

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

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

Gidon T. Frischkorn¹, Claudia C. von Bastian², Alessandra S. Souza¹ & Klaus Oberauer¹

¹University of Zurich, Switzerland

²University of Sheffield, United Kingdom

Word Count:

Abstract: 150; Manuscript: 5077; Total: 5227;

No. of Figures & Tables: 5 & 2

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.

23
24
25
26
27
28
29
30
31
32
33
34
35
36
37

Abstract

Updating is the executive function (EF) previously found to most strongly relate to higher cognitive abilities such as reasoning. However, this relationship could be a methodological artifact: Measures of other EFs (i.e., inhibition and shifting) usually isolate the contribution of EF, whereas updating is measured by overall accuracy in working memory (WM) tasks involving updating. This updating accuracy-score conflates updating-specific individual differences (e.g., removal of outdated information) with variance in WM maintenance. Re-analyzing data ($N = 111$) from von Bastian et al. (2016), we separated updating-specific variance from WM maintenance variance. Updating contributed only 15% to individual differences in performance in the updating tasks, and it correlated neither with reasoning nor with independent WM measures reflecting storage and processing or relational integration. In contrast, the WM 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;

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

40

41 Explaining individual differences in cognitive abilities, such as reasoning or working
 42 memory, by executive functions (EFs) has been popular for some time (Barbey et al., 2012;
 43 Engle, 2002; Kovacs & Conway, 2016). EFs – often conceptualized as attention regulation
 44 mechanisms – are thought to explain why working memory capacity (WMC) and fluid
 45 intelligence (gF) are strongly related constructs (Conway et al., 2002; Shipstead et al., 2016).
 46 Factor-analytic research on individual differences has yielded the distinction of three EFs:
 47 inhibition, shifting, and updating (Karr et al., 2018; Miyake et al., 2000). Inhibition refers to
 48 focusing attention on relevant information while suppressing information irrelevant for the
 49 current task. Shifting refers to flexibly switching between different tasks. Updating refers to
 50 replacing outdated information in working memory (WM) by new, more relevant information.

51 Previous research indicated stronger relationships of updating with gF than for inhibition
 52 and shifting (Friedman et al., 2006; Wongupparaj et al., 2015). However, there were important
 53 differences in the measurement of updating, shifting, and inhibition in these studies. Inhibition
 54 and shifting have been measured by difference scores between an experimental condition
 55 demanding the EF and a control condition demanding it less. These difference scores isolate the
 56 variance due to EF by controlling for confounding processes (e.g., stimulus encoding,
 57 processing, and motor response). By contrast, updating has been measured as the average
 58 performance in WM updating tasks. This average accuracy score conflates updating-specific
 59 variation with individual differences in general WM capacity, as measured by all short-term and
 60 working-memory tasks. The common demand of short-term and working memory tasks is to

61 maintain information over a few seconds, and this demand is the main limiting factor for
62 performance – when maintenance demands are reduced, performance is nearly perfect: Everyone
63 can remember 1 or 2 items, but performance decreases when memory load surpasses 4-5 items.
64 Therefore, researchers agree that this common source of variance reflects primarily maintenance
65 ability, and we will refer to it as *WM maintenance*.

66 The EF demands in updating tasks can be isolated by subtracting performance in a control
67 condition not involving updating from performance in an experimental condition requiring
68 updating. The resulting difference represents the ability to efficiently update information without
69 compromising memory performance. Thus, individuals with high updating abilities should show
70 smaller performance losses between the two conditions than individuals with low updating
71 abilities. Because previous studies did not isolate this updating-specific variance, they might
72 have overestimated the strength of the relationship of updating with gF.

73 The separation of updating from WM maintenance is also relevant from a theoretical
74 perspective: Shipstead et al. (2016) proposed that two different mechanisms contribute to
75 performance in both WM and gF tasks to different degrees: maintenance and disengagement.
76 According to this suggestion, solving reasoning problems, as used to measure gF, involves
77 mainly disengaging from no longer relevant information (e.g., incorrectly deducted or induced
78 rules) and, to a lesser degree, focusing and maintaining relevant information. In contrast, WM
79 measures such as complex span tasks tap mainly maintenance, relying on disengagement only
80 when it comes to avoiding distraction from secondary-task demands. Updating tasks may capture
81 both mechanisms equally: they require maintaining information while also disengaging from
82 outdated information in WM (Ecker et al., 2010). Yet, to investigate the relationships of the

83 maintenance and the disengagement component in updating tasks with *gF* and WMC, variance
84 capturing disengagement needs to be separated from variance due to maintenance.

85

86 **Updating-specific processes and their relation to WMC and *gF***

87 Updating tasks involve a combination of retrieving, transforming, and substituting or
88 removing information stored in WM (Ecker et al., 2010). For instance, in an arithmetic updating
89 task (Oberauer et al., 2000), each updating step involves retrieving one of the digits held in WM,
90 transforming it according to a given arithmetic operation (e.g., “+2”), and substituting the old
91 digit by the result. The most common tasks to assess updating – for instance the *N*-back
92 (Kirchner, 1958), keep-track (Miyake et al., 2000), or running span tasks (Friedman et al., 2006)
93 – require retrieval and substitution of information in WM but no transformation. Specifically,
94 these tasks require selectively accessing some information in WM and substituting it by new
95 information; hence the selective replacing of outdated information is the characteristic feature of
96 WM updating. Whereas the successful retrieval of stored information in these tasks depends
97 primarily on accurate maintenance, substitution requires disengagement from previously encoded
98 or transformed information.

