Psychonomic Bulletin & Review (2021) 28:1657–1667
https://doi.org/10.3758/s13423-021-01930-z

BRIEF REPORT

Is there more to metamemory? An argument for two specialized monitoring abilities

Ian M. McDonough[1] · Tasnuva Enam[1] · Kyle R. Kraemer[1,2] · Deborah K. Eakin[3] · Minjung Kim[4]

Accepted: 9 April 2021 / Published online: 4 May 2021

Abstract

Metamemory is the process of monitoring and controlling one's beliefs, knowledge, and mental processes of memory. One fundamental question is whether the monitoring component of this theory should be considered as only one ability or an umbrella of more specialized abilities. In the current study, we aimed to understand the structure of metamemory monitoring by testing unitary versus specialized measurement models of metamemory. Monitoring accuracy and mean ratings from four common monitoring judgments across different stimulus presentation pairs were calculated to create latent factors for each judgment using structural equation modeling. Our results suggest that although each of the monitoring judgments was correlated with one another, monitoring may be composed of two distinct abilities: one occurring during initial presentation and one occurring at retrieval. These results can help explain prior behavioral and brain dissociations between predictions at encoding and retrieval in terms of experimental and material manipulations. We caution against the conceptualization and use of metamemory monitoring as a unitary construct.

Keywords Episodic memory · Metamemory · Monitoring · Structural equation modeling

Metamemory is the process of monitoring one's beliefs, knowledge, and mental processes to understand how one's memory operates (Nelson & Narens, 1990). This ability can be considered a decision-making process that integrates information from multiple cognitive systems (e.g., Chua et al., 2014) and may be essential for understanding the nature of being human (Metcalfe, 2000). Thus, metamemory is highly relevant to our daily lives, especially for people focused on the efficient use of memory, including students, educators, aging adults, and those with neurocognitive disorders. Despite the importance of metamemory, its mental structure remains vague, thereby hindering our understanding of why certain people have better metamemory than others, how it develops and declines over time, and how it is affected by disease.

One fundamental barrier to understanding metamemory might be that the conceptualization of metamemory has been oversimplified. Here, we ask whether the monitoring component of metamemory should be considered only one ability or an umbrella of multiple specialized abilities. Better understanding this distinction can help improve self-regulation strategies involved in effective assessment of one's learning state and improve learning. If monitoring is comprised of multiple abilities, then educators can effectively leverage their limited resources to focus on improving the specific ability that is lacking.

As a starting point, we sought to determine whether different types of monitoring judgments might reveal a unitary monitoring ability or separable abilities. Monitoring judgments can be made at various points of learning and memory (Nelson & Narens, 1990). Two primary points include initial acquisition or encoding via ease of learning (EOL) and judgments of learning (JOL), and at retention or retrieval via feelings of knowing (FOK) and confidence judgments (CFJ). Each judgment might fall under the umbrella term of "monitoring" to the extent that making the judgment requires people

✉ Ian M. McDonough
immcdonough@ua.edu

[1] Department of Psychology, The University of Alabama, BOX 870348, Tuscaloosa, AL 35487, USA

[2] Department of Psychology, Birmingham Southern College, 900 Arkadelphia Road, Box 549022, Birmingham, AL 35254, USA

[3] Department of Psychology, Mississippi State University, PO Box 6161, Mississippi State, MS 39762, USA

[4] Department of Educational Studies, The Ohio State University, 29 W Woodruff Ave, Columbus, OH 43210, USA

to inspect the contents of their memory and then predict some outcome of the current or future memory state (Hertzog et al., 1990). However, while encoding-based judgments require people to predict future memory accuracy upon initially encountering a stimulus, retrieval-based judgments are made following a retrieval attempt. Nelson and Narens (1990) proposed a "meta-level" structure through which both monitoring and control processes operated and specified that monitoring and control occurred both at encoding and at retrieval. Although they did not explicitly state whether the disparate monitoring judgments represented a unitary process per se, the "meta-level" proposed in their framework leaves open this possibility. Researchers often imply a single unitary process by stating that different types of monitoring judgments "reveal aspects of the monitoring process" (Schwartz & Bacon, 2008, p. 356). Indeed, since their framework was introduced, the question of a unitary or specialized monitoring process has continued to be proposed (Chua et al., 2014; Kelemen et al., 2000; Mazancieux et al., 2020; Schraw et al., 1995).

Individual differences in monitoring ability are helpful to understand similarities and differences across the various judgments. A unitary monitoring perspective predicts that monitoring judgments should exhibit similar patterns of individual differences when using measures at different stages of learning and memory. Evidence for this idea comes from significant correlations among differing magnitudes of monitoring judgments (bias; Kelemen et al., 2000) and metamemory accuracy (Leonesio & Nelson, 1990). Monitoring judgments also are highly correlated among intelligence-based tasks, regardless of performance (Schraw et al., 1995). Even more recently, a high degree of shared variance was found for CFJ magnitude ratings and accuracy across multiple cognitive domains, including episodic memory (Mazancieux et al., 2020). Those authors concluded that metacognitive monitoring consisted of a common algorithm that was engaged across tasks.

Other evidence investigating individual differences suggests that monitoring consists of multiple abilities, either distinct or overlapping. Ironically, the same evidence used to promote a unitary monitoring perspective also can be used to argue against it. For example, although Leonesio and Nelson (1990) found significant correlations between multiple monitoring judgments, the effect sizes were weak (*r*s between .12 and .19), providing evidence against a unitary ability. Similarly, Kelemen et al. (2000) found correlations among multiple encoding-based judgments that ranged from −.01 to .70. However, they did not include retrieval-based monitoring judgments, leaving open a critical test of this account. Lastly, Mazancieux et al. (2020) found that the shared variance that accounted for monitoring ability varied across cognitive domain (7%–48%), suggesting that the remaining variance may be due to domain-specific processes.

From another perspective, dissociations between encoding-based judgments and retrieval-based judgments have been found such as across age groups (Eakin & Hertzog, 2012; Hertzog et al., 1990; Hertzog & Dunlosky, 2011). Encoding-based judgments often are no different between younger and older adults (Hertzog & Dunlosky, 2011), whereas retrieval-based judgments sometimes show stark age differences (e.g., Dodson et al., 2007; Kelley & Sahakyan, 2003). Although some of these differences likely stem from different types of information available during each stage or one's subjective theories (Koriat, 1997), the processes involved in monitoring might also be sufficiently different to lead to these differential effects.

Although the current evidence might appear to favor a multiple-ability view, recent neuroscience work has reinvigorated this question. To the extent that different monitoring judgments converge on the same brain region, variations in monitoring might be explained by differential integrity of that brain region. In reviews of neuroscience approaches, researchers have proposed that both encoding-based and retrieval-based monitoring largely rely on the prefrontal cortex (PFC; Fleming & Dolan, 2012; Pannu & Kaszniak, 2005; Schwartz & Bacon, 2008).

