only the covariates (i.e., WM SP, WM RI, and Gf) were *z*-standardized.

SEM. We used a version of latent change models (McArdle, 2009; McArdle & Hamagami, 2001; Steyer et al., 1997) to isolate updating-specific variance from variance of WM maintenance. Latent-change models or latent-difference models are typically used in longitudinal research to estimate changes in constructs over time (see Figure 2 for an illustration). In these models, one latent factor reflects the intercept that captures initial individual differences in the construct – let's say an ability measured at time-point 1 (let's term this the Intercept factor; see Figure 2A). This Intercept factor predicts another latent factor for the second measurement (typically at a different time) with the second factor capturing individual differences in the intercept and the change. The residual of the second latent factor captures the mean and variance in the change from the initial measurement occasion (McArdle, 2009; McArdle & Hamagami, 2001). Alternatively, such latent-change models can also be specified as bi-factor models that capture variance consistent across different measurement occasions in an *intercept* factor on which all indicators load, and that captures variance induced by the change with a second factor (*change)* on which only the indicators from the second measurement load (see Figure 2B; Steyer et al., 1997).[2]

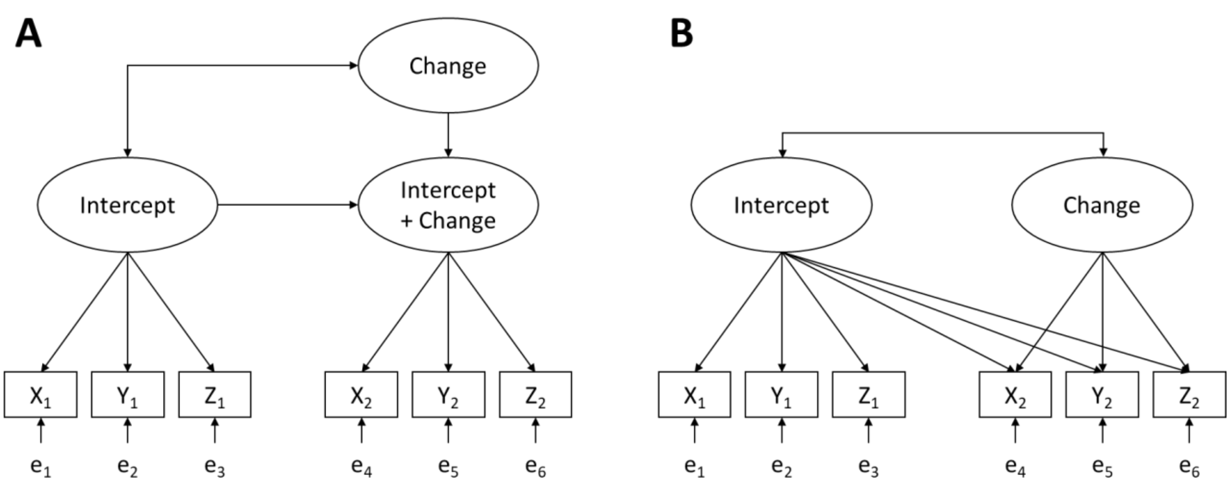
Albeit less conventional, the structure of latent-change models can be applied to estimate latent differences between experimental conditions, by letting the intercept represent individual differences in the control condition, and the change residual represent individual differences in how much the dependent variable in the experimental condition differs from that in the control condition (see Meisel et al., 2019). In EF research, bi-factor models are typically used to capture common and unique variance shared between different EF (for a review, see Karr et al., 2018). In

[2] For recent papers introducing the application of latent change models in developmental cognitive neuroscience or intervention studies see Kievit et al. (2018) or Könen & Karbach (2021).

this context, they are usually not interpreted as change models because measures of different EF (i.e., updating, shifting, and inhibition) are hardly comparable. Yet, specifying a bi-factor model contrasting two experimental conditions within the same task closely resembles a latent change model.

Although the specification of latent change models is similar in both longitudinal and experimental contexts, we note one important difference. In longitudinal applications researchers are interested in measuring the time-related change in a single construct. This necessitates measurement invariance to ensure that changes can be attributed to a change in the same construct over time. By contrast, in experimental contexts we assume that the construct changes due to the different requirements in the experimental conditions. It is exactly this difference that we aim to isolate and, therefore, imposing measurement invariance would invalidate the application of latent change models in experimental contexts. In addition, latent

Figure 2.
Simplified Path Diagrams of Two SEM Isolating Individual Differences in Latent Differences Between Two Measurements.



Note. In both models three indicators are used for repeated measurements, either longitudinal (i.e. the same construct at two measurement occasions), or experimental (i.e. the same indicators in two experimental conditions). The left implementation (A) shows the latent-change, or latent-difference model. It specifies two latent variables for both measurements and isolates variance specific to the second measurement (i.e. the change) through a regression. The right implementation (B) represents a bi-factor approach implementing one factor capturing the shared variance between the first and second measurement and isolates variance specific to the second measurement in a second latent factor. Conceptually both implementations achieve the same estimation of individual differences in a latent difference and only differ in minor details regarding their implied covariance structure.

change/difference models often estimate both mean and covariance structure. However, if we are only interested in individual differences, it is sufficient to focus on the covariance structure. In fact, the mean structure only indicates whether there was an overall performance change between trials that required updating versus trials that did not require updating.

The SEMs were estimated using Bayesian estimation procedures of the package *blavaan* (Merkle & Rosseel, 2018) implemented in *R* (R Core Team, 2018). The benefit of Bayesian SEM over frequentist SEM is that, in combination with adequate priors, they provide more adequate parameter estimation in smaller samples (McNeish, 2016). We used the following priors for BSEM: for variance parameters, we used gamma priors with a shape of 1 and rate of .05, for covariance parameters, we used beta priors with $\alpha = 1$ and $\beta = 1$ extended in range from -1 to +1, and for factor loadings and regression weights, we used normal priors with $\mu = 0$ and $\sigma = 10$. In general, these priors do not severely constrain parameter estimates to specific values, except for the gamma priors for variance estimates that prevent variances from becoming negative, and the beta priors for covariances that prevent Haywood cases (i.e., absolute correlations larger than one). The gamma prior ensures that the credibility interval for variance estimates cannot include zero, and therefore, to test whether a variance is credibly different from zero, we need to compare a model fixing the variance to zero with a model estimating the variance freely.

Parameters were sampled using the *no U-turn* sampler (NUTS) implemented in STAN (Carpenter et al., 2017) with four independent MCMC chains that each consisted of 2000 warmup samples and 5000 samples after warmup. To check convergence of the Bayesian parameter estimation, we required that the potential scale reduction factor (PSRF) was below 1.05. The PSRF (a.k.a. $\hat{R}$) is the ratio of variance within each MCMC chain to the variance between the different chains. PSRF values close to 1.00 indicate perfect convergence, whereas larger values indicate insufficient convergence.

We judged absolute model fit of BSEM using the posterior predictive *p*-value (PP *p*) and a Bayesian implementation of the root-mean square error of approximation (BRMSEA). PP *p*-values close to zero indicate a bad model fit, whereas values close to 0.5 indicate good model fit. We follow the recommendations by Muthén and Asparouhov (2012) in requiring the estimated BSEM to show at least PP $p > .05$ for the model to be retained for interpretation. In addition, we judged relative model fit in comparison to a baseline model – only estimating variances of all manifest indicators and fixing all covariances between indicators to zero – with a Bayesian implementation of the comparative fit index (BCFI). For the BRMSEA and the BCFI, we used the following cutoff criteria to assess model fit: BRMSEA < .05; BCFI > .95. We also report mean posterior estimates and the 95% highest density interval.