99 Only few studies distinguished individual differences specific to these updating processes
100 and related them to other standard WMC measures. Ecker et al. (2010) found that the accuracy of
101 retrieval ($r = .55$) and transformation ($r = .49$) was positively correlated with WMC, but
102 substitution accuracy was not. Individual differences in the speed of updating processes were
103 unrelated to WMC. Similarly, Ecker et al. (2014) observed no correlation between removal
104 efficiency (i.e., the speed with which participants finish updating information in a self-paced
105 updating task) and WMC. However, in a more recent study, Singh et al. (2018) found that

106 removal efficiency was related to WMC. They also found that *gF* was related to removal
107 efficiency, but this relationship was fully mediated by WMC, speaking against the suggestion
108 that disengagement underlies the correlation between updating and *gF*.

109

110 **Present Study**

111 In the present study, we investigated the relationship of updating to *gF* and WMC by re-
112 analyzing data published by von Bastian, Souza, and Gade (2016). The updating tasks in this
113 dataset resemble commonly used keep-track tasks but contain trials with and without updating
114 demands. Thus, these tasks avoid conflating updating with maintenance and also address the
115 limitations of previous efficiency-based paradigms (Singh et al., 2018). By contrasting the
116 updating condition with a control condition requiring no updating at all – similar to inhibition
117 and shifting measures – we isolated updating-specific variance associated with disengagement
118 from variance related to maintenance. Crucially, we circumvent the frequently discussed
119 problem of insufficient reliability of difference scores by separating out trial-noise and isolating
120 only reliable individual differences in the updating effect (Rouder & Haaf, 2019). By isolating
121 updating as a control process separate from WM maintenance, the present study provides a more
122 direct assessment of the predictive power of updating for *gF* and WMC.

123 Furthermore, we differentiated between two aspects of WMC: storage and processing
124 (WM SP), and relational integration (WM RI). WM SP refers to maintaining the representations
125 of several memory items while processing distractors, and this is usually measured with complex
126 span or Brown-Peterson tasks – which are also the paradigms used in this study. WM RI refers to
127 building new relations between elements to create structural representations (Oberauer et al.,
128 2000, 2003). WM RI is usually measured with tasks in which participants have to monitor

129 ensembles of stimuli that change regularly and react when they form a specific constellation
130 (e.g., a square, a rhyme, or some match between several elements). The differentiation of WM
131 SP vs. RI allowed us to explore whether updating is related differently to these two aspects of
132 WMC.

133 **Methods**

134 **Participants**

135 Of the original sample (N = 121) collected by von Bastian et al. (2016), one participant had
 136 to be excluded due to an experimenter error. In addition, we discarded uni- and multivariate
 137 outliers identified by the Mahalanobis distance from the different measures. The present analyses
 138 are thus based on data from 111 participants (67 female, 44 male, $M_{age} = 24.28$, $SD_{age} = 3.71$).

139 **Measures**

140 We analyzed the tasks tapping updating, WM SP, WM RI, and reasoning ability used by
 141 von Bastian et al. (2016). Table 2 displays average performance and reliability estimates for the
 142 tasks tapping these constructs. The covariance matrix of all variables is available in the online

Table 2
 Average performance, descriptive statistics, and reliability estimates for the sample ($N = 111$) and all tasks and measures used in this study.

Construct	Task	Updating	<i>M</i>	<i>SD</i>	<i>Skew</i>	<i>Kurtosis</i>	Min	Max	Est. Rel. ^a
Updating	Figural	no	.70	.22	-.53	-.82	.20	1.00	.94
		yes	.59	.16	-.44	.04	.15	.93	.94
	Numerical	no	.91	.13	-1.54	1.54	.50	1.00	.92
		yes	.72	.19	-.54	-.60	.21	1.00	.95
	Verbal	no	.95	.08	-1.50	1.07	.72	1.00	.84
		yes	.72	.12	-.03	-.86	.47	.97	.90
WM Capacity	Brown-Peterson		.80	.12	-.61	-.18	.45	1.00	.95
	Complex Span		.57	.15	.05	-.87	.27	.88	.92
WM Monitoring	Figural		2.64	.37	-.45	.33	1.43	3.33	.40
	Numerical		2.85	.70	-.04	-.49	1.30	4.36	.70
	Verbal		2.75	.63	-.09	-.04	.80	4.02	.70
Reasoning	Diagramming relationships		.74	.14	-.27	-.58	.33	1.00	.61
	Letter Sets		.84	.14	-1.58	3.09	.27	1.00	.62
	Locations		.68	.18	-.57	-.43	.20	1.00	.64
	Nonsense Syllogisms		.69	.15	.01	-.61	.30	1.00	.41
	Raven's APM		.70	.21	-.49	-.59	.17	1.00	.66

Note. Performance was measured as proportion of correct responses, except for WM Monitoring tasks that used sensitivity (d'). WM = working memory; 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.

143 supplementary material: osf.io/zkd4c.

144 **Updating.** The three updating tasks (both, programs and more detailed information is
145 available online at <http://www.tatool-web.com/#/doc/lib-bat-uzh-ef-updating.html>) were similar
146 in design to the keep-track task used by Miyake et al. (2000). Participants had to remember an
147 initial set of items and subsequently update some of these items one by one, replacing them by
148 new stimuli. At the end of each trial, participants were asked to recall the most recent items.
149 Importantly, in some trials no updating occurred. In these trials, participants were prompted to
150 recall the items directly following their encoding, hence these trials only required WM storage of
151 the initial items.

152 The updating tasks used materials from three different content domains: figural, verbal,
153 and numerical. Figure 1A (see p. 10) illustrates the three tasks. In the *figural updating tasks*,
154 participants had to remember, update, and recall the colors of five different shapes. Each
155 updating step involved the presentation of one of the to-be-remembered shapes in a new color,
156 and participants had to update the color of the respective shape. Using the same procedure, the
157 *numerical updating tasks* used digits ranging from 1 to 9 in four different colors, and the *verbal*
158 *updating tasks* used consonants (except “Y”) presented in five different locations on the screen.
159 Thus, memory set size varied between 4 (numerical updating task) and 5 items (figural and
160 verbal updating tasks). In addition, the number of updating steps in the three different tasks
161 varied from 7 (numerical), through 9 (verbal), to 10 (figural). All tasks comprised 20 trials with
162 updating and 5 trials without updating which were randomly intermixed. Although there were
163 less trials without updating, reliability estimates (see Table 2, p. 8) suggest that performance
164 differences could still be measured adequately.

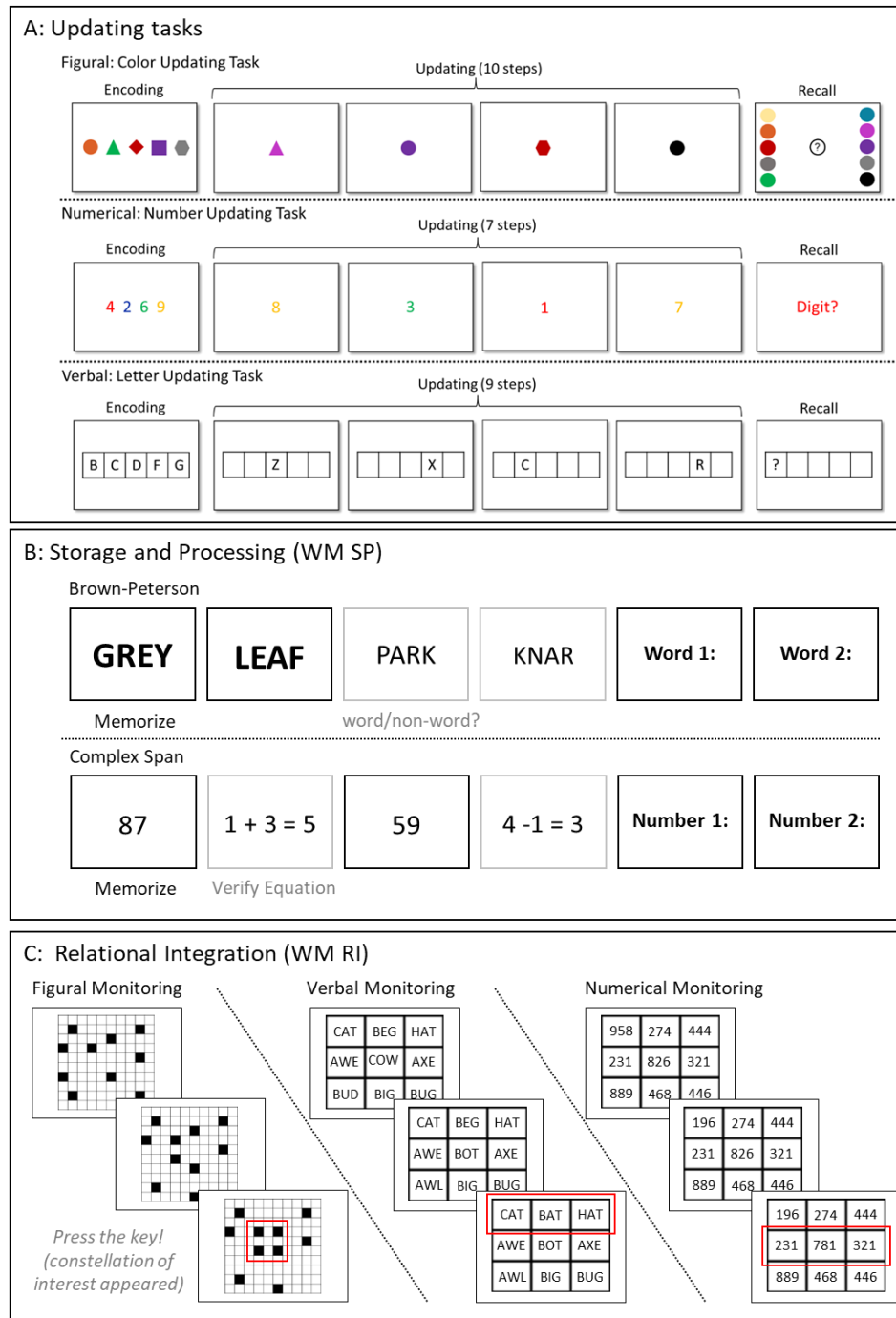


Figure 1. Illustration of the tasks used to measure Updating (A), WM storage & processing (B), and WM relational integration (C). In each of the *updating tasks* (A), participants initially encoded a memory set of 4 to 5 stimuli (colors, digits, or letters). Some trials required replacing one item at a time whenever a new stimulus was displayed, for 7, 9 or 10 updating steps, in the other trials the recall directly followed the 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 interleaving 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 changed sequentially. 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 – participant had to press the space bar. This is illustrated in the figure by the red frame around the relevant constellation.

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

170 **WM SP.** Individual differences in the ability to simultaneously store and process
171 information were measured with two tasks. In the *Brown-Peterson task* (see Figure 1B, p. 10),
172 participants first memorized 3-6 words and then performed five lexical decisions on four-
173 character strings. At the end of each trial, participants had to recall the words in correct serial
174 order. In the *complex span task* (see Figure 1B, p. 10), participants had to remember three to six
175 two-digit numbers while judging the correctness of a mathematical equation in between each of
176 the memoranda. At the end of each trial, participants had to recall the memoranda in correct
177 serial order.

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

182 **WM RI.** The ability to build new relations between multiple elements and integrate them
183 into structural representations was measured by three monitoring tasks (Oberauer et al., 2003;
184 von Bastian & Oberauer, 2013). In these tasks (see Figure 1C, p. 10), participants had to monitor
185 an array of stimuli, some of which were replaced every 2 s, and press the space bar whenever
186 they detected that a critical relation between a subset of the stimuli occurred. Again, the tasks
187 tapped into three different content domains with figural, verbal, and numerical material.

188 In the *figural monitoring tasks*, two of 20 dots changed their position in a 10x10 grid every
189 2 s, and participants had to monitor whether any four dots in the grid formed a square. In the
190 *verbal monitoring task*, 1 of 9 words in a 3x3 grid changed every 2 s, and participants had to
191 monitor whether three words in any direction across the grid (horizontal, vertical, or diagonal)
192 rhymed. In the *numerical monitoring task*, 1 of 9 three-digit numbers in a 3x3 grid changed
193 every 2 s, and participants had to monitor whether three numbers in any direction (horizontal,
194 vertical, or diagonal) had the last digit in common.

195 The performance measure in the monitoring task was the sensitivity d' of the detection
196 performance (i.e., $z(\text{Hits}) - z(\text{False Alarms})$). For participants with a perfect hit or false alarm
197 rate, the rates were corrected to a hit rate with $\frac{1}{2}$ miss and a false alarm rate of $\frac{1}{2}$ false alarm to
198 avoid $d' = \pm \text{Inf}$. Like WM SP measures, the WM RI measures were aggregated by a principal
199 component analysis for Bayesian hierarchical modeling.

200 **gF.** Participants' reasoning ability was assessed with five time-restricted tests. In the short
201 version of the *Raven's Advanced Progressive Matrices* (Arthur et al., 1999; Arthur & Day,
202 1994), participants had to complete a matrix pattern and choose the correct response from eight
203 alternatives. In the *Locations Test* (Ekstrom et al., 1976), participants had to select the correct
204 location of an "X" by identifying the patterns of "X" in four preceding rows of dashes. In the
205 *Letter Sets Test* (Ekstrom et al., 1976), participants had to select one letter set that deviated from
206 a logical pattern among a set of five letter sets. In the *Nonsensical Syllogisms Test* (Ekstrom et
207 al., 1976), participants had to decide whether conclusions drawn from two nonsensical premises
208 were logically valid. Finally, in the *Diagramming Relationships* (Ekstrom et al., 1976),
209 participants had to choose one out of five diagrams that best represented the set relations of three
210 nouns. For all reasoning tasks the performance measures were the proportion of correctly solved

211 items. Again, performance was aggregated by a principal component analysis over all tasks for
212 Bayesian hierarchical modeling.

213 **Statistical Analyses**

214 Raw data and scripts to preprocess and analyze the data can be accessed at: osf.io/zkd4c.

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

220 **SEM.** We estimated latent change models (Steyer et al., 1997) to isolate updating-specific
221 variance from variance of other WM processes with Bayesian SEMs (BSEM) using the package
222 *blavaan* (Merkle & Rosseel, 2018) implemented in *R* (R Core Team, 2018). The benefit of
223 Bayesian SEM is that in combination with adequate priors they provide better parameter
224 estimation in smaller samples (McNeish, 2016). Parameters were sampled using the *no U-turn*
225 sampler implemented in *STAN* (Carpenter et al., 2017) with four independent MCMC chains
226 that each consisted of 1000 warmup samples and 5000 samples after warmup. To check
227 convergence of the Bayesian parameter estimation, we required that the potential scale reduction
228 factor (PSRF) was below 1.05. The PSRF (a.k.a. \hat{R}) is the ratio of variance within each MCMC
229 chain to the variance between the different chains. PSRF values close to 1.00 indicate perfect
230 convergence, while larger values indicate insufficient convergence.

231 We judged absolute model fit of BSEM using the posterior predictive p -value (PP p). PP p -
232 values close to zero indicate a bad model fit, whereas values close to 0.5 indicate good model fit.
233 We follow the recommendations by Muthén and Asparouhov (2012) in requiring the estimated

234 BSEM to show at least PP $p > .05$ for the model to be retained for interpretation. For BSEM
235 model comparisons, we calculated Bayes factors to quantify the extent to which one BSEM is to
236 be favored over the other.

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

244 Bayesian hierarchical generalized linear mixed models (BGLM) were estimated using the
245 *brms* package (Bürkner, 2017). Specifically, parameters were estimated with four MCMC chains
246 each containing 1000 warmup samples and 10,000 samples after warmup. To ensure
247 convergence of the parameter estimation, we again checked that all PSRF values were below
248 1.05. As accuracy of each recall in the updating tasks follows a binomial distribution (0 =
249 incorrect, 1 = correct), we modeled recall performance in each trial with a binomial distribution
250 and a logit link function. In this model, the number of correctly recalled items in each trial in the
251 three updating tasks was predicted by the content domain of the tasks (i.e., figural, verbal,
252 numerical) and the updating factor (i.e., whether a trial contained updating or not).

253 In addition, separate models for each of the three covariates (i.e., WM SP, WM RI, and gF)
254 were estimated to quantify the relation of each covariate with overall accuracy (i.e. the intercept)
255 and, more importantly, the interaction of each covariate with updating (i.e., the slope describing
256 the updating effect). This cross-level interaction captures to what extent each covariate predicts

257 individual differences in the updating effect. To test whether this interaction was credible, we
258 first evaluated whether the 95% credibility interval (CI) of the posterior of the interaction
259 included zero. In addition, we compared a model including the interaction to a model not
260 including the interaction with the covariate, and quantified evidence for or against either of the
261 two models with Bayes Factors (BF) and posterior probabilities (PP) of the two models estimated
262 via bridge sampling (Gronau et al., 2018). To establish the robustness of the BF and the PP
263 estimation we estimated both models and BFs 10 times. In the results we report the smallest BF
264 or PP, so that the values estimate the lower limit for the estimation of the evidence for one or the
265 other model.

266

Results

267 **What is measured by updating tasks?**

268 First, we decomposed the common variance of the three updating tasks into two
 269 components of variance: (a) individual differences in WM maintenance vs. (b) individual
 270 differences related to updating-specific variance. Figure 2 depicts the structure of the latent
 271 change model (Steyer et al., 1997) we used to this end, and the parameters estimated with a
 272 Bayesian structural equation model (BSEM). WM maintenance is captured in a *Maintenance*
 273 factor estimated from trials without updating demands. The *Maintenance + Updating* factor
 274 captures variance in trials with updating demands. By regressing the *Maintenance + Updating*
 275 factor variance on the *Maintenance* factor variance, the residual *Updating* variable reflects
 276 individual variation in updating-specific processes.

277 The model fitted the data well, with a posterior predictive (PP) $p = .654$, and convergence
 278 of the parameter estimation was good for all parameter estimates, PSRFs > 1.01 . The results
 279 suggest that there are considerably smaller individual differences in updating specific processes

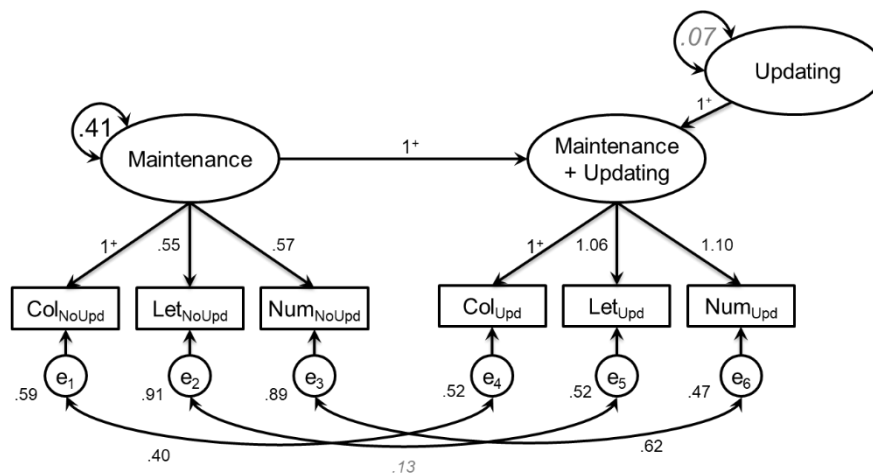


Figure 2. Graphical illustration of the latent change model that separates individual differences in basic memory processes from individual differences in updating-specific processes. Values for parameters refer to the posterior mean of the posterior distribution of parameters. Parameters printed in gray and in italics had 95% credibility intervals including zero. Variances and factor loadings are given as unstandardized parameters. + = Parameter was fixed to the depicted value. Col = Color, Let = Letter, Num = Number, NoUpd = no updating, Upd = updating.

280 ($\sigma_{\text{Upd}} = .07$, 95% CI = [.00, .21]) compared to WM maintenance ($\sigma_{\text{Main}} = .41$, 95% CI = [.19,
281 .70]), $\Delta = .34$, 95% CI = [.07, .61]. Given these estimates, approximately 85% of individual
282 differences in trials with updating across the three tasks (i.e. the *Maintenance* + *Updating* factor)
283 is accounted for by individual differences in general WM capacity, whereas only 15% of
284 variance can be attributed to individual differences in updating-specific processes. Regarding the
285 observed accuracies, individual differences in updating accounted for about 7% of variance in
286 performance in trials with updating, whereas individual differences in WM maintenance
287 accounted for 41 to 48% of variance.

288 Noteworthy, the 95% credibility interval of the updating-specific variance included zero,
289 implying that there might be no true variance in updating across the three different tasks.
290 However, because we explicitly aimed at investigating the relationship of updating with other
291 variables, we did not fix this variance to zero, in order to still be able to estimate relationships of
292 updating with the three covariates.

293

294 **Relationship of Updating with Reasoning and WMC**

295 The main question of this study is whether WM maintenance or updating-specific
296 processes are related to the three covariates: gF, WM SP, and WM RI. To address this question,
297 we estimated four separate BSEM that included the three covariates into the latent change model
298 for the updating tasks. Model I freely estimated the relationship between the *Maintenance* factor,
299 the *Updating* factor and all covariates. Model II fixed the relationship between the updating
300 factor and the covariates to zero. Model III conversely fixed the relationship between the
301 maintenance factor and all covariates to zero. Finally, Model IV fixed the relationship between
302 both the maintenance factor and the updating factor and the covariates to zero.