Although evidence fuels both sides of the debate, few studies have measured multiple monitoring judgments within the same participant and within the same paradigm to allow a direct test of unitary or multiple abilities. Moreover, previous studies relying on zero-order correlations do not clearly adjudicate between these different hypotheses because those methods do not indicate how strong a correlation must be to provide evidence for each perspective. We tested these opposing perspectives by measuring the bias and accuracy from four common monitoring judgments (EOLs, JOLs, FOKs, and CFJs) across different stimulus presentation pairs (word–word, picture–picture or picture–word) to create latent factors for each monitoring judgment in structural equation models (SEM). This approach allows an explicit test of various configurations of metamemory monitoring structure not afforded by previous studies.

Method

Participants

Of 342 participants collected from the Introductory Psychology Subject Pool in two large southern public universities, the final sample consisted of 329 participants (M_{age}= 19.40 years; 84.7% female; 95.3% non-Hispanic; 79.2% White, 15.1% African American, 1.9% Asian, and 2.2% mixed race; $M_{household\ income}$ = $109,960; $M_{parental\ education}$ = 15.13 years). Participants were excluded due to memory performance below chance (<20%) averaging across all blocks or

for failure to vary their judgments in at least one condition. All participants were compensated with course credits. All participants provided written consent as approved by the University of Alabama Institutional Review Board. The study was conducted in accordance with the Declaration of Helsinki.

Materials

A total of 240 word pairs were acquired from the University of South Florida word association norms (Nelson et al., 2004). These word pairs were randomly assigned to one of 12 lists used to generate the study stimuli (20 unique pairs each). Pictures that represented the words were gathered from multiple sources (Brady et al., 2008; Gonsalves & Paller, 2000) with the requirement that the pictures be clear, in color, of single objects, and on a white background. Four lists were composed of word–word pairs, four of picture–picture pairs, and four of picture–word pairs. Lists were tested for association characteristics using the word association norm program, ListChecker Pro 1.2 (Eakin, 2010). Each word was individually equated on set size (strength and number of connections with associated words), concreteness (the degree that the concept denoted by a word refers to a perceptible entity), connectivity (connections among association-set words) and resonance probability (connections from the association-set words back to the word itself). All item pairs were unrelated to each other and unrelated to any other item within their list.

The paired-associates task was divided into 12 blocks corresponding to the 12 stimuli lists mentioned above. Each block assessed one judgment type throughout all pairs within the block, and each participant completed three blocks for each judgment (JOLs, EOLs, FOKs, or CFJs). This fully crossed, within-participant design ensured that each participant made a single metamemory judgment for a single stimulus pair type in given block.

After recall, each block included a five-alternative forced-choice (5-AFC) recognition test phase for each of the 20 pairs. In this test, a cue (word or picture) was presented along with one target and four lure words. To minimize differences in familiarity, lures were constructed using targets from previously seen pairs within the block and never across the blocks. Monitoring judgments were counterbalanced across stimuli lists so that each pair (e.g., BRIDGE–KNIGHT) received a different judgment across participants. All the pairs were randomized within each block, and all blocks were randomized across trials. The correspondence between the monitoring judgments and recognition accuracy served as the dependent variable for the primary SEM analyses. Magnitude judgments served as the dependent variable for the secondary SEM analyses (see Supplemental Material).

Procedure

Each block consisted of a learning phase, a cued recall phase, a distraction phase, and a recognition memory phase (see Fig. 1). First, participants viewed a list of 20 item pairs in randomized blocks—either all word–word pairs, all picture–picture pairs, or all picture–word pairs for each block. During the learning phase, each stimulus pair was presented for 2 seconds. After viewing a pair, one of three procedures occurred depending on the block. For JOL blocks, participants were asked to provide their JOL judgment for the pair, "*How likely will you correctly recognize this item pair on a later test?"* on a scale from 0–100, where 0 indicated "definitely will not" and 100 indicated "definitely will." For EOL blocks, participants were asked to provide their EOL judgment for the pair, *"Compared with other pairs, how easy will it be to learn this item pair?"* on a scale from 0–100, where 0 indicated "most difficult" and 100 indicated "most easy." Both JOLs and EOL judgments are encoding-based judgments. In the retrieval-based blocks (FOK and CFJ), no encoding-based judgment was given. Rather, after viewing the pair, the screen moved on to the cued recall phase.

In the cued recall phase, participants saw the cue on the screen and they were asked to recall the second word when cued by the first word by typing in the answer within 10 seconds. If participants were in the FOK block, after each cued recall attempt, participants were asked to provide their FOK judgment for the pair, *"How likely will you recognize this item pair on a later test?"* on a scale from 0–100, where 0 indicated "least likely" and 100 indicated "most likely." These judgments were self-paced. No monitoring judgments were given at this point if participants were in the JOL, EOL, or CFJ block. After the cued recall phase, participants were asked to solve one multiplication question with a three-digit answer as a distractor task to reduce potential carryover effects from cued recall to the 5-AFC recognition phase. The distractor task was self-paced, and participants had to type in their answer and press "Next" on the screen to move on to the next phase.

In the 5-AFC recognition phase, participants were shown the cue, along with five options (one target intermixed with four lures) from which they were asked to choose the correct target to complete the pair. The 5-AFC test was also self-paced. For CFJ blocks, participants were required to provide their CFJ judgments after making each 5-AFC recognition decision, *"How confident are you of your answer?"* on a scale from 0–100, where 0 indicated "not at all confident" and 100 indicated "extremely confident."

After finishing each block for a pair type, participants could rest for 10 seconds before the next block was presented, or press "Next" to move on to the next block immediately. Before starting a new block, a new prompt appeared on the screen that showed the percentage of study completion to motivate the participants to continue the study. The entire study took 45–60 minutes to complete.