Another benefit of the Bayesian estimation of SEM is that we were able to compare models via Bayes factors (BFs). Specifically, BFs quantify the extent to which one BSEM is to be favored over another, thereby quantifying evidence in favor of a simpler model, unlike non-significant differences between nested models obtained in Chi-Square difference tests. By measuring the strength of evidence on a continuous scale, model comparisons via BFs also indicate whether the evidence might be inconclusive (i.e., BFs close to 1). Given the rather small sample size of the current study, this feature ensures that we do not overinterpret results that lack sufficient evidence to select one model over another.

Bayesian hierarchical models. One recently raised critique of estimating change scores and latent change factors in SEMs is that the aggregation of performance over trials in different experimental conditions fails to separate trial-to-trial noise from true between-subject and experimental-effect variance (Rouder & Haaf, 2019). This might decrease the amount of reliable variation that can be detected in the experimental effect (in this case the updating-specific variance). To address this limitation, we additionally ran Bayesian hierarchical generalized linear mixed models (BGLM) as suggested by Rouder and Haaf (2019).

In the BGLMs the number of correctly recalled items in each trial in the three updating tasks was predicted by the content domain of the tasks (i.e., figural, verbal, numerical) and the updating factor (i.e., whether a trial contained updating or not). These experimental effects represent different task difficulties, namely, lower accuracy in trials requiring updating compared to trials without updating. To model individual differences in the updating effect, we included random slopes for both the effects of task content and of updating requirement. These random effects reflect variation in the experimental effects across individuals. To investigate whether any of the three covariates is related to individual differences in the updating effect, the three covariates (Gf, WM SP, and WM RI) were included separately as additional predictors for performance in the updating tasks.[3] Regarding the question to what extent updating is related to Gf, WM SP, or WM RI, the important parameter in this BGLM is whether the covariate predicts individual differences in the updating effect across the three tasks. This is reflected in the cross-level interaction between the experimental updating effect and individual differences in the covariate. This interaction can also be interpreted as a difference in the correlation of the covariate with performance in trials with and without updating. Specifically, if a covariate such as Gf has a larger (positive) regression weight for predicting updating performance than for predicting no-updating performance, then the size of the updating effect is smaller for people with higher than those with lower Gf, which can be described as an interaction of the updating effect with Gf.

---

[3] As the estimation of BGLM is time consuming and estimating additional correlations between predictors is difficult, we estimated separate models for each of the three covariates (i.e., WM SP, WM RI, and Gf).

Regarding our research question, we thus tested whether this interaction between updating and the respective covariate was credibly different from zero. Specifically, we first evaluated whether the 95% credibility interval (CI) of the posterior of the interaction included zero. In addition, to quantify evidence for the absence of an interaction between updating and the three covariates, we compared a model including that interaction, and the three-way interaction of task content, updating, and the covariate, to a model not including these interactions. Evidence for or against either of the two models was evaluated with BFs and posterior probabilities (PP) of the two models estimated via bridge sampling (Gronau et al., 2018). To establish the robustness of the BF and the PP estimation we estimated models and BFs 10 times. In the results, we report the smallest BF, and the PP for the favored model, so that the values reflect the lower limit for the estimation of the evidence for one or the other model.

The BGLMs were estimated using the *brms* package (Bürkner, 2017). As accuracy of each recall in the updating tasks follows a binomial distribution (0 = incorrect, 1 = correct), we modeled recall performance in each trial with a binomial distribution and a logit link function. For fixed effects (i.e., the intercept and group level effects) we used normal priors with $\mu = 0$ and $\sigma = 1$. For random effects, reflecting variation of effects across individuals, we used half Cauchy priors with a location of zero and a scale of 2. Parameters were estimated with four MCMC chains each containing 1000 warmup samples and 10,000 samples after warmup. To ensure convergence of the parameter estimation, we again checked that all PSRF values were below 1.05.

Results

What Is Measured by Updating Tasks?

First, we decomposed the common variance of the three updating tasks into two components of variance: (a) individual differences in WM maintenance and (b) individual differences related to updating-specific variance. Specifically, we contrasted the two options of specifying latent-difference models (i.e., latent change vs. bi-factor specification) described in the Method section, and estimated parameters for all models. In addition, we tested whether there was credible updating specific variance by fixing the updating-specific variance in the models to zero. Table 2 summarizes the fit indices for the four models and model comparisons via BFs. Figure 3 depicts the path diagrams of the models and their estimated parameters.

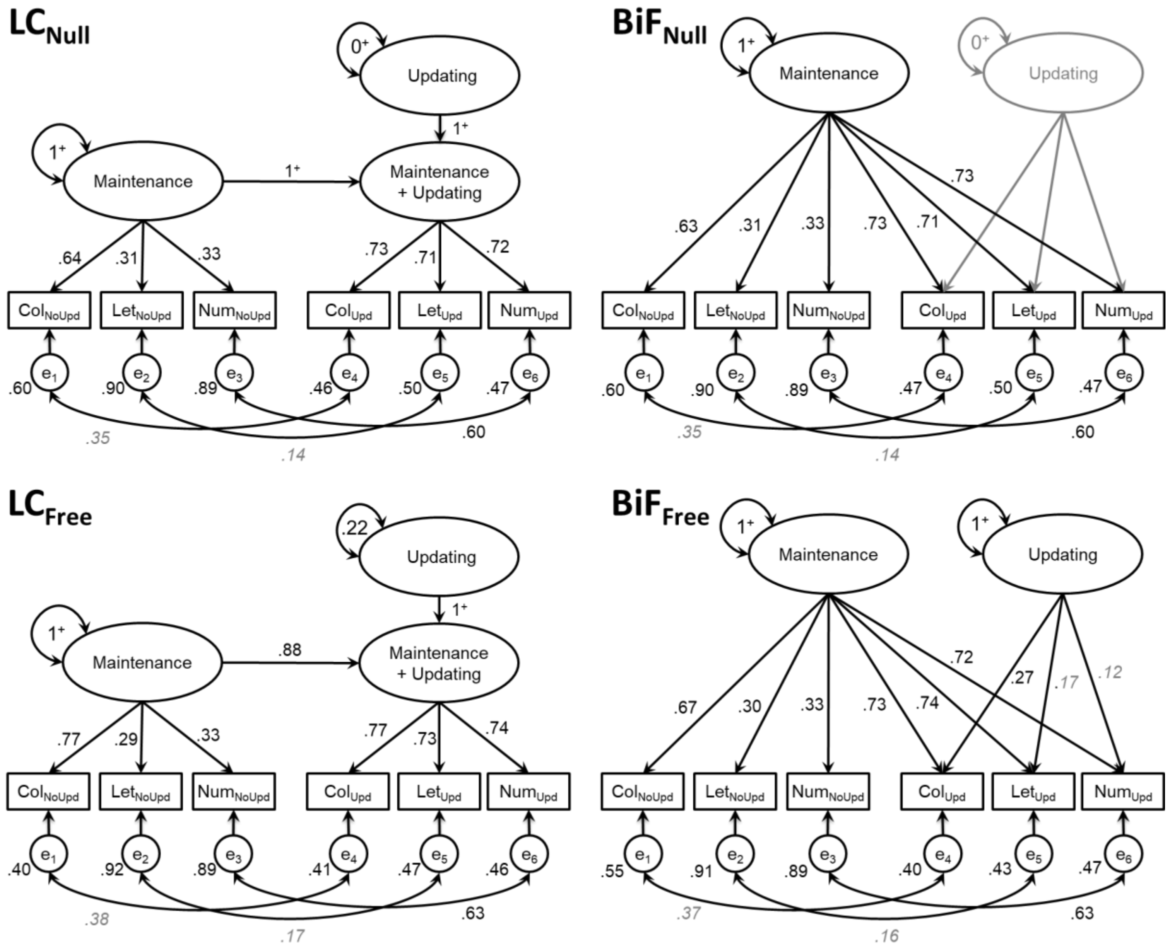
As can be seen from the model comparisons (see BFs in Table 2), the two models fixing the updating-specific variance to zero (Bi-Facto$_{\text{Null}}$ and Latent Change$_{\text{Null}}$) fitted the data best, and equally well (BF ≈ 1). This is also reflected in the variance captured by the updating-specific factor in the two models freely estimating the updating-specific variance (Bi-Factor$_{\text{Free}}$ and Latent Change$_{\text{Free}}$, see Figure 3). In the latent-change model (Latent Change$_{\text{Free}}$), the 95%

Table 2
Model Fit of the Models Isolating Individual Differences Specific to Updating from Individual Differences in Maintenance.