Table 1

Summary of Model Fit indices for the BSEM estimating the relationship between the memory and updating factor with the three covariates.

Model	Main. - Cov	Upd -Cov	PP <i>p</i>	BF
I	free	free	.311	121.5
II	free	0	.252	
III	0	free	.022	9.8 x 10 ⁵
IV	0	0	.002	1.8 x 10 ⁷

Note. Main. = maintenance, Cov = covariates, Upd = updating, PP *p* = posterior predictive *p*-value, BF = Bayes Factor.

Bayes Factors are computed in comparison with the best fitting model that is highlighted in bold.

303 Table 1 summarizes the absolute and relative model fit of these four models. The
 304 comparison of the four models via Bayes factors suggested that Model II, estimating only
 305 covariance between the *Maintenance* factor and the three covariates, provides the best and most
 306 parsimonious description of the observed covariance structure. Thus, Model II (see Figure 3) was
 307 retained for interpretation. Unsurprisingly, the factor capturing WM maintenance in the updating

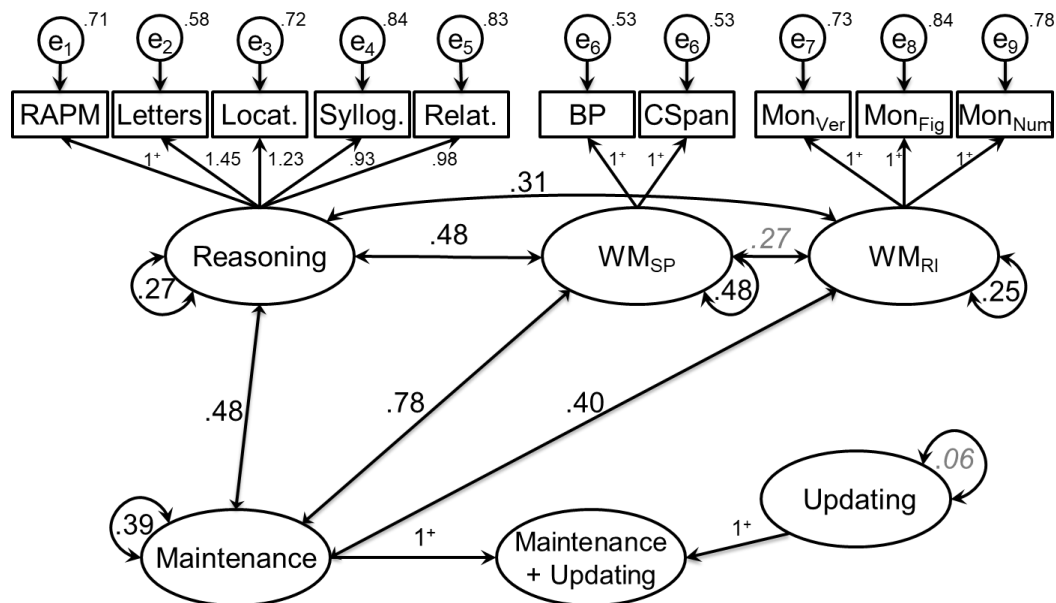


Figure 3. Graphical illustration of BSEM II only estimating the correlation between individual differences in WM maintenance and the covariates. Parameter values refer to the posterior mean. Parameters printed in gray and italics had 95% credibility intervals that included zero. All factor loadings and variances are reported as unstandardized parameters, except for correlations that are standardized. + = Parameter was fixed to the depicted value. 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.

308 tasks showed the largest correlation with WM SP, $r = .78$; correlations with gF , $r = .48$, and
 309 WM RI, $r = .40$, were still substantial. This implies that updating tasks capture, to a large extent,
 310 individual differences shared with tasks tapping WM SP.

311

312 **Alternative Analysis: Bayesian hierarchical generalized linear mixed models**

313 The BGLM results captured the experimental effects across the three updating tasks (i.e.,
 314 accuracy was lower in trials with updating than without updating), and the variation reflecting
 315 individual differences in overall accuracy and in the updating effect across the three tasks (see
 316 supplementary online material). Unlike the BSEMs, the BGLM showed credible variability
 317 across individuals in the updating effect, $\sigma_{\text{Upd}} = 0.35$ (95% CI = [0.26; 0.44]; see Figure 4). This
 318 corresponds to about 6.2% (95% CI = [3.6; 9.1]) of variance in observed accuracies. In contrast,
 319 variation in overall performance (i.e., the intercept) captured about 38.1% (95% CI = [29.7;
 320 46.6]) of variance in observed accuracies. Hence by modeling trial-by-trial data, and thereby
 321 isolating trial noise, the BGLM captured true individual differences in updating.