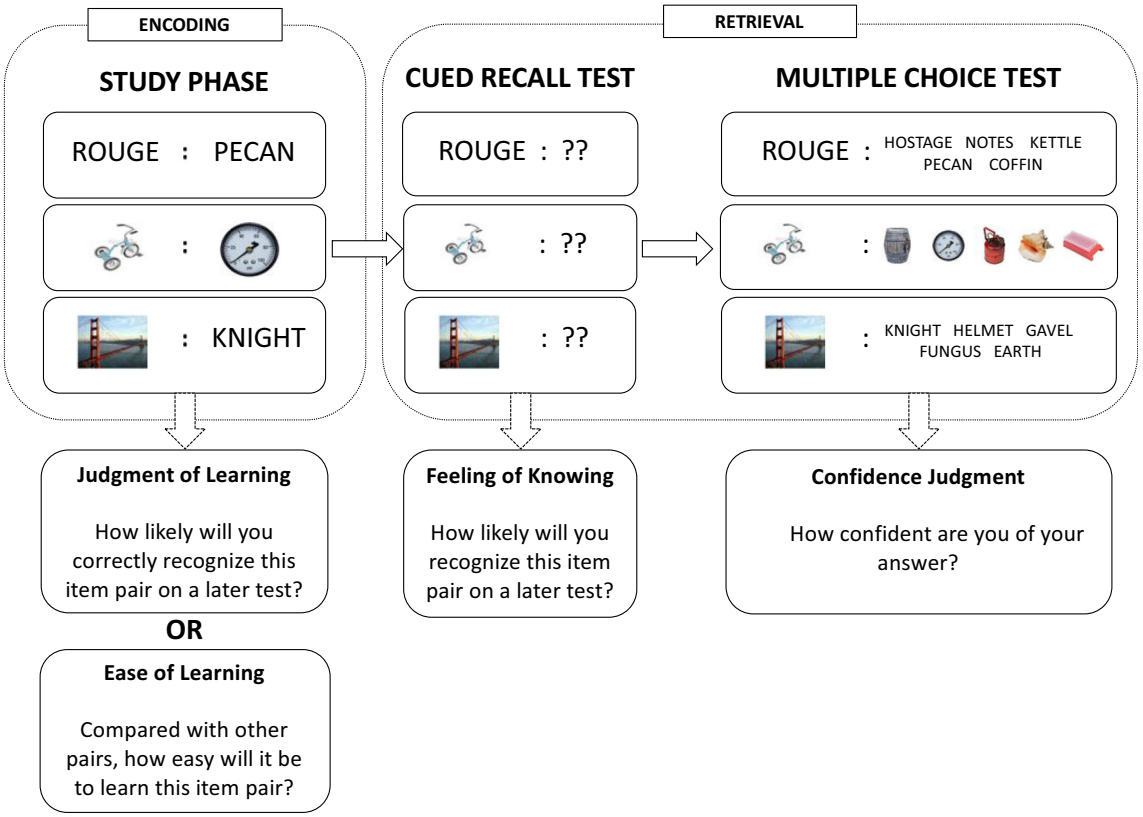

Fig. 1 Schematic of experimental paradigm across three phases. Each participant made a single metamemory judgment for a single stimulus pair type in a given block and thus were presented with 12 blocks that fully crossed each judgment–stimulus pairing. During the study phase (i.e., encoding), pairs of items were presented and immediately following each pair, participants made a judgment of learning, an ease of learning judgment, or were not asked to make a judgment. Then, a cued recall test was given, and following each tested item, participants either were asked to make a feeling of knowing judgment or no judgment at all. The last phase consisted of a multiple-choice test in participants were asked to choose which of five previously seen options was previously paired with the first item. Following each memory decision, participants either made a confidence judgment or no judgment at all

Data analysis

In preparation of data analyses, all monitoring judgment values (EOLs, JOLs, FOKs, and CJs) were centered and converted to *z*-scores within each participant and each stimulus pair. For the monitoring bias models, these values were entered directly into the SEMs. For the monitoring accuracy models, accuracy scores first had to be calculated. For this purpose, we used generalized linear mixed effects modeling (GLMM) via the lme4 package (Bates et al., 2015) in R. This method has the advantage of not eliminating participants due to the lack of variability in measures, such as constant monitoring judgments, and retaining participants who have missing data (Murayama et al., 2014). The GLMM approach used recognition accuracy (correct/incorrect) as the dependent variable and the independent variables included the standardized monitoring judgment and stimulus-pair condition (categorically coded) with their interactions using the picture–word stimulus pair as a reference group. A separate GLMM was conducted for each monitoring judgment. A maximal random effect structure and the "bobyqa" optimizer was used (Barr et al., 2013). Participants and items within trials were modeled as random effects (i.e., both random intercepts and random slopes for each random factor was included in all the comparison model analysis). From these analyses, beta values were extracted for each of the 12 conditions and for each participant that represented how well each monitoring judgment in the picture–picture and word–word conditions predicted recognition memory accuracy relative to the picture-word condition. These relative beta values were then used to create metamemory accuracy scores for each of the three conditions within each of the metamemory assessment types by combining the intercept, participant beta score, and the appropriate estimate from the fixed effects results. The individual metamemory accuracy score was calculated differently for each of the conditions. Because the picture–word condition served as the comparison condition, the intercept from the fixed effects from the GLMM analysis was added to the intercept of each individual participant to indicate the degree to which the participant's slope differed from the aggregate slope in the GLMM. The individual metamemory accuracy score for the picture–picture and word–word conditions were calculated by adding the intercept of the metamemory predictions overall and the intercept for the picture–picture or word–

word fixed effects, respectively, to the individual participant's intercept and overall metamemory beta weight. These metamemory accuracy scores served as the observed variables that were entered into the SEM models.

Before developing the main SEM models, we first examined the descriptive statistics including means, standard deviations, skewness, and kurtosis, to check the normality assumption for using the SEM approach. To account for the nonnormality in our data, we used robust estimation (i.e., MLR) embedded in Mplus Version 6 (Muthén & Muthén, 1998–2011). Correlations among the variables also were checked to verify the theoretically based latent factor structure of our models (see Table 2). Mplus was used to conduct a series of confirmatory factor analyses (CFAs) to develop models that each represented separate hypothesized monitoring structures. All models used 12 individual indicators loading onto four factors for monitoring judgment types (JOL, EOL, FOK, and CFJ; see Fig. 2). The first latent variable was for JOL, comprised of the measured JOL judgments from each stimulus type (word–word, picture–picture, picture–word). The second latent variable was for EOL, the third for FOK, and the fourth for CFJ. Prior to proceeding with our baseline (four-factor) model, a fully saturated model that allowed all variables to covary was optimized to favor the principle of parsimony by removing covariances among similar stimuli.

To test our primary research question, we created two hypothesized models by adding distinct second-order latent variables to the baseline model (Model 1). These models are referred to as the unitary monitoring model (Model 2) and the specialized monitoring model (Model 3). For Model 2, latent structures for all four monitoring judgments (JOL, EOL, FOK, and CFJ) were used to create a unitary second-order latent construct called *Unitary Monitoring*. For Model 3, we created two second-order latent constructs, *encoding-based judgments* indicated by JOL and EOL and *retrieval-based judgments* indicated by FOK and CFJ. Note that other models of theoretical interest could not be modeled based on the design (e.g., third-order models or other nonbalanced metamemory groupings). To assess the fit of our hypothesized models, we used several widely used fit indices including the χ^2 goodness-of-fit statistic, the Comparative Fit Index (CFI; values above .95 indicate a good model fit), the

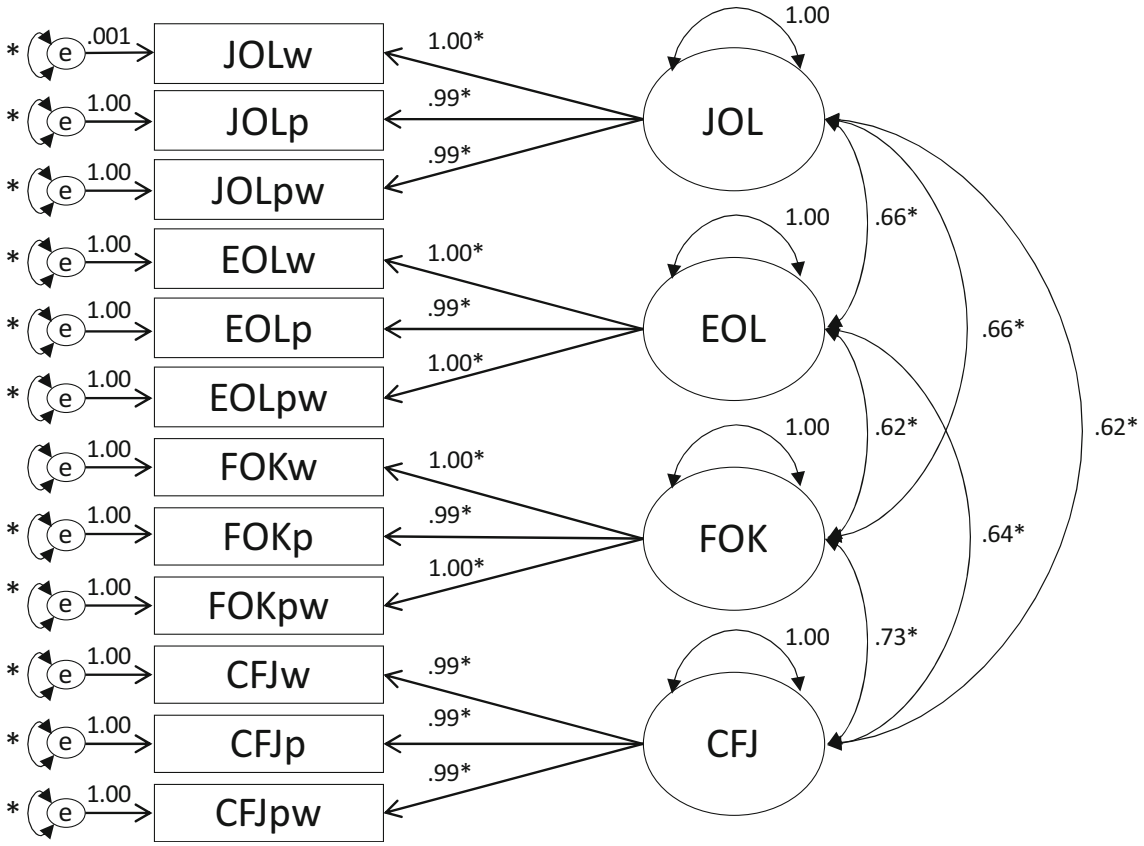

Fig. 2 Model 1: Baseline model. In this model, all four judgment latent constructs, JOL, EOL, FOK, and CFJ were allowed to covary, each with three observed variables. The numbers alongside the arrows are standardized parameter estimates. The numbers alongside the curved arrows are correlation between the latent constructs. All standardized estimates significant at $p < .001$ level. Abbreviations for monitoring judgments are as follows: JOL = judgment of learning; EOL = ease of learning judgment; FOK = feeling of knowing judgment; CFJ = confidence judgment; JOLw = JOL word–word pairs; JOLp = JOL picture–picture pairs; JOLpw = JOL picture–word pairs; EOLw = EOL word–word pairs; EOLp = EOL picture–picture pairs; EOLpw = EOL picture–word pairs; FOKw = FOK word–word pairs; FOKp = FOK picture–picture pairs; FOKpw = FOK picture–word pairs; CFJw = CFJ word–word pairs; CFJp = CFJ picture–picture pairs; CFJpw = CFJ picture–word pairs

root-mean-square error of approximation (RMSEA; values below .05 indicate a good model fit), and the standardized root-mean-square residual (SRMR; values below .08 indicate a good model fit), Akaike information criterion (AIC; smaller values indicate a better fit) and Bayesian information criterion (BIC; smaller values indicate a better fit). Because the robust estimator with MLR was used for analyzing the data, the Satorra–Bentler (S-B) scaled χ^2 difference test (Satorra & Bentler, 2010) was used to compare the models based on the χ^2 goodness-of-fit statistic (Muthén & Muthén, 1998–2011).

Results

Table 1 presents the descriptive statistics for the metamemory variables. Table 2 displays the bivariate correlations across all variables. The baseline model (Model 1) consisted of four separate latent factors (one per judgment). All latent factors (JOL, EOL, FOK, and CFJ) were allowed to covary to understand the relationship between the judgments. This model resulted in a good fit, $\chi^2(36) = 251.45$, $p < .001$; RMSEA = 0.14; CFI = 0.98; SRMR = 0.008, AIC = −107.34, BIC = 97.65. As shown in Fig. 2, each indicator loaded highly onto the latent constructs for each judgment and each judgment had a strong relationship with the others (all >.62). This model indicated that all monitoring judgments share some similar properties, leading to high correlations among them; however, some judgments were more strongly related to certain judgments than others.

We next compared results for the unitary monitoring model (Model 2) and the specialized monitoring model (Model 3) to the baseline model (Model 1). The unitary monitoring model can be found in Fig. 3 and resulted in a good fit, χ^2 (38) = 261.51, $p <$ .001; RMSEA = 0.13; CFI = 0.98; SRMR = 0.025, AIC = −97.19, BIC = 100.20. Each indicator loaded highly onto the latent constructs for each judgment and each latent construct also loaded highly onto a unitary monitoring latent construct (ranging from .77 to .85). However, when directly compared to the baseline model, this model was a poorer fit ($t = 10.74$, $p = .005$).

The specialized monitoring model (Model 3) that distinguished between encoding-based and retrieval-based judgments can be found in Fig. 4. Model 3 showed a good fit, χ^2 (37) = 250.75, $p < .001$; RMSEA = 0.13; CFI = 0.98; SRMR = 0.012, AIC = −107.13, BIC = 94.06. Notably, the loadings for the JOL and EOL latent constructs loaded numerically higher onto the encoding construct and the FOK and CFJ latent constructs loaded numerically higher onto the retrieval construct than on the unitary monitoring construct in Model 2. Model 3 had a comparatively better fit across most of the fit indices than Model 2, including the S-B scaled χ^2 difference test ($t = 10.75$, $p = .001$). This improvement in model fit indicated that monitoring judgments seem not to fall under a unitary ability. We also compared Model 3 with the baseline model. Although many of the fit indices indicated that Model 3 was as good or a better fit than Model 1 (except for SRMR), the S-B scaled χ^2 test was not significant, ($t = 1.44$, $p = .23$). Overall, the analyses clearly reject the notion of a unitary monitoring construct in favor of specialized monitoring constructs and a slight favor toward the two-factor model. This final model differentiates judgments based on the time at which each were made: at encoding or following a retrieval attempt. Parallel SEMs were conducted for monitoring bias and yielded the same patterns of results (see Supplemental Material).

Table 1 Means and standard deviations for observed variables

Indicators	Bias rating		Recognition performance		Monitoring accuracy (beta)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Word–Word JOL	28.40	22.80	0.63	0.29	0.51	1.54
Word–Word EOL	30.41	22.82	0.66	0.29	0.35	1.51
Word–Word FOK	22.84	22.42	0.50	0.27	1.06	1.56
Word–Word CFJ	40.38	30.88	0.51	0.28	-0.08	1.66
Picture–Picture JOL	30.67	22.54	0.64	0.28	0.24	1.44
Picture–Picture EOL	29.47	21.09	0.65	0.30	0.39	1.58
Picture–Picture FOK	22.72	21.45	0.51	0.26	0.83	1.69
Picture–Picture CFJ	44.01	31.38	0.53	0.28	-0.06	1.57
Picture–Word JOL	36.16	22.75	0.80	0.22	0.41	1.54
Picture–Word EOL	35.91	21.60	0.78	0.23	0.44	1.57
Picture–Word FOK	36.09	25.15	0.67	0.26	1.27	1.58
Picture–Word CFJ	58.84	30.40	0.69	0.26	-0.09	1.61

Note. Abbreviations for monitoring judgments are as follows: judgment of learning = JOL; ease of learning judgment = EOL; feeling of knowing judgment = FOK; confidence judgment = CFJ. Stimuli pairs for each judgment either consisted of two words (Word–Word), two pictures (Picture–Picture), or one picture with an associated target word (Picture–Word).

Table 2 Correlations for monitoring accuracy scores and monitoring magnitude ratings

Observed variables	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. JOLw	–	0.83	0.49	0.52	0.64	0.71	0.44	0.49	0.63	0.60	0.44	0.37
2. EOLw	0.66	–	0.50	0.50	0.60	0.73	0.45	0.44	0.62	0.58	0.46	0.36
3. FOKw	0.66	0.62	–	0.66	0.31	0.40	0.59	0.46	0.36	0.37	0.52	0.42
4. CFJw	0.61	0.62	0.71	–	0.39	0.49	0.52	0.58	0.40	0.38	0.47	0.58
5. JOLp	1.00	0.65	0.67	0.61	–	0.76	0.43	0.50	0.41	0.38	0.30	0.30
6. EOLp	0.66	1.00	0.63	0.63	0.66	–	0.52	0.50	0.59	0.59	0.45	0.41
7. FOKp	0.65	0.62	0.99	0.72	0.67	0.63	–	0.58	0.41	0.42	0.50	0.43
8. CFJp	0.59	0.61	0.71	0.98	0.59	0.62	0.72	–	0.43	0.38	0.50	0.46
9. JOLpw	1.00	0.66	0.66	0.61	0.99	0.66	0.65	0.59	–	0.74	0.56	0.42
10. EOLpw	0.66	1.00	0.62	0.62	0.66	1.00	0.62	0.61	0.66	–	0.61	0.42
11. FOKpw	0.66	0.63	1.00	0.71	0.67	0.63	0.99	0.71	0.66	0.63	–	0.53
12. CFJpw	0.63	0.64	0.74	0.99	0.63	0.65	0.75	0.99	0.63	0.64	0.74	–

Note. All correlations are significant at the .001 level (two-tailed). The upper triangle includes correlations for magnitude monitoring judgments and the lower triangle includes correlations for monitoring accuracy judgments. Abbreviations for monitoring judgments are as follows: judgment of learning = JOL; ease of learning judgment = EOL; feeling of knowing judgment = FOK; confidence judgment = CFJ. For each monitoring judgment respectively: w = word–word pairs; p = picture–picture pairs; and pw = picture–word pairs.

Discussion

Metamemory monitoring is commonly referenced as a singular function despite acknowledgment that the stage at which monitoring occurs influences the outcome of such processes. The present results suggest that monitoring may be two distinct abilities: one occurring during encoding and another at retrieval. We found the same patterns whether considering monitoring bias (the decision process) or accuracy (the correspondence between the decision and memory performance). These results are bolstered by behavioral dissociations between predictions at encoding and retrieval using both experimental and material manipulations but may seem counter to

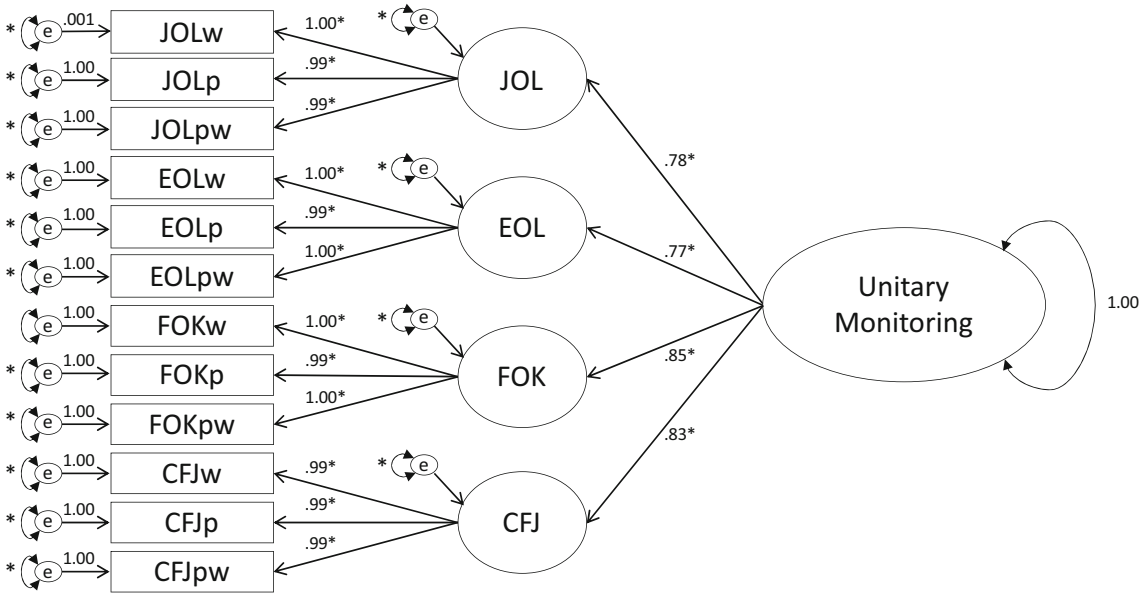

Fig. 3 Model 2: Unitary monitoring model. In this model, performance on 12 tasks loaded onto four latent construct, JOL, EOL, FOK, and CFJ, which were then loaded into one unitary latent construct, monitoring (e = error). The numbers alongside the arrows are standardized parameter estimates. All standardized estimates significant at $p < .001$ level. Abbreviations for monitoring judgments are as follows: JOL = judgment of learning; EOL = ease of learning judgment; FOK = feeling of knowing judgment; CFJ = confidence judgment; JOLw = JOL word–word pairs; JOLp = JOL picture–picture pairs; JOLpw = JOL picture–word pairs; EOLw = EOL word–word pairs; EOLp = EOL picture–picture pairs; EOLpw = EOL picture–word pairs; FOKw = FOK word–word pairs; FOKp = FOK picture–picture pairs; FOKpw = FOK picture–word pairs; CFJw = CFJ word–word pairs; CFJp = CFJ picture–picture pairs; CFJpw = CFJ picture–word pairs

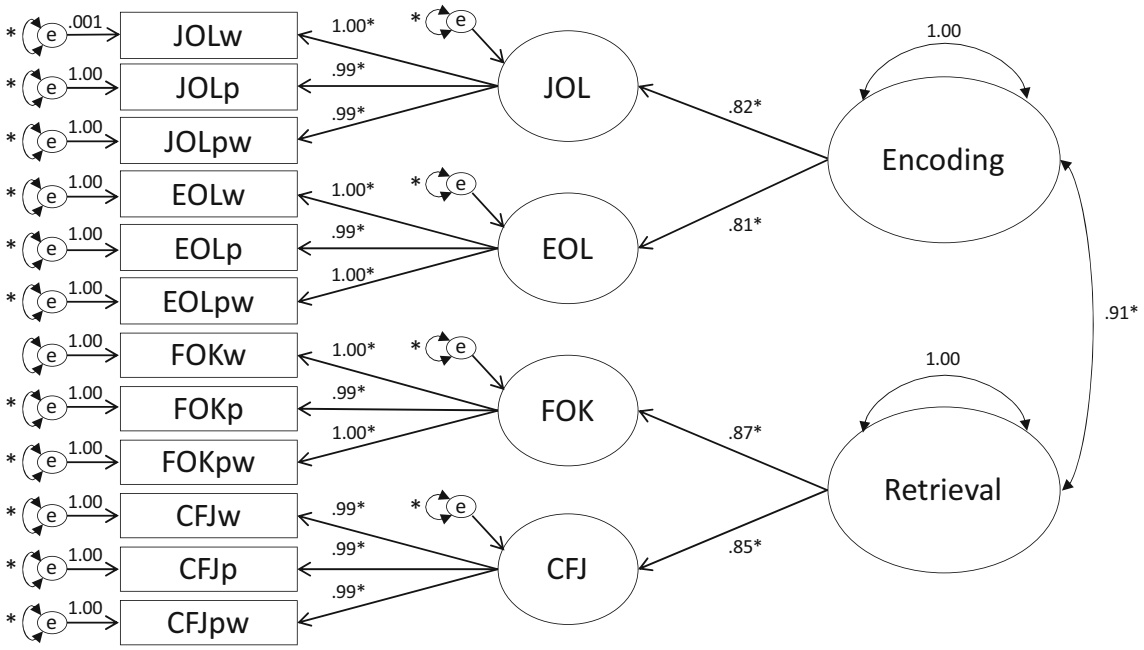

Fig. 4 Model 3: Specialized monitoring model. In this model, performance on 12 tasks were loaded onto two latent constructs, encoding and retrieval (e = error). The numbers alongside the arrows are standardized parameter estimates. All standardized estimates significant at $p < .001$ level. Abbreviations for monitoring judgments are as follows: JOL = judgment of learning; EOL = ease of learning judgment; FOK = feeling of knowing judgment; CFJ = confidence judgment; JOLw = JOL word–word pairs; JOLp = JOL picture–picture pairs; JOLpw = JOL picture–word pairs; EOLw = EOL word–word pairs; EOLp = EOL picture–picture pairs; EOLpw = EOL picture–word pairs; FOKw = FOK word–word pairs; FOKp = FOK picture–picture pairs; FOKpw = FOK picture–word pairs; CFJw = CFJ word–word pairs; CFJp = CFJ picture–picture pairs; CFJpw = CFJ picture–word pairs

notions of a single "meta-level" of monitoring and recent findings in neuroscience.

Interpreting a two-factor construct of metamemory monitoring

A two-factor structure can be interpreted in two, non-mutually exclusive ways. First, the two factors might stem from different inputs into the monitoring decision that vary between encoding and retrieval (Nelson & Dunlosky, 1991). Decision-making can be influenced by the salience of individual factors that form the basis of judgments (Schwarz et al., 1991; Tversky & Kahneman, 1973). In Koriat's (1997) cue-utilization approach, predictions are influenced by three categories of available cues: intrinsic (characteristics of stimuli), extrinsic (encoding instructions), and mnemonic (internally experienced processes). These cues and other sources of knowledge can also inform personal beliefs about how memory should work. Critically, these cues or one's beliefs may not be equally diagnostic of future memory. At encoding, for example, intrinsic characteristics of the to-be-remembered stimulus (e.g., perceptual fluency) often are salient (Koriat, 1997; Koriat & Ma'ayan, 2005) but may not always influence memory performance (Besken & Mulligan, 2013; Rhodes & Castel, 2008). At retrieval, mnemonic characteristics may be most salient, with FOKs and CFJs being more based on factors like prior experience (Koriat & Levy-Sadot, 2000), fluency (Kelley & Lindsay, 1993), familiarity (Koriat & Ma'ayan, 2005), and accessibility (Koriat, 1997). Additionally, perceptual details that were salient during EOLs and JOLs made during encoding may be forgotten at retrieval, removing their influence on FOKs and CFJs. Ultimately, monitoring judgments made at one stage (e.g., encoding) likely share common cues and biases, but across stages (encoding vs. retrieval), these cues and biases should not be assumed to be similar or equally diagnostic of memory.

Second, encoding-based and retrieval-based monitoring are composed of different sets of processes. One well-known example of these different processes is highlighted in the difference between immediate and delayed JOLs. JOLs made immediately after encoding are less accurate than those made after a delay (Nelson & Dunlosky, 1991). This effect has been explained by retrieval-based theories (Spellman & Bjork, 1992); the delay allows for a covert retrieval attempt prior to making the JOL, a process that is not available during encoding. Furthermore, interference processes occur selectively during retrieval-based monitoring, but not encoding-based monitoring (Eakin & Hertzog, 2012). Additionally, different types of inputs encourage different types of processes to be recruited during monitoring. Fluently retrieved information might prevent the initiation of monitoring processes at retrieval (McDonough et al., 2015), but also highly influence

monitoring processes at encoding (Rhodes & Castel, 2008). Retrieval-based monitoring may consist of assessments of difficulty that are more objective (the amount of information retrieved), whereas encoding-based monitoring judgments may consist of subjective levels of difficulty due to fewer diagnostic cues being available (Mazancieux et al., 2020). Relatedly, retrieval-based monitoring affords real-time feedback that can directly influence control processes—such as memory search strategies—whereas encoding-based processes do not (Mazancieux et al., 2020). Future research might systematically assess the degree to which different inputs qualitatively alters the monitoring process and its dependence on stage of monitoring.

The early role of neuroscience in metamemory

Much of the inferences linking monitoring to the PFC has stemmed from lesion patients. Although such studies vary in their locus and breadth of damage, several recent fMRI studies have provided more nuanced underpinnings of the brain's role in monitoring. The medial PFC, specifically, is consistently activated during encoding-based judgments like JOLs (Do Lam et al., 2012; Kao et al., 2005). However, other brain regions (the posterior parietal cortex; PPC) have been more consistently activated during retrieval-based judgments like FOKs (e.g., Maril et al., 2005). Medial PFC is known to support introspection such as self-referencing and mental simulation whereas PPC regions support attention during memory retrieval; both are also involved in monitoring (Bastin et al., 2019; Mitchell & Johnson, 2009). Thus, the bases for making encoding-based judgments may rely more on cue introspection while bases for retrieval-based judgments may rely more on memory retrieval mechanisms. We note that only a few fMRI studies have compared different metamemory judgments as we have done here to critically test shared and unique brain regions at a network level. Doing so would further clarify the structure of metamemory monitoring, especially beyond the PFC.

Other considerations

The distinction between these two monitoring constructs should be considered within the context of the high covariation among them. These high correlations suggest that these two factors are not completely distinct. Although the source of this covariation is not clear, it may be due to shared "meta-level" structures through which monitoring or control processes operate (Nelson & Narens, 1990). Alternatively, the high correlation could be a common-method bias that occurs when the same method is used to measure different constructs (Podsakoff et al., 2003). Relatedly, some of these high correlations could be due to individual differences in response tendencies. A liberal or very confident person might provide high values on every judgment, whereas a conservative or not confident person might provide low values on every judgment. Such biases would inflate the overall correlation among all the measures in a systematic manner and would suggest that the distinction between the two monitoring constructs would likely be even more pronounced. This bias would be most apparent for monitoring bias, but less relevant for accuracy. Given that similar relationships among the monitoring judgments were found regardless of whether monitoring bias or accuracy was assessed, any explanation would have to address both patterns.

This two-factor model of metamemory monitoring provides information that translates to practical considerations. The first is how monitoring is conceptualized in research studies. We propose that a lack of systematic exploration of manipulations and verifications of effects across monitoring judgments represents an implicit notion of a unitary monitoring ability. Even in studies that acknowledge different types of metamemory judgments, investigations of monitoring often use only one judgment and do not explicitly acknowledge potential limits in generalizability to other metamemory judgments (e.g., Pinon et al., 2005; Zawadzka & Higham, 2016). Take the metacognitive illusion of font size in which font size influences JOLs but not memory (Rhodes & Castel, 2008). Despite having hundreds of citations—representing an important discovery—it took 10 years before researchers formally studied the effect on retrieval-based monitoring judgments (Luna et al., 2018). They reported a reduced font size effect for judgments at retrieval, showing the effect is not representative of other monitoring judgments. Moving forward, rather than referring to "monitoring" as a homogeneous construct, scientific inferences would be clearer by specifying the investigation of encoding-based monitoring or retrieval-based monitoring. Additionally, appropriate generalizations—or lack thereof—should be addressed if both types of monitoring were not assessed in a given study.

A second practical consideration is when metamemory should be assessed both in research and in educational settings. Because monitoring judgments are always based on missing and sometimes misleading cues (e.g., metamemory illusions), how can learning be efficiently assessed? Assessing learning at every point in the process takes time and might be redundant, leading to barriers implementing learning assessments in large-scaled applications or for large numbers of learning items. Our findings suggest that educators and trainers should examine monitoring at *both* encoding and retrieval, but different assessments during each stage are unnecessary. For example, students should not rely on monitoring at encoding to be predictive of monitoring at retrieval. This factor is particularly important for skills used in life-and-death situations. For instance, physicians should not rely only

on prior assessment of skill learning, but also monitor retrieval of those skills at the time of use. In eyewitness memory, the timing of monitoring can influence the accuracy of eyewitness identification (Palmer et al., 2013). Our findings might explain why early eyewitness confidence differs from confidence assessments made later.

Conclusion

The present study uniquely measured multiple monitoring judgments among the same individuals, allowing us to model how monitoring judgments relate to one another. Our results suggest that metamemory monitoring is best conceptualized as a set of related abilities, but separated across encoding and retrieval processes. Consequently, we caution against the categorization and use of monitoring as a unitary construct, whether implicitly or explicitly, and encourage the inclusion of both encoding-based and retrieval-based monitoring judgments to efficiently and systematically investigate metamemory monitoring by measuring and perhaps finding new dissociations in monitoring outcomes. By doing so, we may also reach a more accurate understanding of the basis of metacognition and the individual differences within these abilities, and any potential differences in how encoding-based and retrieval-based monitoring judgments relate to actual memory performance.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13423-021-01930-z.

Acknowledgments We thank Elaine Tan for providing help on the generalized linear mixed-effects modeling to calculate the monitoring accuracy scores.

Authors' contributions I.M.M. developed the original idea for the study. T.E. and D.K.E. were in charge of data collection.

T.E., K.R.K., D.K.E., and M.K. performed the data analysis with critical input from I.M.M. I.M.M. and M.K. were responsible for the interpretation of the data. I.M.M., T.E., K.R.K., and D.K.E. wrote the manuscript, and all contributed advice as well as revisions. All authors read and approved the final manuscript.

Data availability The datasets supporting the conclusions of this article can be requested from the corresponding author.

Declarations

Ethics approval and consent to participate This research was conducted in compliance with the ethical guidelines of the APA, all participants were consented before participating, and was approved by the IRBs of the University of Alabama and Mississippi State University.

Consent for publication Not applicable.

Code availability Syntax and output of primary analyses are available in supplemental material.

Competing interests The authors declare no competing interests.

References

Bastin, C., Besson, G., Simon, J., Delhaye, E., Geurten, M., Willems, S., & Salmon, E. (2019). An integrative memory model of recollection and familiarity to understand memory deficits. *Behavioral and Brain Sciences*, 1–66. https://doi.org/10.1017/S0140525X19000621

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Besken, M., & Mulligan, N. W. (2013). Easily perceived, easily remembered? Perceptual interference produces a double dissociation between metamemory and memory performance. *Memory & Cognition*, *41*(6), 897–903. https://doi.org/10.3758/s13421-013-0307-8

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(38), 14325–14329. https://doi.org/10.1073/pnas.0803390105

Chua, E. F., Pergolizzi, D., & Weintraub, R. R. (2014). The cognitive neuroscience of metamemory monitoring: understanding metamemory processes, subjective levels expressed, and metacognitive accuracy. *The cognitive neuroscience of metacognition* (pp. 267-291). Springer.

Dodson, C., Bawa, S., & Krueger, L. (2007) Aging, metamemory, and high-confidence errors: A misrecollection account. *Psychology and Aging, 22*(1), 122–33. https://doi.org/10.1037/0882-7974.22.1.122

Do Lam, A. T., Axmacher, N., Fell, J., Staresina, B. P., Gauggel, S., Wagner, T., Olligs, J., & Weis, S. (2012). Monitoring the mind: the neurocognitive correlates of metamemory. *PLOS ONE*, *7*(1), Article e30009. https://doi.org/10.1371/journal.pone.0030009

Eakin, D. K. (2010). ListChecker Pro 1.2: A program designed to facilitate creating word lists using the University of South Florida word association norms. *Behavior Research Methods, 42*(4), 1012–1021. https://doi.org/10.3758/BRM.42.4.1012

Eakin, D. K., & Hertzog, C. (2012). Immediate judgments of learning are insensitive to implicit interference effects at retrieval. *Memory & cognition*, *40*(1), 8–18. https://doi.org/10.3758/s13421-011-0138-4

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417

Gonsalves, B., & Paller, K. A. (2000). Neural events that underlie remembering something that never happened. *Nature Neuroscience*, *3*(12), 1316–1321. https://doi.org/10.1038/81851

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology and Aging*, *5*, 215–227. https://doi.org/10.1037/0882-7974.5.2.215

Hertzog, C., & Dunlosky, J. (2011). Metacognition in later adulthood: Spared monitoring can benefit older adults' self-regulation. *Current*

Directions in Psychological Science, *20*(3), 167–173. https://doi.org/10.1177/0963721411409026

Kao, Y. C., Davis, E. S., & Gabrieli, J. D. (2005). Neural correlates of actual and predicted memory formation. *Nature Neuroscience*, *8*(12), 1776–1783. https://doi.org/10.1038/nn1595

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition*, *28*(1), 92–107. https://doi.org/10.3758/BF03211579

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*, 1–24. https://doi.org/10.1006/jmla.1993.1001

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language, 48,* 704–721. https://doi.org/10.1016/S0749-596X(02)00504-1

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370.

Koriat, A., & Levy-Sadot, R. (2000). Conscious and unconscious metacognition: A rejoinder. Consciousness and Cognition, 9(2), 193–202. https://doi.org/10.1006/ccog.2000.0436

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of memory and Language*, *52*(4), 478–492. https://doi.org/10.1016/j.jml.2005.01.001

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 464–470. https://doi.org/10.1037/0278-7393.16.3.464

Luna, K., Martín-Luengo, B., & Albuquerque, P. B. (2018). Do delayed judgments of learning reduce metamemory illusions? A meta-analysis. *Quarterly Journal of Experimental Psychology, 71*, 1626–1636. https://doi.org/10.1080/17470218.2017.1343362

Maril, A., Simons, J. S., Weaver, J. J., & Schacter, D. L. (2005). Graded recall success: An event-related fMRI comparison of tip of the tongue and feeling of knowing. *NeuroImage*, *24*(4), 1130–1138. https://doi.org/10.1016/j.neuroimage.2004.10.024

Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. A. (2020). Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *Journal of Experimental Psychology: General, 149*(9), 1788–1799. https://doi.org/10.1037/xge0000746

McDonough, I. M., Bui, D. C., Friedman, M. C., & Castel, A. D. (2015). Retrieval monitoring is influenced by information value: The interplay between importance and confidence on false memory. *Acta Psychologica*, *161*, 7–17.

Metcalfe, J. (2000). Metamemory: Theory and data. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 197–211). Oxford University Press.

Mitchell, K. J., & Johnson, M. K. (2009). Source monitoring 15 years later: What have we learned from fMRI about the neural mechanisms of source memory?. *Psychological Bulletin*, *135*(4), 638–677. https://doi.org/10.1037/a0015849

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1287–1306. https://doi.org/10.1037/a0036914

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus user's guide* (6th ed.). Muthén & Muthén.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science*, *2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 402–407. https://doi.org/10.3758/BF03195588

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 26, pp. 125–173). Academic. https://doi.org/10.1016/S0079-7421(08)60053-5

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*(1), 55–71. https://doi.org/10.1037/a0031602

Pannu, J. K., & Kaszniak, A. W. (2005). Metamemory experiments in neurological populations: A review. *Neuropsychological Review*, *15*, 105–130. https://doi.org/10.1007/s11065-005-7091-6

Pinon, K., Allain, P., Kefi, M. Z., Dubas, F., & Le Gall, D. (2005). Monitoring processes and metamemory experience in patients with dysexecutive syndrome. *Brain and Cognition*, *57*(2), 185–188. https://doi.org/10.1016/j.bandc.2004.08.042

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*, 615–625. https://doi.org/10.1037/a0013684

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243–248. https://doi.org/10.1007/s11336-009-9135-y

Schraw, G., Dunkle, M. E., Bendixen, L. D., & Roedel, T. D. (1995). Does a general monitoring skill exist? *Journal of Educational Psychology*, *87*(3), 433–444. https://doi.org/10.1037/0022-0663.87.3.433

Schwartz, B. L., & Bacon, E. (2008). Metacognitive neuroscience. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 355–371). Psychology Press.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61*(2), 195–202. https://doi.org/10.1037/0022-3514.61.2.195

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, *3*(5), 315–317. https://doi.org/10.1111/j.1467-9280.1992.tb00680.x

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Zawadzka, K., & Higham, P. A. (2016). Recalibration effects in judgments of learning: A signal detection analysis. *Journal of Memory and Language*, *90*, 161–176. https://doi.org/10.1016/j.jml.2016.04.005