Model	N_{par}	PSRFs <	PP *p*	BRMSEA	BCFI	BF_{01}
Latent Change$_{\text{Null}}$	**15**	**1.00**	**.616**	**.02 [.00; .10]**	**.99 [.95; 1.00]**	
Latent Change$_{\text{Free}}$	16	1.00	.613	.02 [.00; 11]	.99 [.95; 1.00]	39.88
Bi-Factor$_{\text{Null}}$	15	1.00	.629	.02 [.00; .10]	.99 [.95; 1.00]	1.07
Bi-Factor$_{\text{Free}}$	18	1.00	.570	.03 [.00; .12]	.99 [.94; 1.00]	1.6×10^4

Note. N_{par} = number of freely estimated parameters in the model, PSRF = potential scale reduction factor, PP *p* = posterior predictive *p*-value, BRMSEA = Bayesian RMSEA, BCFI = Bayesian CFI. For Bayesian fit indices we report the posterior mean and the 95% highest density interval in the squared brackets. Bayes Factors are computed in comparison with the best fitting model, which is highlighted in bold.

Figure 3. Latent change model separating individual differences in WM maintenance from individual differences in updating-specific processes.



Note. Values for parameters refer to the posterior mean of the posterior distribution of parameters. Parameters printed in gray and in italics had 95% credibility intervals including zero. Variances and factor loadings are given as standardized parameters. + = Parameter was fixed to the depicted value. Col = Color, Let = Letter, Num = Number, NoUpd = no updating, Upd = updating.

credibility interval for the updating specific variance did not include zero , 95% CI = [.10; .48], indicating that there was credible updating-specific variance across all three tasks (12 to 13% of variance in the manifest indicators). In the bi-factor model ($\text{Bi-Factor}_{\text{Free}}$), the loadings from the updating-specific factor on indicators from trials requiring updating in the three tasks were small, and only credible for the color updating task (i.e. Col_{Upd}), indicating that there was little credible updating-specific variance across the three tasks (1.5 to 7% of variance in the manifest indicators). The difference in the credibility in the updating-specific variance for the different tasks can be explained by an additional proportionality constraint in the latent difference specification compared to the bi-factor specification. In detail, the latent change specification assumes that the contributions of the *maintenance* and the *updating* factor differ by a constant proportion for the three indicators of the *maintenance + updating* factor. The bi-factor model (Bi-Factor$_{\text{Free}}$) does not include this constraint. Yet, a similar assumption can be included in the bi-factor specification through constrainting the ratio of the loadings from the *updating* and the *maintenance* factor to the same value (i.e. b_M/b_U = constant) for the updating trials of the three tasks. In this bi-factor model (Bi-Factor$_{\text{Const.}}$, see Figure 4)[4], the loadings from the updating factor on trials requiring updating are positive and credibly different from zero for all three tasks, and the amount of updating-specific variance in trials requiring updating is comparable to the latent-difference specification (between 14 to 16% of variance).

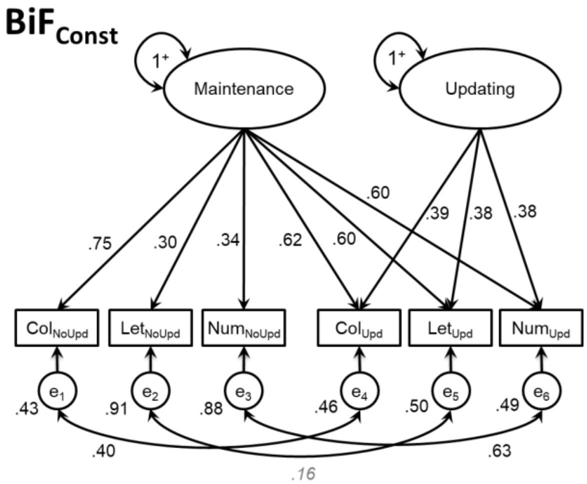
Nonetheless, irrespective of the specification of the BSEM isolating updating-specific variance, the model comparisons indicate that a model without any updating-specific variance is to be preferred over any of the models freely estimating updating specific variance. This was not due to the maintenance factor capturing all variance in the indicators, as this factor explained only between 10 to 40% of variance for indicators not requiring updating[5], and 48 to 51% of

---

[4] Due to limitations in implementing constraints for BSEM estimated in STAN, we had to estimate parameters for this model using JAGS. Therefore, we could not compute Bayes Factors comparing the constrained bi-factor model with the other models. A frequentist estimation of the different models that is included in the online supplementary material illustrates that the constrained bi-factor model and the latent change model freely estimating the updating specific variance imply the same covariance structure and thus fit the data equally well.

[5] The reason that both letter and numerical trials without updating loaded weakly on the maintenance factor is likely a restriction in variance due to ceiling effects. The average proportion correct for both these indicators was >.90. In contrast, the remaining non-updating indicator that loaded more strongly on the maintenance factor had

Figure 4. Path diagram of the bi-factor model including the same proportionality assumption as the latent difference model, freely estimating the updating-specific variance in the three tasks.



Note. The ratio of loadings from the maintenance and updating factors on the three indicators for performance in trials requiring updating (Col_{Upd}, Let_{Upd}, and Num_{Upd}) is constant (.62/.39 ≈ .60/.38 ≈ .60/.38).Values for parameters refer to the mean of the posterior distribution of parameters. Parameters printed in gray and in italics had 95% credibility intervals including zero. Variances and factor loadings are given as standardized parameters.

variance for indicators requiring updating. Hence, there was still a large portion of variance left to be explained in the upating condition of the different tasks. These results indicate that there was little domain-general variability specific to updating across these tasks.

Strictly speaking, these results preclude any further investigation of relationships of updating-specific variance with the other covariates (i.e. Gf, WM SP, and WM RI), because the model not including any updating-specific variance provided a better fit to the data. However, because investigating these relationships was the main aim of the current study, we still estimated the relationships of updating-specific variance with the three covariates using the latent change model (Latent $Change_{Free}$). Yet, if any of these relationships would have been credibly different from zero, they would need to be interpreted carefully and need replication with

lower average proportion correct (.70) and a larger variance (.22), suggesting that with sufficient variance non-updating and updating trials capture individual differences in maintenance to a similar extent.

credible updating variance in alternative analyses (like the Bayesian hierarchical models reported below) or in future studies.

Relationship of Updating with Reasoning and WMC

Our main question was whether WM maintenance or updating-specific processes are related to the three covariates: Gf, WM SP, and WM RI. To address this question, we estimated four separate BSEMs that included the three covariates into the latent-change model for the updating tasks. Specifically, Model I freely estimated the relationship between the *Maintenance*

Table 3

Summary of Model Fit Indices for the Measurement Models of the Three Covariates, And for the Joint BSEMs Estimating the Relationship Between the Maintenance and Updating Factors With the Three Covariates.

MM: Covariates			N_{par}	PSRF*s* <	PP *p*	BRMSEA	BCFI	
G*f*			10	1.00	.732	.00 [.00; .08]	.99 [.89; 1.00]	
WM SP			2	1.00	.821	.00 [.00; .08]	.99 [.91; 1.00]	
WM RI			2	1.00	.866	.00 [.00; .03]	.99 [.94; 1.00]	
Joint Models	Maint. - Cov	Upd -Cov	N_{par}	PSRF*s* <	PP *p*	BRMSEA	BCFI	BF_{01}
LC I	free	free	41	1.00	.314	.04 [.01; .06]	.95 [.91; .1.00]	5.08
LC II	**free**	**0**	**38**	**1.00**	**.260**	**.04 [.02; .06]**	**.95 [.90; .99]**	
LC III	0	free	38	1.00	.027	.06 [.05; .07]	.88 [.83; .92]	1.7×10^4
LC IV	0	0	35	1.00	.002	.07 [.06; .08]	.83 [.79; .87]	3.5×10^8

Note. MM = measurement model, N_{par} = number of freely estimated parameters, PSRF = potential scale reduction factor, PP *p* = posterior predicitve *p*-value, BRMSEA = Bayesian RMSEA, BCFI = Bayesian CFI, Maint. = maintenance, Cov = covariates, Upd = updating, BF = Bayes Factor.
For Bayesian fit indices we reported the posterior mean and the the 95% highest density interval in the squared brackets. Bayes Factors are computed in comparison with the best fitting model, which is highlighted in bold.

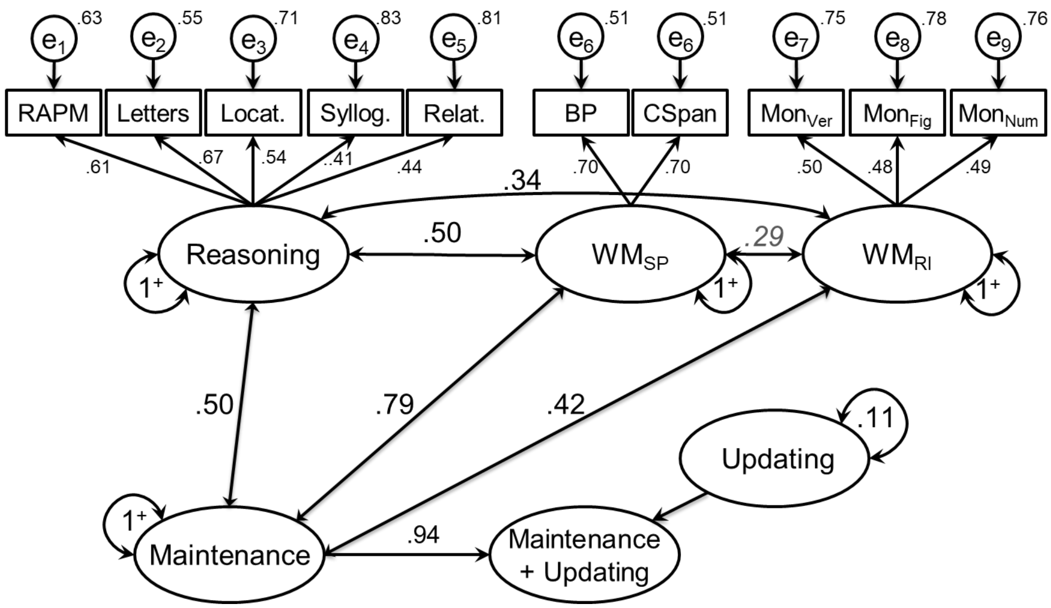
factor, the *Updating* factor, and all covariates. Model II fixed the relationship between the updating factor and the covariates to zero. Conversely, Model III fixed the relationship between the maintenance factor and all covariates to zero. Finally, Model IV fixed the relationship between both the maintenance factor and the updating factor and the covariates to zero.

To rule out potential misfit of the joint models due to inadequate measurement models for any of the three covariates, we estimated the model fit for the measurement models for Gf, WM SP, and WM RI, before estimating the four joint models. As can be seen in the top part of Table 3, all three measurement models fit excellently to the data. In detail, we fit a τ-congeneric model for Gf estimating all loadings from the latent Gf factor on the five indicators freely, and likewise estimating all error variances of the five indicators freely. For both WM SP and WM RI we estimated τ-equivalent measurement models constraining the loadings of all indicators on the latent factor to be equal. In addition, we also constrained the error variances of all indicators to be equal. For WM SP this was necessary to achieve an over-identified measurement model, for WM RI this was the most parsimonious and still well-fitting measurement model.

The bottom part of Table 3 summarizes the absolute and relative model fit of the four joint models estimating the relationship of maintenance and updating with the three covariates. The comparison of the four models via BFs suggested that Model II, allowing only relationships between the *Maintenance* factor and the three covariates, provides the best and most parsimonious description of the observed covariance structure. Specifically, the BF comparison indicates that the model fixing relationships of updating with any of the covariates to zero (Model II) is 5 times more likely than a model freely estimating the relationships of both maintenance and updating with the three covariates (Model I). In line with this, the relationship of the updating factor with the three covariates estimated in Model I were small to moderate, and their 95% credibility intervals included zero (Gf: r = -.33, 95% CI = [-.96; .56]; WM SP: r = -.03, 95% CI = [-.89; .75], WM RI: r = .52, 95% CI = [-.48; .98]). Thus, Model II (see Figure 5) was retained for interpretation. In this model, the factor capturing WM maintenance in the updating tasks showed the largest correlation with WM SP, r = .79 (95% CI = [.58; .96]); the correlations with Gf, r = .50 (95% CI = [.26; .72]), and with WM RI, r = .42 (95% CI = [.11;

72]), were still substantial. This implies that updating tasks capture, to a large extent, individual differences shared with tasks tapping WM SP, and their shared variance reflects the ability to maintain information.

Figure 5. Graphical illustration of LC II, freely estimating only the correlation between individual differences in WM maintenance and the covariates**.**

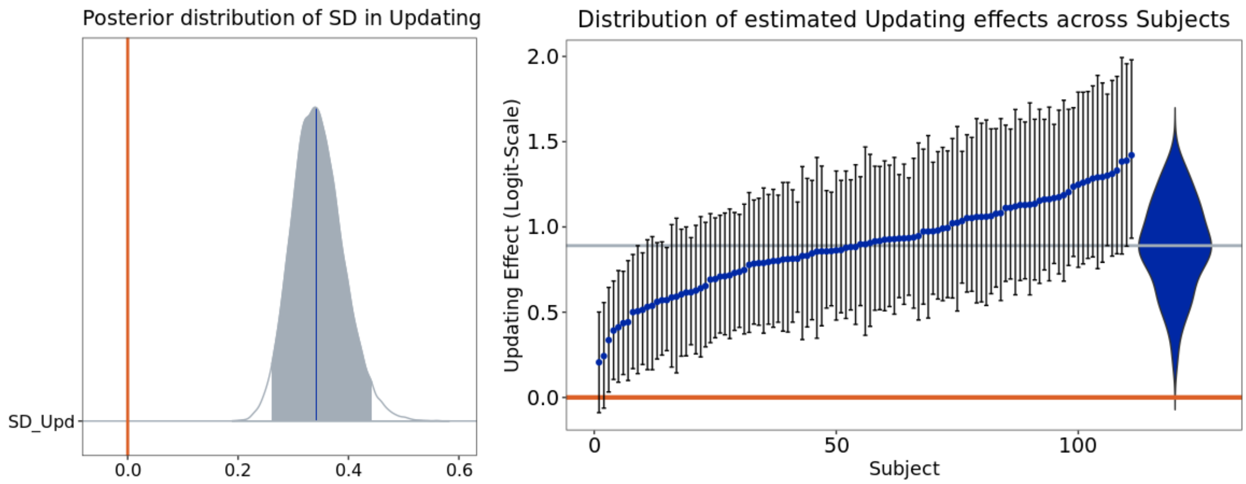


Note. Parameter values refer to the posterior mean. Parameters printed in gray and italics had 95% credibility intervals that included zero. All factor loadings and variances are reported as unstandardized parameters, except for correlations, which are standardized. Variances with superscript + were fixed to 1.
WM = working memory, SP = storage & processing, RAPM = Raven's advanced progressive matrices, Locat. = Locations, Syllog, = Nonsense Syllogisms, Relat. = Diagramming Relationships, Mon = monitoring, BP = Brown-Peterson, CSpan = complex span, Ver = verbal, Fig = figural, Num = numerical.

Alternative Analysis: Bayesian Hierarchical Generalized Linear Mixed Models

The BGLM results captured the experimental effects across the three updating tasks (i.e., accuracy was lower in trials with updating than without updating), and the variation reflecting individual differences in overall accuracy, and in the updating effect, across the three tasks (see supplementary material online at: osf.io/zkd4c). Like the BSEMs, the BGLM showed credible variability across individuals in the updating effect, σ_{Upd} = 0.34 (95% CI = [0.26; 0.44]; see Figure 6). Yet, unlike in BSEM, fixing this variance to zero across participants considerably impaired model fit, BF < 9.4×10^{33}, PP_{full} > .99; $PP_{\text{constrained}}$ < .01. Thus, the BGLM captured variance in the updating effect that could not be fixed to zero. The variation in the updating effect corresponded to about 6.2% (95% CI = [3.6; 9.1]) of the variance in observed accuracies. In contrast, variation in overall performance (i.e., the intercept) captured about 38.2% (95% CI =

Figure 6. Posterior distribution of estimated variance in the updating effect (left side) and distribution of updating effects across all subjects (right side).



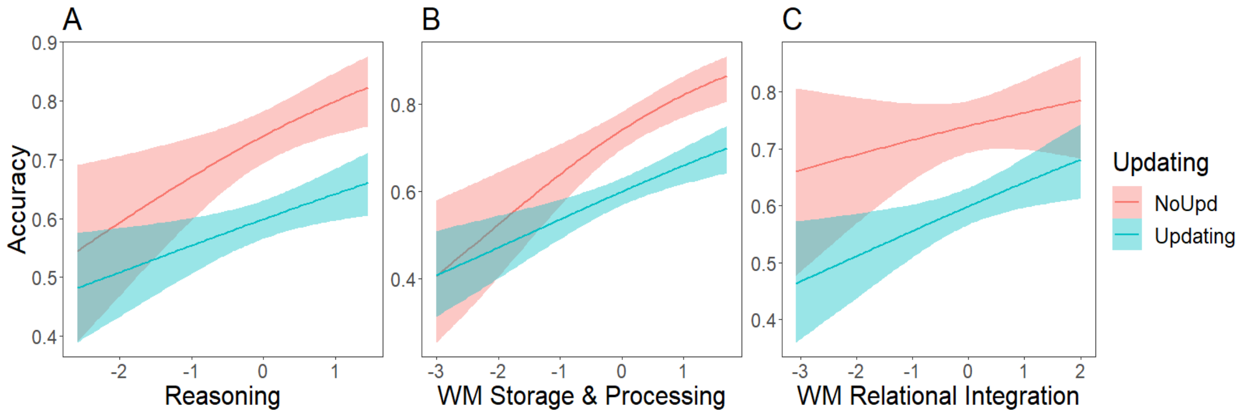
Note. The individual effects displayed on the right refer to the individual difference in performance (on the logit-scale) between trials with and without updating across all three updating tasks. For illustration purposes, they were arranged from the smallest to the largest individual effect. Error bars show the 95% highest density interval of each effect, and the violin plot illustrates the distribution of individual effects.

[29.9; 46.9]) of the variance in observed accuracies. Hence, by modeling trial-by-trial data, and thereby isolating trial noise, the BGLM measured true individual differences in updating.

Relationship of updating with the covariates. To test whether any of the three covariates – Gf, WM SP, or WM RI – was related to individual differences in the updating effect, we estimated BGLMs with each of the three covariates, each including the effects of task (figural, numerical, verbal), updating (trials with vs. without updating demands), and one of the three covariates, as well as interactions between the three effects. Figure 7 illustrates the results.

BGLM: Updating and Gf. As illustrated in Figure 7A, including Gf as predictor for accuracy across the three tasks, and trials with and without updating, showed that people with higher Gf had higher accuracy in the updating tasks, β = 0.31 (95% CI = [0.15; 0.47]). However, there was no credible evidence that Gf predicted performance in trials with and without updating differently, β = 0.06 (95% CI = [-0.03; 0.15]). Thus, we compared the full model to a model without the corresponding interaction of Gf and updating. The BF as well as posterior model probabilities (PP) indicated that the no-interaction model was more likely than the full model, BF

Figure 7. Illustration of the prediction of overall accuracy for trials with and without updating in the three BGLMs including (A) reasoning ability, (B) WM storage & processing, and (C) WM relational integration as predictor.



Note. The shaded red and blue area around the regression lines indicates the 95% credibility area around the regression curve. Please note that we estimated a linear model on the logit scale. As the logit scale does not transform linearly on the accuracy scale the displayed linear regressions are curved on the accuracy scale.

$> 1.1 \times 10^4$; $PP_{full} < .01$; $PP_{no\text{-}interaction} > .99$.[6] If anything, the direction of the interaction effect suggests that participants with lower Gf showed smaller decreases in performance in updating trials compared to no-updating trials.

BGLM: Updating and WM SP. As shown in Figure 7B, people with higher WM SP scores had higher overall accuracy in the updating tasks, β = .42 (95% CI = [0.27; 0.57]). Again, there was no credible evidence that WM SP predicted variations in the updating effect, β = 0.07 (95% CI = [-0.01; 0.16]). Although close to being credible, the direction of this effect implied that, if anything, participants with lower WM SP ability showed smaller deteriorations in performance in updating trials compared to no-updating trials, which is the opposite of what one would

[6] To establish the robustness of the BF and the PP estimation we estimated the models and the corresponding BFs/PPs 10 times. We report the smallest BF, and the smallest PP for the superior model, so that the values estimate the lower limit for the estimation of the evidence for one or the other model. See Method for further details.

expect. The model without the interaction was more likely than the model including the interaction, BF > 17.4; $PP_{full} < .05$; $PP_{no\text{-}interaction} > .95$.

BGLM: Updating and WM RI. Figure 7C illustrates the relationships of WM RI with performance in the updating tasks. Like the other covariates, people better in WM RI had higher overall accuracy in the updating tasks, $\beta = 0.18$ (95% CI = [0.01; 0.35]). WM RI also did not credibly predict variability in the updating effect, $\beta = -0.04$ (95% CI = [-0.12; 0.06]). Again, a model without the interaction was clearly favored over the model including the interaction BF > 3.1×10^4; $PP_{full} < .01$; $PP_{no\text{-}interaction} > .99$.

Discussion

The primary goal of the current study was to investigate whether one EF, namely the ability to update WM, can account for variance in general cognitive ability, as reflected in Gf and WMC. Previous studies (Friedman et al., 2006; Wongupparaj et al., 2015) have found larger relationships of Gf and WMC to updating than to two other executive functions, namely inhibition and shifting. We investigated whether these findings were due to differences in their measurement (average vs. difference scores), or whether updating-specific processes are truly more closely related to Gf and WMC. For this purpose, we isolated individual differences in updating-specific processes in three commonly used memory-updating tasks and estimated their relationship to Gf and two aspects of WMC. Results from Bayesian SEM and mixed-effect models showed that individual differences in updating trials represent mainly WM maintenance ability, whereas updating-specific variance contributes substantially less to individual differences in updating tasks. Measuring credible updating-specific variance was challenging and required a modelling approach that separates out trial noise, as our Bayesian GLM did (Rouder & Haaf, 2019). However, even when measured credibly, the updating-specific variance was related neither to Gf nor to aspects of WMC (i.e., WM SP and WM RI). In contrast, individual differences in the WM maintenance component of the updating tasks were related to both Gf and WMC. This result challenges existing theories assuming a close relationship between EFs and higher cognitive abilities.

Previous work on the relationships among the three commonly distinguished executive functions – inhibition, shifting, and updating – indicates that there is shared variance among these EF that fully absorbs the inhibition factor but leaves some shifting-specific and updating-specific variance to be represented by separate factors (Friedman et al., 2008; Karr et al., 2018). This *common-EF* model could explain previous findings of larger relationships of updating with

Gf and WMC (Friedman et al., 2006; Wongupparaj et al., 2015) by assuming that individual differences in cognitive processes specific to updating are more relevant for Gf and WMC than individual differences captured in the *common-EF* factor. However, the *common-EF* model was developed from individual-differences studies in which updating was measured as overall accuracy in updating tasks. The present results show that this measure reflects predominantly WM maintenance. Therefore, we conclude that the updating-specific factor in the *common-EF* model probably reflects WM maintenance, and maintenance ability is the variance that is shared with Gf and WMC.

Updating Cannot Explain Why WMC and *Gf* Are Related

Contrary to theoretical accounts claiming that executive control explains why Gf and WMC are strongly related constructs (Engle, 2002; Kane & Engle, 2002; Shipstead et al., 2016), the present results add to recent studies showing no relationship of individual differences in the three commonly defined EF factors with Gf or WMC (Frischkorn et al., 2019; Rey-Mermet et al., 2019). Previous studies had consistently found updating to strongly relate to WMC and Gf, unlike shifting and inhibition (Friedman et al., 2006; Wongupparaj et al., 2015). Our study explains why: The use of average performance in updating tasks in these previous studies has conflated the contribution of general WMC, in particular maintenance ability, and updating-specific processes. Variance in updating-specific processes, however, contributes little to individual differences in overall performance in updating tasks. Even when using the best available statistical model to estimate variance in updating free from trial-to-trial noise (Rouder & Haaf, 2019), individual differences in neither Gf nor two other aspects of WMC were related to individual differences in the updating effect. This result also contradicts the specific prediction derived from the theory of Shipstead et al. (2016), which is that maintenance ability is more strongly related to WMC, whereas disengagement ability – represented here by differences in updating-specific processes – is more strongly related to Gf. We found that both WMC and Gf were related only to maintenance ability but not the executive-control component of updating task performance.

Taken together, the relationships of updating with Gf and WMC reported in previous studies were likely driven by variance in WM maintenance. In the present study, WM maintenance and WM SP were strongly related to each other and predicted Gf to a similar degree. This resonates with previous findings indicating that updating tasks and complex span tasks measure WMC to a similar extent (Schmiedek et al., 2009). Likewise, it matches previous results showing that primarily individual differences in short-term memory storage (e.g., encoding and maintaining information in WM) explain the association of WMC and Gf (Colom et al., 2005; Martínez et al., 2011). Other research converges with this conclusion by showing that memory maintenance is the only demand necessary for measuring WMC in a valid manner. In particular, WM measures do not need to require additional attentional regulation (e.g., the filtering of distractors in complex span tasks, or the substitution of information in updating or running span tasks). Measures only requiring WM maintenance are equally well-suited to measure WMC (Wilhelm et al., 2013). Therefore, the mechanisms and processes involved in the formation, maintenance, and retrieval of representations in WM seem to be more relevant regarding individual differences in WMC and Gf than executive processes. One candidate currently discussed in that regard is the ability to form and maintain bindings in WM (Oberauer, 2019).

Regarding the relationship of updating-specific processes with WM maintenance, some previous studies have already provided evidence suggesting that specifically the substitution of information in WM is not related to WMC (Ecker et al., 2010). The present study extended this result to Gf and WM RI. In contrast, Singh et al. (2018) found evidence that the efficiency of removal of outdated information from WM – measured by differences in response latencies to updating stimuli in different conditions – was related to both WMC and Gf (although the latter relation was fully mediated by WMC). Whereas this latency-based measure captured the time that individuals needed to carry out one updating step in WM, it did not capture the overall success of that process over several steps (i.e., final recall accuracy), which is the type of measure used in the present study. The updating efficiency measured by Singh and colleagues may thus represent other aspects of updating (e.g., speed of removing old information from WM) that we did not capture in our paradigm.

Isolating Cognitive Processes

A premise of the present research is that, to measure EF, we need to isolate the variance reflecting EF from variance of basic mechanisms and processes whose functioning is supervised by the EF in question. The most common way of achieving this is through a difference score contrasting two experimental conditions. One issue with isolating cognitive processes that has gained considerable traction, in particular in research on EFs, is that differences between experimental conditions tend to be unreliable (Enkavi et al., 2019; Hedge et al., 2018). Recently, some researchers have even proposed to avoid using difference scores as indicators for individual differences in cognitive processes in general, and instead use measures based on average performance in a single task condition (Draheim et al., 2019). This line of reasoning suggests that, instead of aligning the measurement of updating with that of shifting and inhibition by controlling for variance in basic information processing, we should instead develop average score measures for inhibition and shifting (Draheim et al., 2020) to overcome the so-called *reliability paradox*. Measures of EFs using average scores (e.g., accuracy in the anti-saccade task, or accuracy in WM updating tasks) are attractive because they have better reliability and stronger relationships with fluid intelligence and WMC compared to difference scores (Shipstead et al., 2014; von Bastian et al., 2020). Yet, we maintain that the sweeping dismissal of measures controlling for baseline information processing (i.e., difference scores, latent differences, or trial-noise controlled experimental effects) is not warranted. Although such experimental differences often showed poor reliability, this is not a statistical necessity, and it is not always the case in practice. For instance, with a sufficient number of trials, task-switch costs (von Bastian & Druey, 2017) and conflict costs in inhibition tasks (Rey-Mermet et al., 2018) can be measured with acceptable reliability.

In addition, conceptually, there are few alternatives that allow for isolating variation in a specific cognitive process. For tasks measuring EF, performance necessarily relies on two kinds of processes: (1) those that do the basic information-processing work, such as perceptual decision-making or memory maintenance, and (2) executive processes that control the basic processes and shield them against distraction. Therefore, individual differences in average performance (be it reaction times or accuracy) conflates variance in the success and efficiency of basic processes with variance in EF. Hence, researchers interested in individual differences in EF are left with two options: (a) using cognitive measurement models to separate basic and executive processes reflected in different parameters of the model (Frischkorn & Schubert, 2018), or (b) isolate the variance of executive processes through measures contrasting conditions with equivalent basic processes but different demands on EF (e.g., difference scores, latent differences, or experimental effects cleaned from trial-to-trial noise).

Lacking cognitive measurement models for the present tasks, we avoided the problem of unreliable differences with two statistical methods that isolate variations in updating-specific processes on a latent level. Although latent-change models estimated via BSEM were not able to capture credible variance in updating-specific processes, the BGLMs were able to isolate credible variations in performance decrements due to updating. As the BGLM separates true variance in the updating effect from trial-to-trial noise and task-specific variance, its estimate of the individual updating effect is error-free, analogous to a latent factor in an SEM. This approach circumvents the low-reliability problem. Nonetheless, updating-specific variance was related to neither Gf nor WMC in either BSEM or BGLM analysis. In sum, even when isolating only the reliable proportion of variance in updating-specific processes, there is no relation of updating with Gf or WMC.

Limitations of the current study

The sample size of the present study is low compared to other studies investigating individual differences in behavioral measures. Small sample sizes lead to considerable uncertainty in parameter estimates (Kretzschmar & Gignac, 2019; Schönbrodt & Perugini, 2013) as well as low power for detecting credible differences between statistical models. Regarding the first point, we report 95% credibility intervals that summarize the uncertainty in parameter estimates and allow for a more nuanced interpretation of the results than point estimates do. To address the second problem, we used BFs to compare BSEM and BGLM. Unlike non-significant *p*-values in frequentist model comparison tests, BFs quantify evidence in favor of one model over the other and indicate if there is insufficient evidence to accept either of the models. All BFs reported in the current study provide at least robust evidence (i.e., BFs > 3) for one of the BSEM or BGLMs.

Still, credibility intervals for parameter estimates were wider than desirable and, thus, the present results do not allow specific interpretations regarding the size of the investigated relationships. Rather, they indicate whether the data are better explained by assuming the presence or absence of a relationship between the different constructs. Nonetheless, this does not change the main takeaways from the present study: (1) Average performance in updating tasks predominantly reflects WM maintenance and only little to no updating-specific variance; and (2) there is no relationship of this updating-specific variance to any of the three covariates, even when using an elaborate statistical procedure to isolate credible updating specific variance. Given the strength of evidence for these two conclusions as quantified by the BFs, the present study was able to provide robust evidence despite the small sample size. Nevertheless, a replication of the present findings in future studies with larger sample sizes would be desirable.

A further limitation is that there was some heterogeneity in the breadth of operationalization for the different constructs. Specifically, the updating tasks, Gf, and WM RI measures tapped both verbal and figural domains using verbal, numerical, and figural material. The WM SP measures only tapped the verbal domain using numerical and verbal material. Therefore, the reported relationships of WM SP with Gf and WM RI might be underestimated due to a lack of representation of figural material for WM SP. Yet, WM SP still correlated strongly with the maintenance factor from the updating tasks that summarized variance from both content domains. Therefore, we think that this difference in the breadth of operationalization is not critical with respect to the interpretation of the results.

Conclusion

Previous studies suggesting a strong relationship of WM updating with Gf and WMC conflated variance of general WM ability with updating-specific variance and, thereby, overestimated the contribution of updating – or, in Shipstead et al.'s (2016) terminology, disengagement – to individual differences in Gf and WMC. Instead of updating-specific variance, average performance in updating tasks captures individual differences similar to WM

SP measures. Previous research has already established that two of the three established EF abilities – inhibition and shifting – share little, if any, variance with Gf (Friedman et al., 2006; Wongupparaj et al., 2015). Here we show that the third EF ability – updating – also fails to account for variance in Gf and two aspects of WMC.

References

Allom, V., & Mullan, B. (2014). Individual differences in executive function predict distinct eating behaviours. *Appetite*, *80*, 123–130. https://doi.org/10/f6czvh

Arthur, W., & Day, D. V. (1994). Development of a Short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement*, *54*(2), 394–403. https://doi.org/10/fgtkhd

Arthur, W., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-Sample Psychometric and Normative Data on a Short Form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment*, *17*(4), 354–361. https://doi.org/10/frmvvf

Barbey, A. K., Colom, R., Solomon, J., Krueger, F., Forbes, C., & Grafman, J. (2012). An integrative architecture for general intelligence and executive function revealed by lesion mapping. *Brain*, *135*(4), 1154–1164. https://doi.org/10/gfvn33

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10/gddxwp

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10/b2pm

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*(2), 163–183. https://doi.org/10/frs9t8

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10/ddq83h

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. https://doi.org/10/gbvchk

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42–100. https://doi.org/10/c2tc4x

Diamond, A. (2013). Executive Functions. *Annual Review of Psychology*, *64*(1), 135–168. https://doi.org/10/b2m2

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508–535. https://doi.org/10/ggc7kb

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2020). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General*. https://doi.org/10/gg9p63

Ecker, U. K. H., Lewandowsky, S., & Oberauer, K. (2014). Removal of information from working memory: A specific updating process. *Journal of Memory and Language*, *74*, 77–90. https://doi.org/10/ggrw2k

Ecker, U. K. H., Lewandowsky, S., Oberauer, K., & Chee, A. E. H. (2010). The components of working memory updating: An experimental decomposition and individual differences.

Journal of Experimental Psychology: Learning, Memory, and Cognition, *36*(1), 170–189. https://doi.org/10.1037/a0017891

Ecker, U. K. H., Oberauer, K., & Lewandowsky, S. (2014). Working memory updating involves item-specific removal. *Journal of Memory and Language, 74*, 1–15. https://doi.org/10/gd3vs9

Ekstrom, R. B., French, J. M., Harman, H. H., & Derman, D. (1976). *Manual for Kit of Factor-Referenced Cognitive Tests*. Educational Testing Service. https://www.ets.org/Media/Research/pdf/Manual_for_Kit_of_Factor-Referenced_Cognitive_Tests.pdf

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science*, *11*(1), 19–23. JSTOR. https://doi.org/10/b5qkt3

Enkavi, A. Z., Eisenberg, I. W., Bissett, P. G., Mazza, G. L., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences*, *116*(12), 5472–5477. https://doi.org/10/gfwtvc

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not All Executive Functions Are Related to Intelligence. *Psychological Science*, *17*(2), 172–179. https://doi.org/10/bmb68s

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, 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/b62mcp

Frischkorn, G. T., & Schubert, A.-L. (2018). Cognitive Models in Intelligence Research: Advantages and Recommendations for Their Application. *Journal of Intelligence*, *6*(3), 34. https://doi.org/10/gd3vqn

Frischkorn, G. T., Schubert, A.-L., & Hagemann, D. (2019). Processing speed, working memory, and executive functions: Independent or inter-related predictors of general intelligence. *Intelligence*, *75*, 95–110. https://doi.org/10/gf3sxs

Gronau, Q. F., Wagenmakers, E.-J., Heck, D. W., & Matzke, D. (2018). A Simple Method for Comparing Complex Models: Bayesian Model Comparison for Hierarchical Multinomial Processing Tree Models Using Warp-III Bridge Sampling. *Psychometrika*. https://doi.org/10/gft3ck

Hedden, T., & Yoon, C. (2006). Individual differences in executive processing predict susceptibility to interference in verbal working memory. *Neuropsychology*, *20*(5), 511–528. https://doi.org/10/bdgg7g

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10/gddfm4

Himi, S. A., Bühner, M., Schwaighofer, M., Klapetek, A., & Hilbert, S. (2019). Multitasking behavior and its related constructs: Executive functions, working memory capacity, relational integration, and divided attention. *Cognition*, *189*, 275–298. https://doi.org/10/gh3d69

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*(4), 637–671. https://doi.org/10/bwh9mt

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., & Garcia-Barrera, M. A. (2018). The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin*. https://doi.org/10/gd3vsx

Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A.-L., de Mooij, S. M. M., Moutoussis, M., Goodyer, I. M., Bullmore, E., Jones, P. B., Fonagy, P., Lindenberger, U., & Dolan, R. J. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, *33*, 99–117. https://doi.org/10/gfvqmd

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/bwtsjn

Könen, T., & Karbach, J. (2021). Analyzing Individual Differences in Intervention-Related Changes. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920979172. https://doi.org/10/gjf7zt

Kovacs, K., & Conway, A. R. A. (2016). Process Overlap Theory: A Unified Account of the General Factor of Intelligence. *Psychological Inquiry*, *27*(3), 151–177. https://doi.org/10/gd3vr6

Kretzschmar, A., & Gignac, G. E. (2019). At what sample size do latent variable correlations stabilize? *Journal of Research in Personality*, *80*, 17–22. https://doi.org/10/gfzhkx

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/bxmdv4

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10/cqwttd

Martínez, K., Burgaleta, M., Román, F. J., Escorial, S., Shih, P. C., Quiroga, M. Á., & Colom, R. (2011). Can fluid intelligence be reduced to 'simple' short-term storage? *Intelligence*, *39*(6), 473–480. https://doi.org/10/b9h36d

McArdle, J. J. (2009). Latent Variable Modeling of Differences and Changes with Longitudinal Data. *Annual Review of Psychology*, *60*(1), 577–605. https://doi.org/10/dhxt7h

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins, A. G. Sayer, L. M. Collins (Ed), & A. G. Sayer (Ed) (Eds.), *New methods for the analysis of change.* (2001-01077-005; pp. 139–175). American Psychological Association. https://doi.org/10.1037/10409-005

McNeish, D. (2016). On Using Bayesian Methods to Address Small Sample Problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

Meisel, S. N., Fosco, W. D., Hawk, L. W., & Colder, C. R. (2019). Mind the gap: A review and recommendations for statistically evaluating Dual Systems models of adolescent risk behavior. *Developmental Cognitive Neuroscience*, *39*, 100681. https://doi.org/10/ggdzcc

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian Structural Equation Models via Parameter Expansion. *Journal of Statistical Software*, *85*(1), 1–30. https://doi.org/10/gf7fkx

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10/fhpgvb

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/bkksp2

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10/f396t7

Oberauer, K. (2009). Design for a Working Memory. In *Psychology of Learning and Motivation* (Vol. 51, pp. 45–100). Academic Press. https://doi.org/10.1016/S0079-7421(09)51002-X

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—Facets of a cognitive ability construct. *Personality and Individual Differences*, *29*(6), 1017–1045. https://doi.org/10/btrs9h

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, *31*(2), 167–193. https://doi.org/10/fs4vfj

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(4), 501–526. https://doi.org/10/gcx8pf

Rey-Mermet, A., Gade, M., Souza, A. S., von Bastian, C. C., & Oberauer, K. (2019). Is executive control related to working memory capacity and fluid intelligence? *Journal of Experimental Psychology: General*, *148*(8), 1335–1372. https://doi.org/10/gfz43z

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452–467. https://doi.org/10/gfxsct

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/c3pt67

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10/f496x4

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/f9hdxt

Shipstead, Z., Lindsey, D. R. B., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language*, *72*, 116–141. https://doi.org/10/gd3vsp

Singh, K. A., Gignac, G. E., Brydges, C. R., & Ecker, U. K. H. (2018). Working memory capacity mediates the relationship between removal and fluid intelligence. *Journal of Memory and Language*, *101*, 18–36. https://doi.org/10/gfdbz5

Snyder, H. R., Miyake, A., & Hankin, B. L. (2015). Advancing understanding of executive function impairments and psychopathology: Bridging the gap between clinical and cognitive approaches. *Frontiers in Psychology*, *6*. https://doi.org/10/f66j67

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online*, *2*(1).

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10/b77m95

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—And a little bit more. *Intelligence*, *30*(3), 261–288. https://doi.org/10/bcsx4d

von Bastian, C. C., Blais, C., Brewer, G. A., Gyurkovics, M., Hedge, C., Kałamała, P., Meier, M. E., Oberauer, K., Rey-Mermet, A., Rouder, J. N., Souza, A. S., Bartsch, L. M., Conway, A. R. A., Draheim, C., Engle, R. W., Friedman, N. P., Frischkorn, G. T., Gustavson, D. E., Koch, I., … Wiemers, E. A. (2020). Advancing the understanding of individual differences in attentional control: Theoretical, methodological, and analytical considerations. *PsyArXiv*, 1–81.

von Bastian, C. C., & Druey, M. D. (2017). Shifting between mental sets: An individual differences approach to commonalities and differences of task switching components. *Journal of Experimental Psychology: General*, *146*(9), 1266–1285. https://doi.org/10/gchkd3

von Bastian, C. C., Locher, A., & Ruflin, M. (2013). Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods*, *45*(1), 108–115. https://doi.org/10/f4nn3j

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language*, *69*(1), 36–58. https://doi.org/10/gf88hv von Bastian, C. C., Souza, A. S., & Gade, M. (2016). No evidence for bilingual cognitive advantages: A test of four hypotheses. *Journal of Experimental Psychology: General*, *145*(2), 246–258. https://doi.org/10/f792cx

Wilhelm, O., Hildebrandt, A., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, *4*, 433. https://doi.org/10/gd3vs7

Wongupparaj, P., Kumari, V., & Morris, R. G. (2015). The relation between a multicomponent working memory and intelligence: The roles of central executive and short-term storage functions. *Intelligence*, *53*, 166–180. https://doi.org/10/gd3vsr

Appendix

Table A1.
Correlation matrix of the manifest indicators used for Bayesian structural equation models.

			Updating Tasks						Reasoning					WM SP		WM RI		
			No Updating			Updating												
			Color	Letter	Number	Color	Letter	Number	RAPM	Locat	Letter	Relat.	Syllog.	BP	CS	Verbal	Figural	Numeric
Updating	No Updating	Color		.16	.27	.64	.46	.46	.28	.23	.28	.15	.05	.37	.32	.08	-.02	.06
		Letter	.16		.20	.13	.31	.32	.08	.19	.16	.08	.22	.35	.28	.03	.00	.05
		Number	.27	.20		.26	.16	.63	.21	.07	.14	.01	-.05	.10	.15	.08	-.02	.30
	Updating	Color	.64	.13	.26		.52	.54	.25	.30	.14	.19	.11	.35	.48	.24	-.01	.17
		Letter	.46	.31	.16	.52		.49	.24	.20	.14	.15	.28	.49	.35	.32	.05	.19
		Number	.46	.32	.63	.54	.49		.31	.23	.11	.13	.15	.36	.26	.22	.06	.32
Reasoning		RAPM	.28	.08	.21	.25	.24	.31		.26	.47	.25	.22	.14	.07	.17	.05	.19
		Locat	.23	.19	.07	.30	.20	.23	.26		.37	.20	.30	.21	.23	.07	.06	.17
		Letter	.28	.16	.14	.14	.14	.11	.47	.37		.28	.20	.29	.24	.10	.11	.11
		Relat.	.15	.08	.01	.19	.15	.13	.25	.20	.28		.24	.21	.20	.08	.07	.15
		Syllog.	.05	.22	-.05	.11	.28	.15	.22	.30	.20	.24		.26	.18	-.06	.00	.07
WM SP		BP	.37	.35	.10	.35	.49	.36	.14	.21	.29	.21	.26		.49	.28	.21	.06
		CS	.32	.28	.15	.48	.35	.26	.07	.23	.24	.20	.18	.49		.08	-.10	.07
WM RI		Verbal	.08	.03	.08	.24	.32	.22	.17	.07	.10	.08	-.06	.28	.08		.28	.30
		Figural	-.02	.00	-.02	-.01	.05	.06	.05	.06	.11	.07	.00	.21	-.10	.28		.15
		Numerical	.06	.05	.30	.17	.19	.32	.19	.17	.11	.15	.07	.06	.07	.30	.15	

Note. WM SP = working memory storage & processing; WM RI = working memory relational integration; RAPM = Raven advanced progressive matrices; Locat = locations test; Letter = letter sets task; Relat = relations test; Syllog = syllogisms task; BP = Brown-Peterson task; CS = complex span task