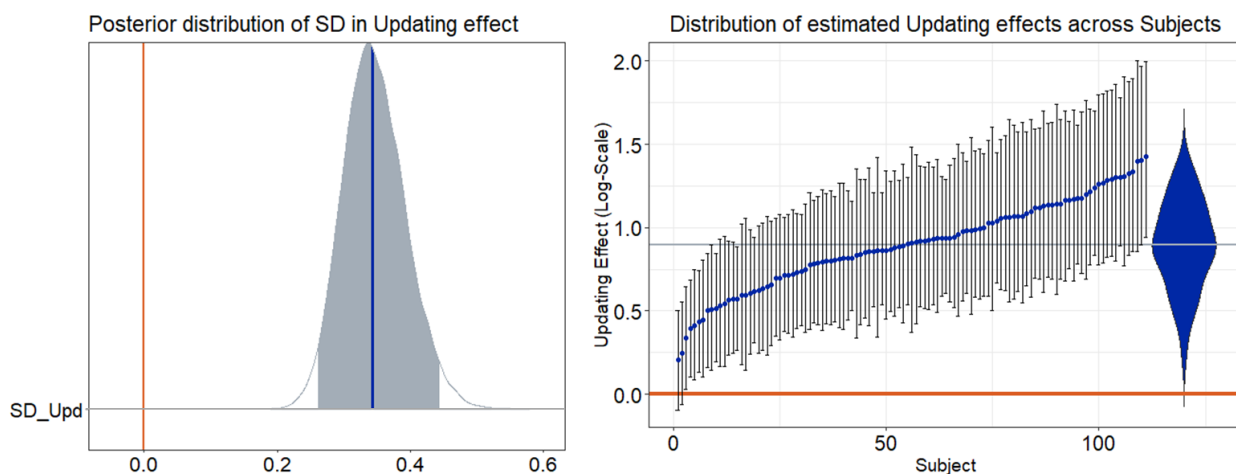


Figure 4. Posterior distribution of estimated variance in the updating effect (left side) and distribution of updating effects across all subjects (right side). The individual effects displayed on the right refer to the individual difference in performance (on the log-scale) between trials with and without 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.

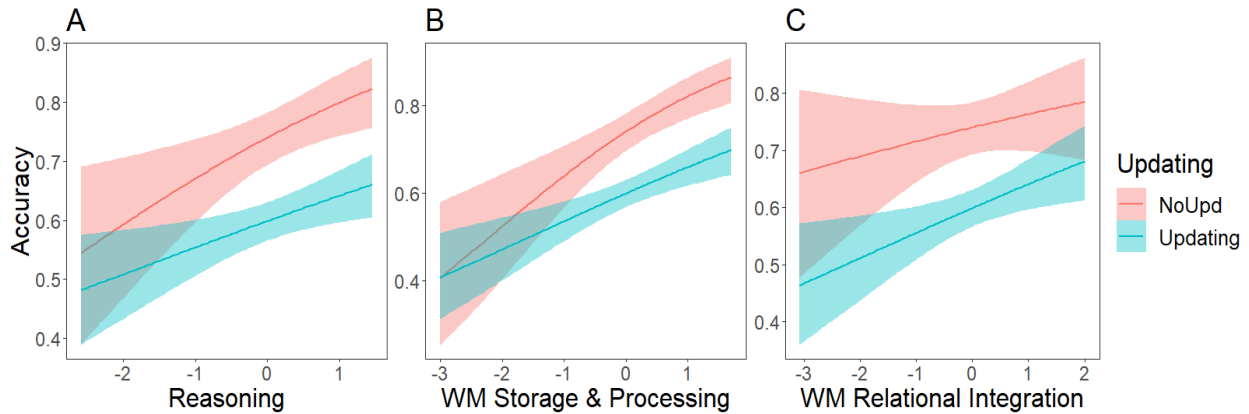


Figure 5. Illustration of the prediction of overall accuracy for trials with and without updating in the three BGLM including (A) reasoning ability, (B) WM storage & processing, and (C) WM relational integration as predictor. The shaded red and blue area around the regression lines indicates the 95% 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 somewhat curved on the accuracy scale.

322 **Relationship of Updating with the covariates.** To test whether any of the three covariates
 323 – gF, WM SP, or WM RI – was related to individual differences in the updating effect, we
 324 estimated BGLMs for each of the three covariates, each including the effects of task (figural,
 325 numerical, verbal), updating (trials with vs. without updating demands), and one of the three
 326 covariates as well as all interactions between the three effects. Figure 5 illustrates the results.

327 *BGLM: Updating and gF.* As illustrated in Figure 5A, including gF as predictor for
 328 accuracy across the three tasks and trials with and without updating showed that people with
 329 higher gF had higher accuracy in the updating tasks, $\beta = 0.31$ (95% CI = [0.15; 0.47]). However,
 330 there was no credible evidence that gF was related to variations in the updating effect, $\beta = 0.06$
 331 (95% CI = [-0.03; 0.15]). Thus, we compared the full model to a model without the interaction of
 332 gF and updating. The Bayes factor as well as posterior model probabilities (PP) indicated that the
 333 no-interaction model is more likely than the full model, $BF > 8.9 \times 10^5$; $PP_{full} < .01$;

334 $PP_{\text{no-interaction}} > .99$.¹ If anything, the interaction effect suggests that participants with lower gF
335 showed smaller decreases in performance in updating trials compared to no-updating trials.

336 *BGLM: Updating and WM SP.* As shown in Figure 5B, people with higher WM SP scores
337 had higher overall accuracy in the updating tasks, $\beta = .42$ (95% CI = [0.27; 0.57]). However,
338 again there was no credible evidence that WM SP predicted variations in the updating effect, $\beta =$
339 0.08 (95% CI = [-0.01; 0.16]). Although close to credibility, this effect implied that, if anything,
340 participants with lower WM SP ability showed smaller deteriorations in performance in updating
341 trials compared to no-updating trials. The Bayes factor as well as PP indicated that a model
342 without the interaction was more likely than the model including the interaction, $BF > 2.3 \times 10^3$;
343 $PP_{\text{full}} < .01$; $PP_{\text{no-interaction}} > .99$.

344 *BGLM: Updating and WM RI.* For an illustration of the relationships of WM RI with
345 performance in the updating tasks see Figure 5C. Similar to the other covariates, people better in
346 WM RI scores had higher overall accuracy in the updating tasks, $\beta = 0.18$ (95% CI = [0.02;
347 0.35]). But WM RI also did not credibly predict variability in the updating effect, $\beta = -0.03$ (95%
348 CI = [-0.12; 0.05]). Again, a model without the interaction was clearly favored over the model
349 including the interaction $BF > 3.4 \times 10^5$; $PP_{\text{full}} < .01$; $PP_{\text{no-interaction}} > .99$.

¹ To establish the robustness of the BF and the PP estimation we estimated both models and BFs/PPs 10 times. We report the smallest BF or PP, so that the values estimate the lower limit for the estimation of the evidence for one or the other model. See Methods for further details.

350

Discussion

351 We isolated individual differences in updating-specific processes in three commonly used
352 memory updating tasks and estimated their relationship to gF and two aspects of WMC. Results
353 from Bayesian SEM and mixed-effect models showed that individual differences in updating
354 trials represent mainly WM maintenance ability, whereas updating-specific variance contributes
355 substantially less to individual differences in updating tasks. The credible measurement of this
356 updating-specific variance was challenging, requiring a modelling approach that was capable of
357 parceling out trial noise. However, even when measured credibly, the updating-specific variance
358 was related neither to gF nor to aspects of WMC (i.e., WM SP and WM RI). In contrast,
359 individual differences in the WM maintenance component of the updating tasks were related to
360 both gF and WMC. This result challenges existing theories about the relationship between EFs
361 and higher cognitive abilities.

362

Updating cannot explain why WM and gF are related

364 Contrary to theoretical accounts claiming that executive attention explains why gF and
365 WMC are strongly related constructs (Engle, 2002; Shipstead et al., 2016), the present results
366 add to recent studies showing no relationship of individual differences in the three commonly
367 defined EF facets with gF or WMC (Frischkorn et al., 2019; Rey-Mermet et al., 2019). Previous
368 studies had consistently found updating to strongly relate to WMC and gF, unlike the EF facets
369 of shifting and inhibition (Friedman et al., 2006; Wongupparaj et al., 2015). Our study explains
370 why: The use of average performance in updating tasks in previous studies has conflated the
371 contribution of general WM capacity (i.e., maintenance ability, and perhaps other variables) and
372 updating-specific processes. Variance in updating-specific processes, however, contributes little

373 to individual differences in overall performance in updating tasks. Even when using the best
374 available statistical model to estimate variance in updating free from trial-to-trial noise (Rouder
375 & Haaf, 2019), individual differences in neither gF nor two other aspects of WMC were related
376 to individual differences in the updating effect. Therefore, the relationships reported in previous
377 studies were likely driven by variance in WM maintenance. WM maintenance and WM SP were
378 strongly related to each other and predicted gF to a similar degree in the present study. This
379 resonates with previous findings indicating that updating tasks and complex span tasks measure
380 WMC similarly (Schmiedek et al., 2009).

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

392

393 **Isolating cognitive processes: To subtract or not to subtract?**

394 One issue with isolating cognitive processes that has gained considerable traction is that
395 differences between experimental conditions tend to be unreliable (Hedge et al., 2018). Recently,

396 some researchers have even proposed to avoid using difference scores as indicators for
397 individual differences in cognitive processes in general (Draheim et al., 2019). We maintain that
398 this sweeping dismissal of difference scores is not warranted. Although difference scores often
399 showed poor reliability, this is not a statistical necessity, and it is not always the case in practice.
400 For instance, with a sufficient number of trials, task-switch costs (von Bastian & Druet, 2017)
401 and conflict costs in inhibition tasks (Rey-Mermet et al., 2018) can be measured with acceptable
402 reliability.

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

414 Lacking cognitive measurement models for the present tasks, we avoided the problem of
415 unreliable differences with two statistical methods that isolate variations in updating-specific
416 processes on a latent level. Although latent-change models estimated via BSEM were not able to
417 capture credible variance in updating-specific processes, BGLMs were able to isolate credible
418 variations in performance decrements due to updating. As the BGLM separates true variance in

419 the updating effect from trial-to-trial noise and task-specific variance, its estimate of the
420 individual updating effect is error-free, analogous to a latent factor in an SEM. This approach
421 circumvents the low-reliability problem. Nonetheless, updating-specific variance was related to
422 neither *gF* nor WMC in either BSEM or BGLM analysis. In sum, even when isolating only the
423 reliable proportion of variance in updating-specific processes, there is no relation of updating
424 with *gF* or WMC.

425

426 **Conclusion**

427 Previous studies suggesting a strong relationship of WM updating with *gF* and WMC
428 conflated variance of general WM ability with updating-specific variance and, thereby,
429 overestimated the contribution of updating – or, in Shipstead et al.’s (2016) terminology,
430 disengagement – to individual differences in *gF* and WMC. Instead of updating-specific
431 variance, average performance in updating tasks captures individual differences similar to WM
432 SP measures. Previous research has already established that two of the three established EF
433 abilities – inhibition and shifting – share little, if any, variance with fluid intelligence (Friedman
434 et al., 2006; Wongupparaj et al., 2015). Here we show that the third EF ability – updating – also
435 fails to account for variance in *gF*.

References

- 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., Theriault, 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>
- 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>
- 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/dn563p>
- 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>
- 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>

- 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>
- 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>
- 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>
- 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>
- 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>
- 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/gf794c>
- 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>
- 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., 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>
- 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>
- 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>
- Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online*, *2*(1).
- 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., & 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>
- 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>