Metacognition and Learning
https://doi.org/10.1007/s11409-020-09254-4

# How to assess the contributions of processing fluency and beliefs to the formation of judgments of learning: methods and pitfalls

**Chunliang Yang [1] • Rongjun Yu [2] • Xiao Hu [3] • Liang Luo [1,3] • Tina S.-T. Huang [4] • David R. Shanks [4]**

Received: 29 April 2019 / Accepted: 22 November 2020/Published online: 08 January 2021


**Abstract**
Judgments of learning (JOLs) play a fundamental role in helping learners regulate their study strategies but are susceptible to various kinds of illusions and biases. These can potentially impair learning efficiency, and hence understanding the mechanisms underlying the formation of JOLs is important. Many studies have suggested that both processing fluency and metamemory beliefs can contribute substantially to the construction of JOLs. However, in recent years another body of evidence has accumulated apparently demonstrating that beliefs play a dominant role, whereas processing fluency plays little or even no role in JOL formation. In the current article, we review the experimental and analytic methods employed in this field to measure the contributions of processing fluency and beliefs to the formation of JOLs. We then illustrate several potential disadvantages and pitfalls of those research methods. Suggestions about how to solve or avoid such problems are discussed. We make several proposals for future research to shed additional light on the illusions and biases that have been documented in JOLs.

**Keywords** Judgments of learning · Processing fluency · Metamemory beliefs · Experimental and analytic methods

✉ Chunliang Yang
chunliang.yang@bnu.edu.cn

[1] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University, 19 Xinjiekouwai Street, Haidian District, Beijing 100875, China

[2] Department of Management, School of Business, Hong Kong Baptist University, Hongkong, China

[3] Collaborative Innovation Center of Assessment Toward Basic Education Quality, Beijing Normal University, Beijing, China

[4] Division of Psychology and Language Sciences, University College London, London, UK

Judgments of learning (JOLs; metacognitive estimates of the likelihood that a given item will be remembered on a future occasion) are susceptible to various kinds of illusions and biases (e.g., Rhodes and Castel 2008; Yang et al. 2018a). Given that JOLs likely play a causal role in study strategy regulation (Kornell and Bjork 2008; Li et al. 2015), understanding the mechanisms responsible for the formation of JOLs is of considerable importance for exploring practical interventions to calibrate JOLs, reduce biases, and optimize learning efficiency. Accordingly, shedding light on the mechanisms underlying the construction of JOLs is a major goal for researchers and educators (e.g., Frank and Kuhlmann 2016; Koriat 1997; Mueller, Dunlosky and Tauber 2016; Yang et al. 2018a).

Previous studies have employed diverse experimental and analytic methods to investigate the roles of processing fluency (an experience-based cue) and metamemory beliefs (a theory-based cue) in JOL formation. The research findings derived from different studies are inconsistent (e.g., Mueller et al. 2014; Mueller et al. 2013; Undorf and Erdfelder 2014; Yang et al. 2018a).

The current review aims to: (1) briefly introduce some of the key empirical findings documented by the relevant studies; (2) summarize the experimental and analytic methods employed by this body of studies; (3) highlight several disadvantages and pitfalls of those methods; and finally (4) offer some suggestions regarding how to solve or mitigate those problems in future research.

## Two models of the relative contributions of processing fluency and beliefs in JOL formation

Processing fluency (i.e., the ease with which a given item is mentally processed) plays an important role in several domains of metacognitive judgment. For instance, more fluently processed items are more likely to be judged true (Schwarz and Reber 1999) and grammatically correct (Kinder et al. 2003); stimuli against highly contrasting backgrounds are judged more likable (Reber et al. 2004); and words presented in easy-to-read fonts are rated as more familiar (Reber and Zupanek 2002) (for a review of fluency effects on metacognitive judgments, see Alter and Oppenheimer 2009). Along the same lines, processing fluency feeds into JOLs: fluently processed items are judged more likely to be remembered than less fluently processed ones – the fluency effect on JOLs, a phenomenon documented in the bulk of the literature (e.g., Besken 2016; Dunlosky et al. 2014; Hertzog et al. 2003; Undorf and Erdfelder 2011; Undorf et al. 2017; Yang et al. 2018a). (The Appendix provides a brief summary of the possible ways in which processing fluency can influence JOLs. It explains that although processing fluency is typically regarded as an experienced (non-analytic) cue by comparison with the analytic basis of beliefs, there are open questions regarding how "experienced fluency" contributes to JOL formation and whether some of that influence might in fact be analytic.)

Besides processing fluency, people's metamemory beliefs (that is, beliefs about how memory operates) also contribute substantially to the construction of JOLs. For example, people tend to believe that words presented in large fonts (e.g., 48-pt) are easier to remember than ones in small fonts (e.g., 18-pt), and they apply this belief when forming JOLs by giving higher learning judgments to large than to small words (Hu et al. 2015; Mueller et al. 2014; Rhodes and Castel 2008). Similarly, people tend to believe that concrete words (e.g., *apple*) are more likely to be remembered than abstract ones (e.g., *loyalty*), and apply this belief so as to give higher JOLs to concrete than to abstract words (Witherby and Tauber 2017a). There are dozens of studies observing an influence of such beliefs on JOLs, such as the effects of beliefs about volume, animacy, and relatedness (e.g., Frank and Kuhlmann 2016; Jia et al. 2015; Li et al. 2016; Mueller et al. 2013, 2014, 2016; Susser and Mulligan 2015; Undorf and Erdfelder 2014).

Koriat's (1997) *dual-basis* model assumes that metacognitive judgments (such as JOLs and subjective confidence ratings) are formed through implicitly applying heuristics to a holistic range of cues in order to infer future memory performance. These cues include the subjective experience of performing the task (e.g., processing fluency) as well as beliefs about how various factors (e.g., stimulus features, encoding procedures, test formats, one's own memory ability, etc.) affect memory. Put differently, the model hypothesizes that both subjective processing experience (processing fluency) and theory-based cues (beliefs) can play important roles in JOL formation. Nonetheless, this model has been met with some skepticism on the basis of recent research on underlying causal processes that has been taken to suggest that processing fluency contributes much less to JOL formation than beliefs (e.g., Jia et al. 2015; Li et al. 2016; Mueller et al. 2014, 2016; Witherby and Tauber 2017a).

Witherby and Tauber (2017a), for instance, found that word concreteness drives JOLs mainly through metamemory beliefs (i.e., the belief that concrete words are easier to remember than abstract words) rather than processing fluency. They showed, specifically, that participants predicted better recall of concrete than abstract words when given a description of a hypothetical experiment. Because no words were actually studied, fluency cannot have played any role in this effect. Similarly, both font size (Mueller et al. 2014) and semantic relatedness (Mueller et al. 2013) were found to affect JOLs mainly via beliefs (i.e., people believe that large words or semantically related word pairs are easier to remember than small words or unrelated pairs), instead of processing fluency. There are many other recent studies providing evidence that various factors (e.g., concreteness, word frequency) affect JOLs mainly through beliefs instead of processing fluency (e.g., Jia et al. 2015; Li et al. 2016; Mueller et al. 2016).

These findings jointly support an *analytic processing* model, recently proposed by Mueller et al. (2016), which hypothesizes that JOLs are predominantly driven by beliefs, whereas processing fluency plays little or even no role in JOL formation. Specifically, the analytic processing model assumes that learners exert conscious control over the factors that might enter into metacognitive judgments: they explicitly monitor features of the stimulus or task to search for cues that could be diagnostic for making rational judgments, they then form beliefs based on the anticipated memorial impact of those cues, and subsequently apply those beliefs to make judgments (Kelley and Jacoby 1990). According to Mueller and Dunlosky (2017, p. 246):

> The critical new twists to AP [analytic processing] theory are that it emphasizes (a) that people first explicitly search for cues that will allow them to reduce their uncertainty in predicting future memory performance and (b) that people will develop beliefs on-line – as they are participating in an experiment – about how different variables may help them to accurately predict performance. Either these newly formed beliefs or a priori beliefs will in part drive JOLs… Importantly, AP theory does not rule out the contribution of processing fluency to JOLs. If people do not construct beliefs (or retrieve a priori ones) relevant to the prediction context, then the subjective experience of fluency that differs across items may influence JOLs. However, in contrast to other dual-process models of JOLs (e.g., cue-utilization framework, Koriat 1997), AP theory emphasizes the

dominant role of beliefs in constructing JOLs and provides a description of processes for how beliefs may be developed and influence JOLs.

Because the analytic processing model does not reject the possible contribution of processing fluency, it would be erroneous to regard the dual-basis and analytic processing models as being in opposition. Indeed the analytic processing model can be regarded as a subclass of dual-basis models in which the contribution of processing fluency is significantly attenuated (for a detailed comparison of these two models, see Mueller and Dunlosky 2017).

Researchers have adopted various methods to explore how a given factor (e.g., font size, volume, relatedness, concreteness) affects JOLs: Does it affect JOLs through processing fluency or through beliefs (or a combination of both)? To offer an overview of the empirical findings, we summarise them in Table 1, which includes the majority (if not all) of the JOL phenomena that researchers have studied to reveal the roles of processing fluency and beliefs in JOL formation.[1] As shown in Table 1, previous studies have demonstrated that beliefs consistently contribute to JOLs in the majority of cases, whereas only in a minority are JOLs affected by both processing fluency and beliefs. We warn readers however to interpret the research findings listed in Table 1 cautiously because, as we will illustrate below, some conclusions reached in previous research are subject to a number of significant problems.

In this article we review the research methods used in previous studies, discuss their potential disadvantages, and attempt to offer some suggestions for future research. At the outset, we acknowledge that several methodological problems may be difficult to settle because of the limitations of current research methods; it is, however, important to highlight them in order to equip future researchers to take those pitfalls into account when developing experimental designs and when interpreting their research findings.

For the sake of exposition and given that the main topic of the current review is methodological, we organize the following sections by research methods (i.e., discussing the problems method-by-method) instead of by studies (i.e., discussing the problems study-by-study). It should be emphasized, however, that some (although certainly not all) previous studies employed several different experimental methods to evaluate the roles of processing fluency and beliefs in the construction of JOLs (e.g., Mueller et al. 2014; Witherby and Tauber 2017a). In doing so they based their major conclusions on convergent evidence as well as on the findings of individual experiments.

## Experimental methods for measuring processing fluency

There are different types of processing fluency which may affect JOLs, such as conceptual fluency (ease of accessing a given item's conceptual meaning), perceptual fluency (ease of perceiving an item), imaginative fluency (ease of forming a mental image to represent an item), naming fluency (ease of naming), retrieval fluency (ease of retrieval from memory), and so on. A variety of experimental methods have been employed to explore the influence of processing fluency on JOLs including but not limited to: lexical decision, self-regulated study time

[1] There are other factors assumed to affect JOLs through processing fluency and/or beliefs, such as instructor fluency (Carpenter et al. 2013), perceptual interference (Besken and Mulligan 2013), task experience (Dunlosky and Hertzog 2000), and so on. Critically, authors have assumed but not experimentally documented that processing fluency and/or beliefs contribute to the effects of those factors on JOLs. Without direct tests, it is premature to draw such inferences, and hence they are not included in the table.

**Table 1** Empirical findings from previous studies on the roles of processing fluency and beliefs in the formation of JOLs

| Factors | JOL phenomena | Beliefs? | Fluency? | Sample references |
|---|---|---|---|---|
| Font size | Higher JOLs are given to large (48-pt) than to small (18-pt) words, despite font size having little effect on recall. | Yes | Yes | (Hu et al. 2015; Rhodes and Castel 2008; Yang et al. 2018a) |
| Relatedness | Higher JOLs are given to semantically related word pairs (e.g., *dog-cat*) than to unrelated pairs (e.g., *box-head*), and related pairs are better recalled than unrelated ones. | Yes | Yes | (Mueller et al. 2013; Undorf and Erdfelder 2014) |
| Generation | Higher JOLs are given to intact items (e.g., *rain-umbrella*) than to items requiring a response to be generated (e.g., *door-w_nd_w*?), whereas generated items are equally well or even better recalled than intact ones. | Yes | Yes | (Besken 2016; Froger et al. 2011; Matvey et al. 2001) |
| Handedness | Higher JOLs are given to words written by the dominant hand than to words written by the non-dominant hand, despite hand selection having no effect on recall. | Yes | Yes | (Susser and Mulligan 2015; Susser et al. 2017) |
| Semantic coherence | Higher JOLs are given to coherent triads (i.e., compound remote associates of a single solution word such as *silk–cream–even*, solution: *smooth*) than incoherent triads (i.e., ones with no common associates such as *deck–stool–pocket*), and coherent triads are better recalled. | Yes | Yes | (Undorf and Zander 2017) |
| Errorful generation | Higher JOLs are given to intact word pairs than to errorfully generated pairs (for which people generate incorrect responses and receive corrective feedback), but errorfully generated pairs are better recalled. | Yes | Unknown | (Potts and Shanks 2014; Yang et al. 2017b) |
| Volume | Higher JOLs are given to loud than to quiet words, despite volume having no effect on recall. | Yes | Unknown | (Frank and Kuhlmann 2016; Rhodes and Castel 2009) |
| Emotion | Higher JOLs are given to emotional faces than to neutral ones, whereas emotional and neutral faces are equally well recognized at a later memory test. | Yes | Unknown | (Witherby and Tauber 2017b) |
| Study opportunity | Greater study opportunities enhance memory retention, but people tend to lack awareness of the benefits of multiple study opportunities. | Yes | Unknown | (Ariel et al. 2014) |
| Concreteness | Higher JOLs are given to concrete (e.g., *apple*) than to abstract (e.g., *idea*) words, and concrete words are better recalled than abstract words. | Yes | No | (Witherby and Tauber 2017a) |
| Identity | Higher JOLs are given to identical word pairs (e.g., *dog-dog*) than to semantically related pairs (e.g., *pond-lake*), whereas semantically related word pairs are better recalled. | Yes | No | (Mueller et al. 2016) |
| Word frequency | Higher JOLs are given to high-frequency words (e.g., *apple*) than to low-frequency words (e.g., *caste*), and high-frequency words are better recalled. | Yes | No | (Jia et al. 2015) |

**Table 1** (continued)

| Factors | JOL phenomena | Beliefs? | Fluency? | Sample references |
|---|---|---|---|---|
| Animacy | Higher JOLs are given to animate words (e.g., *dog*) than to inanimate words (e.g., *road*), and animate words are better recalled than inanimate ones. | Yes | No | (Li et al. 2016) |
| Matched priming | Words (e.g., *phone*) which are preceded by matched primes (e.g., *phone*) are given higher JOLs than ones preceded by mismatched primes (e.g., *doctor*), despite primes having no effect on recall. | Unknown | No | (Susser et al. 2016) |
| Age-related memory decline | People believe that memory ability declines with aging in adulthood, but item-by-item JOLs are insensitive to this decline. | No | Unknown | (Tauber et al. 2019) |
| Clarification speed | Higher JOLs are given to quickly clarifying items than to slowly clarifying ones, despite clarification speed having no effect on recall. | No | Yes | (Undorf et al. 2017)) |

"No" means that published research has documented little or no evidence supporting the role of processing fluency or beliefs in a JOL phenomenon; "Unknown" indicates that the role of processing fluency or beliefs has not been experimentally assessed as yet

allocation, continuous identification (CID), retrieval latency, mental image formation, and naming latency. As different experimental methods are sensitive to different types of processing fluency (Ferrand et al. 2011; Yang et al. 2018a), it is essential to employ a sensitive method to assess a given type of processing fluency.

Below we first demonstrate how insensitive methods can lead to mischaracterization of the role of a given type of processing fluency by considering a pair of studies on the "font size" effect: Mueller et al. (2014) and Yang et al. (2018a). Then, we demonstrate that self-regulated study time allocation tasks and study trials tasks, which have been widely-used in previous studies, may at least in some situations be insensitive or invalid instruments for measuring processing fluency.

## Task sensitivity to measure perceptual fluency: CID vs. lexical decision

The font size effect on JOLs refers to the fact that participants reliably judge that words printed in large fonts will be easier to remember than those in small fonts, despite the fact that font size generally has minimal effect on recall. Mueller et al. (2014) employed a lexical decision task in their Experiment 1 to explore the role of perceptual fluency in the font size effect on JOLs. In this task, words (e.g., *computer*) and non-words (e.g., *thate*) were sequentially presented in a random order, and in either a large or small font. Participants' task was to judge, as quickly and accurately as they could, whether the on-screen letter string was a word or non-word. The results revealed no difference in lexical decision RTs between large and small words, whereas higher JOLs were given to large than to small words. Similarly, Undorf and Zimdahl (2019) found a null difference in RTs between 18 pt. and 48 pt. words in their lexical decision task.

It is important to record that Mueller et al. did not entirely reject any role of processing fluency. Instead, they proposed that "processing fluency, as measured by the lexical decision task, is not mediating the font-size effect" (p. 4), and they encouraged future research to explore the potential contributions from other kinds of processing fluency (p. 9). In a recent study, Yang et al. (2018a) proposed that the lexical decision task employed by Mueller et al. might lack sensitivity to measure perceptual fluency.

Previous studies have shown that lexical decisions are not solely driven by *perceptual* processing but that *conceptual* processing is also involved (e.g., Chumbley and Balota 1984). It is worth noting that the first study documenting the font size effect explicitly proposed that the effect may result from "people us[ing] fluently processed perceptual information that is highly accessible at encoding when they make memory predictions" (Rhodes and Castel 2008, p. 624). Hence, it is unclear to what extent Mueller et al.'s results, derived from the lexical decision task, could disapprove the claim that *perceptual* fluency contributes to the font size effect on JOLs.

To further explore this, Yang et al. (2018a) employed a CID task in their Experiment 1 to measure the difference in perceptual fluency between large and small words. In the CID task, a word (e.g., *sheep*) and a mask (#####) were alternately presented, with the duration of the word increasing and that of the mask decreasing across a series of rapid cycles. Thus, the word gradually became easier to perceive as time elapsed. Participants were required to respond as soon as they could identify the word and they then made a JOL following each correct identification. Yang et al. observed that participants identified large words more quickly than small ones and moreover this was a very large effect (Cohen's $d = 1.25$), with 27 out of 28 participants responding faster on average to large than to small words. In addition, a mediation analysis found that font size affected JOLs, at least partially, through its effect on perceptual fluency. In their Experiment 2, Yang et al. directly compared the sensitivity of the CID and lexical decision tasks to perceptual fluency by using the same participants and materials. The results confirmed directly that CID is more sensitive to variations in perceptual fluency than lexical decision (for related findings, see Ferrand et al. 2011; Grainger and Segui 1990).

The above pair of studies (Mueller et al. 2014; Yang et al. 2018a) clearly demonstrate the importance of employing sensitive methods to identify the role of a given type of processing fluency in the formation of JOLs: CID, by comparison with lexical decision, is more sensitive to perceptual fluency, and perceptual fluency does contribute to the font size effect on JOLs. We note that the above results do not imply that lexical decision is completely insensitive to perceptual fluency, merely that lexical decision is measurably less sensitive than CID. It should also be noted that Undorf and Zimdahl (2019) recently documented that when the difference in font size is sufficiently large (e.g., 6 pt. vs. 120 pt), lexical decision can detect a perceptual fluency difference between large and small words.

## Self-regulated study time allocation

We now move to evaluate the validity of the self-regulated time allocation task as a measure of processing fluency. Self-regulated study time allocation has been widely-used to explore the role of processing fluency in the effects of various factors on JOLs, such as word frequency (Jia et al. 2015), animacy (Li et al. 2016), font size (Mueller et al. 2014), concreteness (Witherby and Tauber 2017a), and identity (Mueller et al. 2016). All the aforementioned studies either observed null differences between study times allocated to different types of materials (e.g., high- vs. low-frequency words) or found that the difference in study times did not significantly mediate the factor's effect on JOLs. These studies assumed that self-regulated study time is a measure of processing fluency (i.e., that fluent items can be studied more rapidly) and based their conclusions about processing fluency on this unverified assumption.

The above findings provide little direct support for their conclusions (that is, little or no contribution of processing fluency to their documented JOL phenomena), because the self-regulated study time allocation task is (at least in some situations) a potentially insensitive or invalid measure of processing fluency. Study time allocation is, by definition, a decision, and is known to be affected by a variety of other factors in a goal-driven manner (e.g., motivation, serial position, perceived importance, etc.) besides processing fluency. For instance, Yang et al. (2017a) observed that participants systematically decreased their study times across a study phase in a self-regulated study time allocation task, indicating that, besides processing fluency, serial position manifestly affects time allocation.

Undorf and Ackerman (2017) observed an inverse "U" shaped function relating study times and JOLs in the self-regulated study time allocation task (for related findings, see Koriat et al. 2006; Metcalfe and Kornell 2005). Specifically, Undorf and Ackerman observed that participants allocated less time to study the items with the lowest JOLs (which were assumed to be difficult items and associated with the lowest processing fluency) than to the medium-JOL items (assumed to be medium-difficulty items and associated with medium processing fluency). Metcalfe and Kornell (2005) proposed that participants might have realized that more encoding effort (i.e., longer study times) invested in difficult items produced little or no improvement in their mastery ("*labor in vain*"), and therefore stopped studying these items quickly (and prematurely) and switched to studying the medium-difficulty ones. The inverse "U" shape between JOLs and study times is consistent with the idea that, besides processing fluency, perceived learning rate moderates study time allocation. This inverse "U" shape is also in line with Koriat's (2012) proposal of interactive influences between metamemory monitoring and control. Specifically, not only are JOLs informed by study time, but JOLs also inversely contribute to regulation of study time.

Besides the above factors, numerous studies have established that study time allocation is also affected by the perceived importance (or value) of study materials (e.g., Castel 2007; DeLozier and Dunlosky 2015), which further challenges self-regulated study time allocation as a valid measure of processing fluency. Take the font size effect on JOLs as an example. Mueller et al.'s (2014) Experiment 2 observed a null difference in study time between large and small words in a self-regulated study time allocation task, and their Experiments 3a and 3b observed that some participants thought large words were more important to remember than small ones. Hence, large words might have both invited longer study because they were believed to be more important, as well as shorter study because they were processed more fluently, thus leading to an overall null difference in study time between large and small words. In summary, self-regulated study time allocation could be driven by various factors in a goal-driven manner, casting doubt on its validity as a measure of processing fluency (for related discussion, see Su et al. 2018, p. 10).

Moreover, results from measures of study time allocation are sometimes inconsistent with those from other putative measures of processing fluency. For example, although Jia et al. (2015) found no difference in study time between high- and low-frequency words in their self-regulated study time allocation task, numerous other studies have established that people process high-frequency words much faster than low-frequency ones in many other tasks, such as lexical decision, item naming latency, CID, and so on (e.g., Balota and Chumbley 1984; Grainger 1990; Liu and Reichle 2017; Schilling et al. 1998). Mueller et al. (2014) observed no difference in self-allocated study times between large and small words, but as noted above Yang et al. (2018a) observed that participants identified large words much faster than small ones in their CID task. While Witherby and Tauber (2017a) obtained a null difference in study times between concrete and abstract words in their self-regulated study time allocation task, they observed that participants processed concrete words faster than abstract ones in their lexical decision and mental image formation tasks. Hence, the aforementioned findings, based on self-regulated study time allocation, cannot be taken as proving that processing fluency plays no role in the construction of JOLs.

## Study trials

Besides the experimental methods discussed above, some studies have employed a study trials task to measure processing fluency (e.g., Undorf and Erdfelder 2014; