Simultaneous utilization of multiple cues in judgments of learning

Monika Undorf¹ · Anke Söllner¹ · Arndt Bröder¹

Published online: 11 January 2018
© Psychonomic Society, Inc. 2018

Abstract

There is much evidence that metacognitive judgments, such as people's predictions of their future memory performance (judgments of learning, JOLs), are inferences based on cues and heuristics. However, relatively little is known about whether and when people integrate multiple cues in one metacognitive judgment or focus on a single cue without integrating further information. The current set of experiments systematically addressed whether and to what degree people integrate multiple extrinsic and intrinsic cues in JOLs. Experiment 1 varied two cues: number of study presentations (1 vs. 2) and font size (18 point vs. 48 point). Results revealed that people integrated both cues in their JOLs. Experiment 2 demonstrated that the two word characteristics concreteness (abstract vs. concrete) and emotionality (neutral vs. emotional) were integrated in JOLs. Experiment 3 showed that people integrated all four cues in their JOLs when manipulated simultaneously. Finally, Experiment 4 confirmed integration of three cues that varied on a continuum rather than in two easily distinguishable levels. These results demonstrate that people have a remarkable capacity to integrate multiple cues in metacognitive judgments. In addition, our findings render an explanation of cue effects on JOLs in terms of demand characteristics implausible.

Keywords Metamemory · Judgments of learning · Cue integration

Metacognition—the ability to think about one's own thoughts and cognitions—is among the most fascinating aspects of the human mind. In experimental psychology, the study of metacognition is inextricably linked to obtaining metacognitive judgments. Classic metacognitive judgments include retrospective confidence judgments, feeling of knowing judgments, and judgments of learning (e.g., Dunlosky & Tauber, 2014; Koriat, 2007). In using these judgments, much has been learned about metacognition. Most importantly, research demonstrated that people do not have privileged access to their cognitions. Instead, people infer the state of their own cognitive systems using cues and heuristics. For instance, Koriat's (1997) cue-utilization theory suggests that people base predictions of future memory performance (judgments of learning, JOLs) on different types of cues. Intrinsic cues, such as word frequency or concreteness, are inherent to study materials. In contrast, extrinsic cues, such as presentation time or number of study presentation, are bound to specific study conditions.

Based on a review of early JOL studies, Koriat (1997) argued that JOLs are usually sensitive to intrinsic cues but often insensitive to extrinsic cues (but see Dunlosky & Matvey, 2001; Jang & Nelson, 2005).

Over the years, considerable evidence for the impact of various specific cues on metacognitive judgments has accumulated. For instance, Rhodes and Castel (2008) found that the font size of to-be-remembered words influences JOLs, Zimmerman and Kelley (2010) demonstrated that JOLs take account of the word characteristic emotionality, and Mueller, Tauber, and Dunlosky (2013) summarized results showing that associative relatedness between members of a word pair affects JOLs. By comparison, the question of whether and how multiple cues combine to affect metacognitive judgments has received less attention (Rhodes, 2016). Resolving this question is essential for two reasons. First, evidence that a cue affects JOLs when manipulated in isolation does not necessarily imply that this cue affects JOLs in situations where multiple cues are available (e.g., Undorf & Erdfelder, 2013). As important, focusing on single cues falls short of understanding metacognition in everyday situations that usually provide learners with multiple potentially relevant cues (Rhodes, 2016). In the current set of experiments, we therefore empirically evaluate the integration of multiple cues in JOLs.

✉ Monika Undorf
undorf@uni-mannheim.de

¹ Department of Psychology, School of Social Sciences, University of Mannheim, Schloss, Ehrenhof-Ost, 68131 Mannheim, Germany

Cue integration in judgments

The idea that judgments are based on cues and heuristics is not unique to metacognitive judgments but is commonly held in research on judgment and decision-making (JDM; e.g., Koriat, 1997, 2015). In JDM research, the question of how and when people integrate multiple pieces of information in their judgments has been a focus for more than seven decades. Following the seminal work of Brunswik (1944, 1955), numerous studies have investigated the integration of multiple cues in diverse judgmental contexts, including clinical judgments, diagnoses, lie detection, predictions, and staff decisions. Reviews of this research (e.g., Karelaia & Hogarth, 2008) documented high multiple correlation coefficients for predicting judgments by the cues that are available in the judgment context. These correlations are interpreted as indicating additive integration or compensatory use of cues (e.g., Brehmer, 1994; Einhorn, Kleinmuntz, & Kleinmuntz, 1979).

This conclusion, however, has been challenged for theoretical and methodological reasons since the heuristics and biases program (cf. Tversky & Kahneman, 1974). Theoretically, it has been argued that human judgment focuses on one cue in a noncompensatory fashion and follows this cue without integrating further information because people's cognitive resources are limited (Gigerenzer, Todd, & the ABC Research Group, 1999; Todd, Gigerenzer, & the ABC Research Group, 2012). Methodologically, it has been argued that high multiple correlation coefficients of the linear model can occur even when people actually use one-cue strategies (Bröder, 2000; Martignon & Hoffrage, 2002).

There is an emerging consensus that people use simple one-cue heuristics under certain circumstances, such as when they are under time pressure or when they have to retrieve information from memory (for an overview, see Pachur & Bröder, 2013). At the same time, additive integration of multiple cues occurs in various judgment domains (e.g., Anderson, 1981, 2013).

From the perspective of JDM research, there is good reason to suspect that JOLs integrate multiple cues in an additive manner. This is because cue information is often readily available when making JOLs, meaning that there are no explicit search costs for retrieving cue information from the environment or from memory (e.g., Platzer, Bröder, & Heck, 2014). For instance, information about whether an item is printed in large or small font, is presented for study for 2 or 8 s, or evokes an emotional reaction is an integral part of studying it (salience differences between cues notwithstanding). However, it is also plausible to predict that JOLs rely on simple one-cue heuristics. The reason for this is that JOLs refer to a relatively noisy environment (i.e., one's own memory performance). Consequently, cue validities are hard to learn, and people might rely on the

most obvious and valid cues. Moreover, in judgment domains with high stochastic uncertainty, simple and frugal heuristics were found to be particularly robust (Gigerenzer & Brighton, 2009). In sum, there are two possibilities, both of which are plausible from a JDM perspective: When making JOLs, people might integrate multiple cues or focus on single cues.

Cue integration in JOLs

Prior JOL studies that focused on cue integration investigated JOLs made during multitrial learning (e.g., Hertzog, Hines, & Touron, 2013; Hines, Hertzog, & Touron, 2015; Tauber & Rhodes, 2012). In these experiments, people studied the same material in two or more study-test trials. Results revealed that JOLs from later study trials were based on multiple cues, including prior JOLs, prior memory performance, and prior recognition confidence judgments. However, it is possible that all these cues are specific instantiations of a single general cue (e.g., subjective item difficulty). These studies therefore do not provide strong evidence for the integration of multiple cues in JOLs outside of multitrial learning.

At the same time, several observations are relevant to the issue of cue integration in JOLs, although they were obtained in studies that addressed different research questions. All these observations came from experiments that manipulated two or more cues in within-participant designs. In an early experiment with extrinsic cues, Zechmeister and Shaughnessy (1980) demonstrated that number of study presentations and repetition of items in a massed or distributed fashion both affected JOLs (notably, JOLs were higher for massed items than for distributed items, whereas the opposite pattern was found for memory performance; see also Son & Simon, 2012). More recent studies revealed that JOLs were sensitive to the jointly manipulated cues font format (standard vs. aLtErNaTiNg) and font size¹ (Rhodes & Castel, 2008), number of study presentations and font size (Kornell, Rhodes, Castel, & Tauber, 2011), or font style and font size (Price, McElroy, & Martin, 2016). In contrast, Benjamin (2005) demonstrated that unspeeded JOLs for word pairs were affected by target duration but not by cue duration, whereas the reverse

¹ While most cues can be easily classified as either intrinsic or extrinsic in the sense of Koriat's (1997) cue-utilization theory of JOLs, classification of font size is ambiguous (see also Dunlosky & Matvey, 2001). Considering font size as an intrinsic cue is justified, in one sense, by the fact that it concerns properties inherent to the item itself (Rhodes, 2016; Rhodes & Castel, 2008). However, font size fits at least as well in the category of extrinsic cues inasmuch as it is a feature that is randomly assigned to items and therefore specific to particular study conditions. For the present study, we reserve the term *intrinsic* to cues that are inseparable from the item (such as concreteness or emotionality) and therefore consider font size an extrinsic cue. Importantly, the classification of font size is not critical to our general conclusion.

was true for speeded JOLs (for similar findings, see Metcalfe & Finn, 2008). Similarly, Pyc and Rawson (2012) found that JOLs were sensitive to criterion level (number of correct recalls during practice) but were insensitive to the lag between practice trials, and Tauber and Rhodes (2010, Experiment 4) revealed that list length but not list order affected JOLs. With pictorial stimuli, Besken (2016, Experiment 3) found an effect of presentation format (intact vs. degraded) on JOLs, but no effect of the matching of a preceding contour. In summary, while some studies hinted that people integrated two extrinsic cues in JOLs, other studies showed that one cue affected JOLs while another cue was ignored.

A similar picture emerges from experiments that manipulated two or more intrinsic cues. In paired-associate learning, Begg, Duft, Lalonde, Melnick, and Sanvito (1989) found that cue concreteness and target concreteness both affected JOLs, whereas Illman and Morrison (2011) reported that JOLs were sensitive to cue imageability and target age of acquisition but not to cue age of acquisition and target imageability. Also, Hourihan, Fraundorf, and Benjamin (2017) revealed that JOLs were sensitive to word frequency but insensitive to valence and arousal.

Finally, most studies that manipulated one extrinsic and one intrinsic cue alluded to cue integration. When manipulating presentation time and relatedness (Jang & Nelson, 2005; Koriat, 1997), number of study presentations and relatedness (Jang & Nelson, 2005), font size and relatedness (Price & Harrison, 2017; Rhodes & Castel, 2008), announced retention interval and relatedness (Koriat, Bjork, Sheffer, & Bar, 2004, Experiment 3b), and font format and relatedness (Mueller et al., 2013), both cues affected JOLs. Similarly, Mueller and Dunlosky (2017, Experiment 6) showed that both font color (associated with an induced belief) and stimulus category (word vs. non-word) affected JOLs. In Mueller, Dunlosky, Tauber, and Rhodes's (2014) Experiment 1, an interactive effect of font size and stimulus category on JOLs indicated cue integration. Peynircioglu, Brandler, Hohman, and Knutson (2014) revealed a similar interaction between presentation modality (visual vs. auditory) and syntax (coherent vs. re-arranged) on JOLs for musical pieces. Other studies manipulating extrinsic and intrinsic cues showed that cues were ignored. Magreehan, Serra, Schwartz, and Narciss (2016) reported that JOLs were insensitive to readability but sensitive to relatedness. Susser and Mulligan (2015, Experiment 2) found that writing hand (dominant vs. non-dominant) but not word frequency affected JOLs. Studies manipulating reward and relatedness showed that college students and ninth-graders based JOLs on both cues (e.g., Koriat, Ackerman, Adiv, Lockl, & Schneider, 2014; Soderstrom & McCabe, 2011). In contrast, fifth-graders and sixth-graders usually focused on the one cue that was more salient and integrated the two cues only after a

training procedure designed to foster cue integration (Koriat et al., 2014).²

Taken together, previous JOL studies that varied two or more cues suggest that people sometimes ignore available cues when making JOLs. However, for some of the cues, it remains unclear whether people ignored them because multiple cues were available or whether people would have ignored them even when manipulated in isolation (e.g., lag between practice trials, readability, matching of a preceding contour). More importantly, the studies revealing that two cues affected JOLs do not strictly warrant the conclusion of individual participants integrating cue information in their JOLs. The reason is that cue effects were tested at the aggregate level, and therefore may occur even if each individual participant based his or her JOLs on only one cue but different individuals focused on different cues. As an illustration, consider Kornell et al.'s (2011) experiment, in which number of study presentations and font size both affected JOLs. It is of course possible that all or most participants based their JOLs on both cues (i.e., integrated the two cues). However, the general pattern of results is also consistent with some participants basing their JOLs on number of study presentations only and other participants basing their JOLs on font size only. Consequently, prior studies do not provide conclusive evidence for cue integration in JOLs. Furthermore, most of the studies did not specifically target the question of cue integration.

The current study

The experiments reported here systematically investigated whether and to what degree people integrate multiple cues in JOLs. We only selected cues that are well known to affect JOLs when manipulated in isolation. Thus, if people ignore cues, it is reasonable to conclude that this is due to multiple cues being available. We tested the generality of our findings in four ways. First, we simultaneously varied two cues in Experiments 1 and 2 and more than two cues in Experiments 3 and 4. Second, we manipulated extrinsic cues in Experiment 1, intrinsic cues in Experiment 2, and combinations of intrinsic and extrinsic cues in Experiments 3 and 4. Also, we used two discrete cue levels in Experiments 1 to 3, but varied all cues continuously in Experiment 4. Finally, we included not only cues that are known to have similar effects on JOLs and memory performance (number of study presentations, concreteness, emotionality) but also a cue that is

² In their analyses, Koriat et al. (2014) did not differentiate between related and unrelated word pairs but between pairs with below-median and above-median self-paced study time. We interpret this median split as indicative of relatedness because several studies demonstrated shorter study time for related pairs than for unrelated pairs (e.g., Undorf & Ackerman, 2017; Undorf & Erdfelder, 2015).

known to affect JOLs but usually has no or minimal effects on memory performance (font size).

If the information provided by multiple cues is integrated in JOLs, (1) JOLs should be sensitive to the cues at the aggregate level and (2) individual-level analysis should reveal that a large number of participants base their JOLs on at least two cues. In contrast, if cue integration in JOLs does not occur, individual-level analysis should reveal that participants base their JOLs on only one cue.

Experiment 1

In Experiment 1, we varied the two extrinsic cues number of study presentations and font size. Participants studied words that were presented either once or twice. Half the once and twice presented words were printed in a smaller font and the rest were printed in a larger font. Participants made a JOL after the presentation of each word and, after the study phase, completed a free recall test. This procedure was similar to that of a previous experiment by Kornell et al. (2011, Experiment 1). Kornell et al., however, focused on the stability bias—namely, the observation that JOLs underestimate the beneficial effect of future study opportunities—and therefore asked participants to make JOLs only after the first study presentation of each word, knowing whether or not they would study it again. In contrast, in Experiment 1, we focused on JOLs that were made after the study of each item was completed. We expect that both number of study presentations and font size affect JOLs at the aggregate level. If information from these two extrinsic cues is integrated in JOLs, individual-level analyses should reveal that a large number of participants base their JOLs on both cues.

Method

Participants and materials Participants were 53 University of Mannheim undergraduates. Stimuli were 56 German 5–10 letter nouns. All normed values were taken from Võ et al. (2009). Words were of moderate concreteness ($M = 4.05$, $SD = 0.62$), neutral valence ($M = 0.03$, $SD = 0.24$), and moderate arousal ($M = 2.71$, $SD = 0.41$). Four additional words served as primacy buffers and were not included in the analysis.

Procedure The experiment consisted of a study phase and a free recall test. Instructions informed participants that they would study 60 words and would be asked to recall as many words as they could remember in a final test. Participants were also told that after the presentation of each word, they would be asked to estimate the probability of recalling it in the test phase. For each participant, 30 randomly chosen words (two buffer and 28 target items) were presented once for study, and the remaining 30 words were presented twice. Of both the words presented once and the words presented twice, a randomly

selected half was displayed in a small Arial font (18 point), and the other half was displayed in a large Arial font (48 point). Each word remained on the screen for 3 s. Immediately after studying each word, the JOL prompt “The chance to recall (0%–100%)?” appeared on the screen, and participants typed any whole number from 0 to 100. Consequently, we obtained one JOL for words that were presented once and two JOLs for words that were presented twice. A 100-ms blank screen preceded the presentation of each word. Following the study phase, participants performed a numerical filler task for 3 min. Finally, they were asked to write down as many of the words from the study phase as they could remember, in any order. Participants were given 5 min for free recall.

Results and discussion

Figure 1 presents JOLs and recall performance. For words that were studied twice, the figure shows JOLs from the second study presentation (JOLs from the first study presentation can be found in the [Appendix](#)). JOLs were submitted to an ANOVA, with number of study presentations (1, 2) and font size (small, large) as within-subjects factors. A significant main

Fig. 1 Mean judgments of learning (JOL) and percentage of correctly recalled words (recall) in Experiment 1, separately for words presented once (1×) or twice (2×) in a small (18 pt) or a large (48 pt) font. Error bars represent one standard error of the mean

effect of number of study presentations revealed higher JOLs for words studied twice than for words studied once, $F(1, 52) = 17.86, p < .001, \eta_p^2 = .26$. A significant main effect of font size revealed higher JOLs for words presented in a large font than for words presented in a small font, $F(1, 52) = 35.76, p < .001, \eta_p^2 = .41$. The interaction was not significant, $F < 1$.

A 2 (number of study presentations) $\times$ 2 (font size) ANOVA on recall performance revealed that actual memory was significantly better for words studied twice than for words studied once, $F(1, 52) = 150.06, p < .001, \eta_p^2 = .74$. Neither the main effect of font size nor the interaction were significant, both F s < 1 .

To analyze cue integration on the individual level, we coded participants as having based JOLs on number of study presentations if their JOLs were higher for words studied twice than for words studied once. Likewise, participants were coded as having based JOLs on font size if their JOLs were higher for words presented in a large font than for words presented in a small font. Results revealed that 32 participants (60.38%) integrated number of study presentations and font size in their JOLs (binomial test: $p < .001$). The remaining participants based their JOLs on either number of study presentations (7 participants) or font size (12 participants) or on neither cue (2 participants).

Obviously, this analysis ignores effect sizes and requires that all participants exhibit cue effects in the expected directions (i.e., predict better memory for twice-presented than for once-presented words and better memory for words printed in a large than in a small font). We therefore conducted a complementary

individual-level analysis that focused on effect sizes. Figure 2 depicts individual participants' Cohen's d for number of study presentations and font size. It reveals that the majority of participants are located in the upper right quadrant, indicating that they predicted better memory for twice-presented than for once-presented words and better memory for words printed in a large than for words printed in a small font. However, participants can be also found in the other three quadrants. When Cohen's (1977) convention of $|d| \geq .2$ for small effects is used as a criterion for reliable cue effects on JOLs, 16 participants (30.19%) focused on only one cue, as indicated by $|d| < .2$ for one cue and $|d| \geq .2$ for the other cue. In 8 participants (15.09%), neither cue had reliable effects on JOLs, as indicated by both $|d|$ s $< .2$. However, the majority of participants ($n = 29, 54.72\%$) revealed $|d|$ s $\geq .2$ for both cues, which is indicative of cue integration.

In summary, replicating Kornell et al. (2011), both number of study presentations and font size affected JOLs at the aggregate level. While number of study presentations affected recall performance, font size left recall performance unchanged. As in previous studies (e.g., Kornell et al., 2011; Rhodes & Castel, 2008), a large font increased overconfidence in JOLs. Two complementary individual-level analyses revealed that a tight majority of participants integrated both cues in their JOLs. Experiment 2 investigated whether cue integration in JOLs would extend to two intrinsic cues that have similar effects on JOLs and recall performance.

Experiment 2

In Experiment 2, we manipulated the two intrinsic cues concreteness and emotionality. Participants studied words that were either abstract or concrete. Half the words of each level of concreteness were neutral and the rest were emotional. Neutral words were low in arousal and neutral in valence, whereas emotional words were high in arousal and either positive or negative in valence. Hence, as has been done in prior studies, we manipulated arousal and valence jointly to maximize effects of emotionality on JOLs (Tauber & Dunlosky, 2012; Zimmerman & Kelley, 2010; but see Hourihan et al., 2017). If information from the two cues is integrated in JOLs, individual-level analyses should reveal that a large number of participants base their JOLs on both cues.

Method

Participants and materials Participants were 55 University of Mannheim undergraduates. Stimuli were 56 German 5–10 letter nouns. All normed values were taken from Vö et al. (2009). Half the words (28 words) were abstract and half were concrete. Mean imagery value was 2.49 ($SD = 0.35$) for abstract words and 5.65 ($SD = 0.54$) for concrete words (rated on a 7-point scale, 1 = low imageability to 7 = high imageability).

Fig. 2 Scatterplot of individual effect sizes (Cohen's d) measuring the effects of number of study presentations (x -axis) and font size (y -axis) on JOLs in Experiment 1

Half the abstract words (14 words) were low in arousal ($M = 2.10$, $SD = 0.23$; rated on a 5-point scale, 1 = *low arousal* to 5 = *high arousal*) and neutral in valence ($M = 0.10$, $SD = 0.21$; 7-point scale, $-3 = \text{very negative}$ through $0 = \text{neutral}$ to $3 = \text{very positive}$), while the rest were high in arousal and either positive (seven words, arousal: $M = 2.76$, $SD = 0.31$, valence: $M = 2.19$, $SD = 0.15$) or negative (seven words, arousal: $M = 4.05$, $SD = 0.50$, valence: $M = -2.12$, $SD = 0.13$). The same was true for concrete words: Half (14 words) were low in arousal and neutral (arousal: $M = 1.88$, $SD = 0.19$, valence: $M = 0.12$, $SD = 0.24$), while the rest were high in arousal and either positive (seven words, arousal: $M = 3.51$, $SD = 0.48$, valence: $M = 2.22$, $SD = 0.12$) or negative (seven words, arousal: $M = 3.84$, $SD = 0.36$, valence: $M = -2.09$, $SD = 0.11$). Four additional words that differed in concreteness and emotionality served as primacy buffers and were not included in the analysis.

Procedure The procedure was identical to that of Experiment 1, except that each word was presented only once in a 26-point font.

Results and discussion

Figure 3 presents JOLs and recall performance for abstract and concrete words that were neutral or emotional. JOLs were submitted to an ANOVA with concreteness (abstract, concrete) and emotionality (neutral, emotional) as within-subjects factors. A significant main effect of concreteness revealed higher JOLs for concrete words than for abstract words, $F(1, 54) = 38.83$, $p < .001$, $\eta_p^2 = .42$. A significant main effect of emotionality revealed higher JOLs for emotional words than for neutral words, $F(1, 54) = 40.91$, $p < .001$, $\eta_p^2 = .43$. The interaction was also significant, $F(1, 54) = 6.02$, $p = .017$, $\eta_p^2 = .10$, showing a more pronounced emotionality effect for concrete than for abstract words. Post hoc t tests confirmed that for both concrete and abstract words, JOLs were reliably higher for emotional than for neutral words, concrete: $t(54) = 6.80$, $p < .001$, $d = .93$, abstract: $t(54) = 4.80$, $p < .001$, $d = .65$.

A 2 (concreteness) $\times$ 2 (emotionality) ANOVA on recall performance revealed that actual memory was significantly better for concrete words than for abstract words, $F(1, 54) = 57.35$, $p < .001$, $\eta_p^2 = .52$. Recall performance was also significantly higher for emotional words than for neutral words, $F(1, 54) = 17.41$, $p < .001$, $\eta_p^2 = .24$. The interaction was not significant, $F < 1$.

An individual-level analysis based on simple mean differences suggested that 41 participants (74.55%) integrated concreteness and emotionality in their JOLs (binomial test: $p < .001$). The remaining participants based their JOLs on either concreteness (7 participants) or emotionality (5 participants) or on neither cue (2 participants). Figure 4 depicts individual participants' Cohen's d for concreteness and emotionality effects on

Fig. 3 Mean judgments of learning (JOL) and percentage of correctly recalled words (recall) in Experiment 2, separately for abstract and concrete words that were neutral or emotional. Error bars represent one standard error of the mean

JOLs. The figure reveals that most participants were located in the upper right quadrant, indicating that they predicted high concreteness and high emotionality to help memory. When $|d| \geq .2$ is used as a criterion for reliable cue effects, 2 participants (3.64%) used neither cue (i.e., both $|d|s < .2$), 28 participants (50.91%) used only one cue (i.e., one $|d| \geq .2$), and 25 participants (45.45%) integrated both cues (i.e., both $|d|s \geq .2$).

Data from Experiment 2 demonstrated that both concreteness and emotionality affected JOLs and recall performance at the aggregate level. Individual-level analyses revealed that about one half of the participants integrated both cues in their JOLs. Experiment 2 thus replicated Experiment 1's finding of cue integration. It is worth noting that, unlike in Experiment 1, both cues were intrinsic and had similar effects on JOLs and actual memory. This illustrates the robustness of additive cue integration in JOLs. Nevertheless, it is still possible that cue integration in JOLs is limited to two cues and does not occur with more than two cues. Experiment 3 therefore tested whether JOLs would integrate information from the four cues used in Experiments 1 and 2 when manipulated simultaneously.

Fig. 4 Scatterplot of individual effect sizes (Cohen's d) measuring the effects of concreteness (x-axis) and emotionality (y-axis) on JOLs in Experiment 2

Experiment 3

In Experiment 3, we varied the extrinsic cues from Experiment 1 (number of study presentations, font size) and the intrinsic cues from Experiment 2 (concreteness, emotionality). This allowed us to investigate whether people integrate up to four extrinsic and intrinsic cues in JOLs. As in the previous experiments, individual-level analyses are used to assess cue integration.

Method

Participants and materials Participants were 50 University of Mannheim undergraduates. Stimuli were 64 German 5–10 letter nouns. As in Experiment 2, half the words (32 words) were abstract ($M = 2.48$, $SD = 0.36$) and half were concrete ($M = 5.66$, $SD = 0.52$). Half the abstract words were of low arousal and neutral (16 words, arousal: $M = 2.09$, $SD = .22$; valence: $M = 0.14$, $SD = 0.22$), while the rest were of high arousal and either positive (eight words, arousal: $M = 2.74$, $SD = 0.29$, valence: $M = 2.19$, $SD = 0.14$) or negative (eight words, arousal: $M = 3.96$, $SD = 0.41$, valence: $M = -2.10$, $SD = 0.13$). The same was true for concrete words: Half were low in arousal and neutral (16 words, arousal: $M = 1.89$, $SD = 0.18$, valence: $M = 0.15$, $SD = 0.24$), while the rest were high in arousal and either positive (eight words, arousal: $M = 3.43$, $SD = 0.49$, valence: $M = 2.27$, $SD = 0.18$) or negative (eight words,

arousal: $M = 3.85$, $SD = 0.47$, valence: $M = -2.10$, $SD = 0.10$). Four additional words served as primacy buffers and were not included in the analysis.

Procedure The procedure was the same as in Experiment 1, except that for each participant, one randomly chosen fourth of the words from each of the four levels of concreteness and arousal were presented (1) once in a small font, (2) once in a large font, (3) twice in a small font, and (4) twice in a large font.

Results

Figure 5 presents JOLs and recall performance. For twice-presented words, the figure shows JOLs from the second presentation (JOLs from the first presentation can be found in the Appendix). JOLs were submitted to an ANOVA with number of study presentations (1, 2), font size (small, large), concreteness (abstract, concrete), and emotionality (neutral, emotional) as within-subjects factors. A significant main effect of number

Fig. 5 Mean judgments of learning (JOL) and percentage of correctly recalled words (recall) in Experiment 3, separately for abstract and concrete words that were neutral (neut) or emotional (emo) and studied once (1×) or twice (2×) in a small (18 pt) or a large font (48 pt). Error bars represent one standard error of the mean

of study presentations revealed higher JOLs for words studied twice than for words studied once, $F(1, 49) = 50.83, p < .001, \eta_p^2 = .51$. Concerning font size, a significant main effect revealed higher JOLs for words presented in a large font than for words presented in a small font, $F(1, 49) = 17.74, p < .001, \eta_p^2 = .27$. A significant main effect of concreteness revealed higher JOLs for concrete words than for abstract words, $F(1, 49) = 45.76, p < .001, \eta_p^2 = .48$. Finally, a significant main effect of emotionality revealed higher JOLs for emotional words than for neutral words, $F(1, 49) = 80.48, p < .001, \eta_p^2 = .62$. No interactions were significant, all F s < 3.10 .

A 2 (number of study presentations) $\times$ 2 (font size) $\times$ 2 (concreteness) $\times$ 2 (emotionality) ANOVA on recall performance revealed that actual memory was significantly better for words studied twice than for words studied once, $F(1, 48)^3 = 193.98, p < .001, \eta_p^2 = .80$. A significant main effect of font size revealed better memory for words presented in a large font than for words presented in a small font, $F(1, 48) = 8.98, p = .004, \eta_p^2 = .16$. Concerning concreteness, memory was better for concrete words than for abstract words, $F(1, 48) = 51.83, p < .001, \eta_p^2 = .52$. Finally, a significant main effect of emotionality revealed better memory for emotional words than for neutral words, $F(1, 48) = 11.70, p = .001, \eta_p^2 = .20$. No other effects were significant, all F s < 2.96 .

An individual-level analysis based on simple mean differences suggested that 26 participants (52.00%) integrated all four cues in their JOLs (binomial test: $p < .001$). Eighteen participants (36.00%) integrated three cues in their JOLs (2 participants: number of study presentations, font size, and concreteness; 5 participants: number of study presentations, font size, and emotionality; 6 participants: number of study presentations, concreteness, and emotionality; 5 participants: font size, concreteness, and emotionality). The remaining 6 participants (12.00%) integrated two cues in their JOLs (2 participants: number of study presentations and emotionality; 2 participants: number of study presentations and concreteness; 1 participant: font size and emotionality; 1 participant: concreteness and emotionality). An individual-level analysis of effect sizes with the criterion of $|d| \geq .2$ revealed that 22 participants (44.00%) integrated all four cues, that 17 participants (34.00%) integrated three cues, and that 10 participants (20.00%) integrated two cues. The remaining participant (2.00%) used only one cue for making JOLs.

Discussion

Experiment 3 showed that all four cues affected JOLs and recall performance at the aggregate level. Individual-level analyses revealed that cue integration in JOLs is exceptional.

First, 88.00% (simple mean difference analysis) or 78.00% (effect size analysis) of participants based their JOLs on three or four cues. Moreover, all (or all but one) participants integrated at least two cues in their JOLs.

However, although these impressive results apparently demonstrate a high capacity for cue integration, two issues may be raised. First, we manipulated each cue in two easily distinguishable levels. There is a danger, then, that judgments are responses to the demand characteristics of the situation rather than reflections of the psychological variable of interest (Orne, 1962). This means that participants might base their JOLs on a plausible ad hoc hypothesis about the pattern of memory predictions that the experimenter expects (for the impact of ad hoc theories on JOLs, see Mueller & Dunlosky, 2017). Obviously, this is of concern not only for the current experiments but also for other JOL studies. When compared to experiments that vary a single cue, our Experiment 3 alleviated this concern because simultaneously manipulating four cues probably reduced the saliency of each cue. Nevertheless, dichotomous cues might have triggered demand characteristics.

Moreover, in Experiment 3, the selection of words was nonrepresentative in that we (a) excluded words of intermediate concreteness and (b) varied concreteness and emotionality orthogonally. Orthogonal manipulations of stimulus features destroy the correlational structure of the environment. From a Brunswikian perspective, this may produce processes different from those in a natural domain (e.g., Dhimi, Hertwig, & Hoffrage, 2004). For instance, strong positive correlations between cues render some information redundant and enhance the accuracy of one-cue judgment strategies, whereas negative correlations increase conflict and boost compensatory decision making (e.g., Bettman, Johnson, Luce, & Payne, 1993; Fasolo, McClelland, & Lange, 2005). Hence, it is worthwhile to see whether cue integration would also occur in a representative design. Thus, Experiment 4 investigated cue integration in JOLs with nonorthogonal cues that vary on a continuum.

First, though, an unexpected result from Experiment 3 deserves comment: Words presented in a large font were better recalled than words presented in a small font (for a similar finding, see Price et al. 2016). However, this finding did not replicate in Experiments 1 or 4 and thus is not considered further.

Experiment 4

In Experiment 4, words were presented in eight different font sizes between 18 point and 48 point. We selected a representative sample of words that varied across a wide range of concreteness and emotionality. Thus, Experiment 4 assessed whether or not participants base their JOLs on multiple nonorthogonal cues that vary on a continuum. Individual differences were taken into account by using multilevel regression

³ One recall sheet was lost before recall performance was coded.

models that estimated subject-to-subject variation in mean JOLs or recall and in the impact of font size, concreteness, and emotionality on JOLs or recall. To facilitate comparisons across experiments, we also report individual-level analyses that are parallel to those provided for Experiments 1 to 3.

Method

Participants and materials Participants were 48 University of Mannheim undergraduates. We selected a representative sample of 5–10 letter nouns from Võ et al. (2009). To this end, we first divided all nouns into six levels of concreteness (i.e., the one sixth of nouns with lowest concreteness, one sixth of nouns with next lowest concreteness, etc.). Second, we divided all nouns into six levels of arousal (i.e., the one sixth of nouns with lowest arousal, one sixth of nouns with next lowest arousal, etc.). We then selected 64 words while ensuring that the percentage of words in each combination of concreteness and arousal level matched the respective percentage in the Võ et al. (2009) word norms. Mean values of concreteness and arousal and the concreteness–arousal correlation were similar for the selected words and for all words (concreteness: 4.25 vs. 4.29, arousal: 2.76 vs. 2.78, correlation: $-.10$ vs. $-.12$). Four additional words served as primacy buffers and were not included in the analysis.

Procedure The procedure was the same as in Experiment 3, except that all words were presented once and that, for each participant, eight randomly chosen words each were presented in font sizes of 18, 21, 24, 27, 31, 36, 41, and 48 point.

Results and discussion

We used multilevel regression (R packages lme4 and lmerTest; Bates, Maechler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2016; R Core Team, 2016) to evaluate the impact of concreteness, emotionality, and font size on JOLs and recall performance. We specified random intercepts for participants and uncorrelated random effects for concreteness, emotionality, font size, and their interactions. All predictors were centered. Recall performance was modeled with a logistic regression model.

Regressing JOLs on font size, concreteness, emotionality, and their interactions revealed significantly positive unstandardized regression coefficients for all three predictors (see Fig. 6), font size: $t(49) = 4.85$, $p < .001$; concreteness: $t(48) = 6.61$, $p < .001$; emotionality: $t(49) = 5.96$, $p < .001$. A significant interaction between concreteness and emotionality indicated that the impact of emotionality on JOLs decreased with increasing concreteness, $t(2833) = 2.55$, $p = .011$. No other interactions were significant, all $ts < 1$. To follow up the Concreteness $\times$ Emotionality interaction, words were divided at the median of concreteness. Separate regression models

Fig. 6 Fixed effects regression weights b for judgments of learning (JOLs) and recall performance (recall) for font size, concreteness (conc), emotionality (emo), and the interaction between concreteness and emotionality (Conc $\times$ Emo) in Experiment 4. Error bars represent one standard error. * $p < .05$. *** $p < .001$

revealed that emotionality increased JOLs for words with low concreteness, $b = 5.99$ ($SE = 0.98$), $t(48) = 6.09$, $p < .001$, and, to a lesser extent, for words with high concreteness, $b = 2.70$ ($SE = 0.79$), $t(48) = 3.42$, $p = .001$. In sum, JOLs increased with increasing font size, concreteness, and emotionality.

A multilevel logistic regression of recall performance revealed that both concreteness and emotionality but not font size affected actual memory, concreteness: $z = 6.04$, $p < .001$; emotionality: $z = 3.47$, $p < .001$; font size: $z = 1.54$, $p = .123$. On average, each one-unit increase in concreteness increased odds of recall by 1.24 times, and each one-unit increase in emotionality increased odds of recall by 1.27 times. No other effects were significant.

As an analog of the individual-level analyses based on simple mean differences, we coded participants as having based JOLs on a particular cue if that cue revealed a positive regression weight in a multiple linear regression predicting their JOLs from all three cues. Results revealed that 29 participants (60.42%) integrated all three cues in their JOLs (binomial test: $p < .001$). Thirteen participants (27.08%) integrated two cues in their JOLs (5 participants: concreteness and emotionality; 2 participants: font size and emotionality; 6 participants: font size and concreteness). Five participants (10.42%)

based their JOLs on only one cue (2 participants respectively: font size, emotionality; 1 participant: concreteness). The remaining participant used none of the cues. For the individual analysis based on effect size, we tested each participant's standardized regression weight for the respective cue against Cohen's effect size convention for small effects in measures of association ($|r| \geq .10$). According to this criterion, 9 participants (18.75%) integrated all three cues, and 16 participants (33.33%) integrated two cues, whereas 13 participants (27.08%) used only one cue, and 10 participants (20.83%) used none of the cues.

In Experiment 4, the majority of participants integrated two or three cues in their JOLs. This showed that people can integrate nonorthogonal cues that vary on a continuum. However, the number of participants that integrated all available cues in their JOLs was lower in Experiment 4 than in Experiment 3 (according to both classification criteria), indicating that nonorthogonal cues that vary on a continuum are more difficult to integrate than orthogonal cues that vary in two discrete levels. Nevertheless, Experiment 4 showed that, in representative designs, additive integration of multiple cues in JOLs occurs.

Our finding that emotionality affected JOLs when manipulated on a continuum is somewhat at odds with Hourihan et al.'s (2017) finding that JOLs were sensitive to word frequency but insensitive to valence and arousal when all three cues were manipulated on a continuum. Importantly, in our Experiment 4, effects of font size, concreteness, and emotionality on JOLs remained when we included word frequency in the analysis. Given that ranges of word frequency were comparable in Hourihan et al.'s and in our experiment, differences in results are probably due the fact that Hourihan et al. decomposed emotionality into arousal and valence. Maybe, higher salience of emotionality as a whole fosters its integration in JOLs (for salience effects on JOLs, see Castel, McCabe, & Roediger, 2007; Dunlosky & Matvey, 2001; Koriati et al., 2014). However, regardless of the specific reasons for the discrepancy in findings, our results are compatible with Hourihan et al.'s conclusion that emotionality affects JOLs through a cognitive rather than a physiological mechanism. Similarly, our conclusion that emotionality can affect JOLs when manipulated on a continuum holds in the light of Hourihan et al.'s findings.

General discussion

In the current study, we investigated whether and to what degree multiple cues combine to affect JOLs. The question of cue integration has been a major focus in judgment research, but has received relatively little attention in metacognition research. In fact, it was mostly tangential to the focus of prior JOL studies (for notable exceptions, see Hertzog et al., 2013; Hines et al., 2015; Tauber & Rhodes, 2012). In four experiments, we simultaneously varied up to four intrinsic

and extrinsic cues of diverse validities in both orthogonal (Experiments 1–3) and representative designs (Experiment 4).

At the aggregate level, every single cue that was manipulated affected JOLs. This pattern of results closely mirrored findings from JOL experiments that manipulated each cue in isolation. In contrast, only three of the cues (number of study presentations, concreteness, and emotionality) consistently affected recall performance. The remaining cue, font size, affected recall performance in Experiment 3 but not in Experiments 1 or 4. With the exception of Price et al. (2016), previous studies found that font size did not affect recall performance (e.g., Kornell et al., 2011; Rhodes & Castel, 2008). This demonstrates that cue integration held regardless of whether cues had similar effects on JOLs and actual memory or not. In the current research, another difference between predicted and actual memory was that concreteness and emotionality interacted to influence JOLs but not recall performance in Experiments 1 and 4.

Crucially, although effects of two or more simultaneously varied cues on JOLs at the aggregate level allude to cue integration, such findings do not strictly warrant the conclusion that individual participants integrate cues in their JOLs. The reason is that significant effects of two or more cues at the aggregate level may also occur if individual participants based their JOLs on only a single cue and different participants based their JOLs on different cues. However, individual-level analyses on the basis of mean differences and effect size measures argued against this possibility. In all experiments, about half of the participants or more based their JOLs on two or more cues, depending on the criterion used to classify participants. One important finding of the present study was that increasing the number of manipulated cues from two to four did not impair cue integration in JOLs. In contrast, a representative design with nonorthogonal cues that varied on a continuum reduced cue integration in JOLs as compared to orthogonal designs with only two cue levels.

In addition to demonstrating that metacognitive judgments integrate information from multiple cues, the present study has relevance for other theoretical issues. First, when manipulating a single cue in two easily distinguishable levels, strong demand characteristics might threaten the validity of JOLs. Specifically, cue effects on JOLs might be based on participants' hypotheses about what pattern of recall predictions would make a success of the experiment (Orne, 1962; see also Mueller & Dunlosky, 2017). The current Experiment 3 alleviates this concern in demonstrating that effects of four different cues on JOLs persisted when manipulated in the context of multiple varying cues. Even more compelling evidence came from Experiment 4, in which three nonorthogonal cues that varied on a continuum were integrated in JOLs. Although we cannot rule out the possibility that demand effects played a role in Experiments 3 and 4, we think that varying multiple cues greatly reduces the danger of demand effects on JOLs.

As a related point, the availability of multiple cues may affect the extent to which explicit beliefs about memory govern JOLs. Some previous studies that manipulated single cues such as concreteness (Witherby & Tauber, 2016) or font size (Mueller et al., 2014) have found that metamemory beliefs were the sole basis of JOLs (but see, e.g., Besken & Mulligan, 2013; Undorf & Erdfelder, 2015; Undorf & Zander, 2017; Undorf, Zimdahl, & Bernstein, 2017). These findings support the notion that people deliberately search for variability across items and base their JOLs on activated or newly formed beliefs about how item characteristics or experimental manipulations may affect memory (analytic processing theory; Dunlosky & Tauber, 2014; Mueller et al., 2013). It remains to be seen how analytic processing theory may fare in experiments with multiple varying cues, in which relevant beliefs are probably harder to activate or develop. Maybe, the availability of multiple cues fosters the reliance of JOLs on nonanalytic, experience-based processes such as fluency (Koriat, 1997; see also Undorf & Erdfelder, 2015). This speculation awaits further research.

Finally, our experiments touched on the idea that JOLs are sensitive to intrinsic cues but insensitive to extrinsic cues (Koriat, 1997). The finding that JOLs equally integrated intrinsic and extrinsic cues converges with previous results that questioned this hypothesis (Dunlosky & Matvey, 2001; Jang & Nelson, 2005).

The current study clearly demonstrated that people have the ability to integrate multiple cues in JOLs. Still, much remains to be learned about cue integration in metacognitive judgments. For instance, it will need to be explored whether integrating multiple cues fosters JOL accuracy. In the current study, we found a positive correlation between resolution and the number of cues used as bases for JOLs in Experiment 4 ($r = .34, p = .017, d = 0.37$). In Experiments 1 ($r = .25, p = .08, d = 0.26$), 2 ($r = .04, p = .752, d = 0.04$), and 3 ($r = .21, p = .152, d = 0.21$), however, correlations were numerically positive but not statistically different from zero. Clearly, more research is needed determine whether cue integration has positive effects on JOL accuracy.⁴ Also, it would be worthwhile to more thoroughly examine the limits of cue integration in JOLs. Moreover, it remains to be determined what factors influence how much weight people give to individual cues when integrating multiple cues in their JOLs. Finally, future research may explore how task features and individual differences impact on cue integration in JOLs.

In summary, the current work demonstrated that people integrate multiple cues in JOLs. Theoretically, these findings are relevant for understanding how JOLs are formed. Practically, the present results advance knowledge about metacognition in everyday learning, where multiple potentially relevant cues are available. Also, they demonstrate parallels between

metacognitive judgments and judgments about the external world and may therefore contribute to developing an integrative theoretical view of cognition and metacognition.

Acknowledgements This research was supported by a starter grant from the University of Mannheim to the first and second authors. Preparation of this article was completed in part while the first author was a visiting scholar in the Faculty of Industrial Engineering and Management of the Technion–Israel Institute of Technology, Haifa, Israel. We thank Rakefet Ackerman, Edgar Erdfelder, Ido Erev, and Asher Koriat for helpful comments on this research.

Appendix

Table 1 shows JOLs from the first study presentation of words that were studied twice in Experiments 1 and 3. Comparing Table 1 to Figs. 1 and 5 reveals that these JOLs were very similar to JOLs for once-studied words in both experiments.

In Experiment 1, first JOLs for twice-studied words were higher for words presented in a large font than for words presented in a small font, $t(52) = 3.37, p = .001, d = 0.47$.

In Experiment 3, first JOLs for twice-studied words were submitted to an ANOVA with font size, concreteness, and emotionality as within-subjects factors. A significant main effect of font size revealed higher JOLs for words presented in a large font than for words presented in a small font, $F(1, 49) = 7.97, p = .007, \eta_p^2 = .14$. A significant main effect of concreteness revealed higher JOLs for concrete words than for abstract words, $F(1, 49) = 33.89, p < .001, \eta_p^2 = .41$. A significant main effect of emotionality revealed higher JOLs for emotional words than for neutral words, $F(1, 49) = 62.63, p < .001, \eta_p^2 = .56$. No interactions were significant, all F s < 1.

Table 1 Means and standard deviations of judgments of learning from the first study presentations of twice-studied words in Experiments 1 and 3

Experiment				<i>M</i>	<i>SD</i>	
1	Font size	18		35.60	13.74	
		48		40.12	13.75	
3	Font size	18	Abstract	Neutral	23.97	13.75
				Emotional	35.23	16.25
			Concrete	Neutral	33.44	17.40
				Emotional	41.66	19.93
	48	Abstract	Neutral	31.58	18.09	
			Emotional	38.34	16.40	
		Concrete	Neutral	36.35	20.12	
			Emotional	43.61	17.54	

⁴ We thank an anonymous reviewer for suggesting this analysis.

References

- Anderson, N. H. (1981). *Foundations of information integration theory*. New York, NY: Academic Press.
- Anderson, N. H. (2013). Unified psychology based on three laws of information integration. *Review of General Psychology*, 17(2), 125–132. <https://doi.org/10.1037/a0032921>
- 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>
- Begg, I. M., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, 28(5), 610–632. [https://doi.org/10.1016/0749-596X\(89\)90016-8](https://doi.org/10.1016/0749-596X(89)90016-8)
- Benjamin, A. S. (2005). Response speeding mediates the contributions of cue familiarity and target retrievability to metamnemonic judgments. *Psychonomic Bulletin & Review*, 12(5), 874–879. <https://doi.org/10.3758/BF03196779>
- Besken, M. (2016). Picture-perfect is not perfect for metamemory: Testing the perceptual fluency hypothesis with degraded images. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9), 1417–1433. <https://doi.org/10.1037/xlm0000246>
- 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>
- Bettman, J. R., Johnson, E. J., Luce, M. F., & Payne, J. W. (1993). Correlation, conflict, and choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(4), 931–951. <https://doi.org/10.1037/0278-7393.19.4.931>
- Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica*, 87(2/3), 137–154. [https://doi.org/10.1016/0001-6918\(94\)90048-5](https://doi.org/10.1016/0001-6918(94)90048-5)
- Bröder, A. (2000). A methodological comment on behavioral decision research. *Psychologische Beiträge*, 42(4), 645–662.
- Brunswik, E. (1944). Distal focussing of perception: Size-constancy in a representative sample of situations. *Psychological Monographs*, 56(1), i–49. <https://doi.org/10.1037/h0093505>
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193–217. <https://doi.org/10.1037/h0047470>
- Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review*, 14(1), 107–111. <https://doi.org/10.3758/BF03194036>
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.
- Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, 130(6), 959–988. <https://doi.org/10.1037/0033-2909.130.6.959>
- Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic-extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(5), 1180–1191. <https://doi.org/10.1037/0278-7393.27.5.1180>
- Dunlosky, J., & Tauber, S. K. (2014). Understanding people's metacognitive judgments: An isomechanism framework and its implications for applied and theoretical research. In T. J. Perfect & D. S. Lindsay (Eds.), *The Sage handbook of applied memory* (pp. 444–464). Los Angeles, CA: Sage.
- Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, 86(5), 465–485. <https://doi.org/10.1037/0033-295X.86.5.465>
- Fasolo, B., McClelland, G. H., & Lange, K. A. (2005). The effect of site design and interattribute correlations on interactive web-based decisions. In C. P. Haugtvedt, K. A. Machleit, & R. F. Yalch (Eds.), *Online consumer psychology: Understanding and influencing consumer behavior in the virtual world* (pp. 325–342). Mahwah, NJ: Erlbaum.
- Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1), 107–143. <https://doi.org/10.1111/j.1756-8765.2008.01006.x>
- Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York, NY: Oxford University Press.
- Hertzog, C., Hines, J. C., & Touron, D. R. (2013). Judgments of learning are influenced by multiple cues in addition to memory for past test accuracy. *Archives of Scientific Psychology*, 1(1), 23–32. <https://doi.org/10.1037/arc0000003>
- Hines, J. C., Hertzog, C., & Touron, D. R. (2015). Younger and older adults weigh multiple cues in a similar manner to generate judgments of learning. *Aging, Neuropsychology, and Cognition*, 22(6), 693–711. <https://doi.org/10.1080/13825585.2015.1028884>
- Hourihaan, K. L., Fraundorf, S. H., & Benjamin, A. S. (2017). The influences of valence and arousal on judgments of learning and on recall. *Memory & Cognition*, 45(1), 121–136. <https://doi.org/10.3758/s13421-016-0646-3>
- Illman, N. A., & Morrison, C. M. (2011). The role of age of acquisition in memory: Effects on judgements of learning and recall. *Quarterly Journal of Experimental Psychology*, 64(9), 1665–1671. <https://doi.org/10.1080/17470218.2011.591495>
- Jang, Y., & Nelson, T. O. (2005). How many dimensions underlie judgments of learning and recall? Evidence from state-trace methodology. *Journal of Experimental Psychology: General*, 134(3), 308–326. <https://doi.org/10.1037/0096-3445.134.3.308>
- Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, 134(3), 404–426. <https://doi.org/10.1037/0033-2909.134.3.404>
- 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. <https://doi.org/10.1037/0096-3445.126.4.349>
- Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *Cambridge handbook of consciousness* (pp. 289–325). New York, NY: Cambridge University Press.
- Koriat, A. (2015). Metacognition: Decision making processes in self-monitoring and self-regulation. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 356–379). Malden, MA: Wiley-Blackwell.
- Koriat, A., Ackerman, R., Adiv, S., Lockl, K., & Schneider, W. (2014). The effects of goal-driven and data-driven regulation on metacognitive monitoring during learning: A developmental perspective. *Journal of Experimental Psychology: General*, 143(1), 386–403. <https://doi.org/10.1037/a0031768>
- Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, 133(4), 643–656. <https://doi.org/10.1037/0096-3445.133.4.643>
- Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, 22(6), 787–794. <https://doi.org/10.1177/0956797611407929>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). *lmerTest: Tests in linear mixed effects models* (R Package Version 2.0-33) [Computer software]. Retrieved from <https://CRAN.R-project.org/package=lmerTest>
- Magreehan, D. A., Serra, M. J., Schwartz, N. H., & Narciss, S. (2016). Further boundary conditions for the effects of perceptual disfluency

- on judgments of learning. *Metacognition and Learning*, 11(1), 35–56. <https://doi.org/10.1007/s11409-015-9147-1>
- Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, 52(1), 29–71. <https://doi.org/10.1023/A:1015516217425>
- Metcalf, J., & Finn, B. (2008). Familiarity and retrieval processes in delayed judgments of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(5), 1084–1097. <https://doi.org/10.1037/a0012580>
- Mueller, M. L., & Dunlosky, J. (2017). How beliefs can impact judgments of learning: Evaluating analytic processing theory with beliefs about fluency. *Journal of Memory and Language*, 93, 245–258. <https://doi.org/10.1016/j.jml.2016.10.008>
- Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language*, 70(1), 1–12. <https://doi.org/10.1016/j.jml.2013.09.007>
- Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, 20(2), 378–384. <https://doi.org/10.3758/s13423-012-0343-6>
- Orme, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, 17(11), 776–783. <https://doi.org/10.1037/h0043424>
- Pachur, T., & Bröder, A. (2013). Judgment: A cognitive processing perspective. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(6), 665–681. <https://doi.org/10.1002/wcs.1259>
- Peynircioglu, Z. F., Brandler, B. J., Hohman, T. J., & Knutson, N. (2014). Metacognitive judgments in music performance. *Psychology of Music*, 42(5), 748–762. <https://doi.org/10.1177/0305735613491999>
- Platzer, C., Bröder, A., & Heck, D. W. (2014). Deciding with the eye: How the visually manipulated accessibility of information in memory influences decision behavior. *Memory & Cognition*, 42(4), 595–608. <https://doi.org/10.3758/s13421-013-0380-z>
- Price, J., & Harrison, A. (2017). Examining what prestudy and immediate judgments of learning reveal about the bases of metamemory judgments. *Journal of Memory and Language*, 94, 177–194. <https://doi.org/10.1016/j.jml.2016.12.003>
- Price, J., McElroy, K., & Martin, N. J. (2016). The role of font size and font style in younger and older adults' predicted and actual recall performance. *Aging, Neuropsychology, and Cognition*, 23(3), 366–388. <https://doi.org/10.1080/13825585.2015.1102194>
- Pyc, M. A., & Rawson, K. A. (2012). Are judgments of learning made after correct responses during retrieval practice sensitive to lag and criterion level effects? *Memory & Cognition*, 40(6), 976–988. <https://doi.org/10.3758/s13421-012-0200-x>
- R Core Team. (2016). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.
- Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 65–80). New York, NY: Oxford University Press.
- Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, 137(4), 615–625. <https://doi.org/10.1037/a0013684>
- Soderstrom, N. C., & McCabe, D. P. (2011). The interplay between value and relatedness as bases for metacognitive monitoring and control: Evidence for agenda-based monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5), 1236–1242. <https://doi.org/10.1037/a0023548>
- Son, L. K., & Simon, D. A. (2012). Distributed learning: Data, metacognition, and educational implications. *Educational Psychology Review*, 24(3), 379–399. <https://doi.org/10.1007/s10648-012-9206-y>
- Susser, J. A., & Mulligan, N. W. (2015). The effect of motoric fluency on metamemory. *Psychonomic Bulletin & Review*, 22(4), 1014–1019. <https://doi.org/10.3758/s13423-014-0768-1>
- Tauber, S. K., & Dunlosky, J. (2012). Can older adults accurately judge their learning of emotional information? *Psychology and Aging*, 27(4), 924–933. <https://doi.org/10.1037/a0028447>
- Tauber, S. K., & Rhodes, M. G. (2010). Does the amount of material to be remembered influence judgements of learning (JOLs)? *Memory*, 18(3), 351–362. <https://doi.org/10.1080/09658211003662755>
- Tauber, S. K., & Rhodes, M. G. (2012). Multiple bases for young and older adults' judgments of learning in multitrial learning. *Psychology and Aging*, 27(2), 474–483. <https://doi.org/10.1037/a0025246>
- Todd, P. M., Gigerenzer, G., & the ABC Research Group. (2012). *Ecological rationality: Intelligence in the world*. New York, NY: Oxford University Press.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Undorf, M., & Ackerman, R. (2017). The puzzle of study time allocation for the most challenging items. *Psychonomic Bulletin & Review*, 24(6), 2003–2011. <https://doi.org/10.3758/s13423-017-1261-4>
- Undorf, M., & Erdfelder, E. (2013). Separation of encoding fluency and item difficulty effects on judgements of learning. *The Quarterly Journal of Experimental Psychology*, 66(10), 2060–2072. <https://doi.org/10.1080/17470218.2013.777751>
- Undorf, M., & Erdfelder, E. (2015). The relatedness effect on judgments of learning: A closer look at the contribution of processing fluency. *Memory & Cognition*, 43(4), 647–658. <https://doi.org/10.3758/s13421-014-0479-x>
- Undorf, M., & Zander, T. (2017). Intuition and metacognition: The effect of semantic coherence on judgments of learning. *Psychonomic Bulletin & Review*, 24(4), 1217–1224. <https://doi.org/10.3758/s13423-016-1189-0>
- Undorf, M., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language*, 92, 293–304. <https://doi.org/10.1016/j.jml.2016.07.003>
- Vö, M. L.-H., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods*, 41(2), 534–538. <https://doi.org/10.3758/BRM.41.2.534>
- Witherby, A. E., & Tauber, S. K. (2016). The concreteness effect on judgments of learning: Evaluating the contributions of fluency and beliefs. *Memory & Cognition*, 43(4), 639–650. <https://doi.org/10.3758/s13421-016-0681-0>
- Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society*, 15(1), 41–44. <https://doi.org/10.3758/BF03329756>
- Zimmerman, C. A., & Kelley, C. M. (2010). "I'll remember this!" Effects of emotionality on memory predictions versus memory performance. *Journal of Memory and Language*, 62(3), 240–253. <https://doi.org/10.1016/j.jml.2009.11.004>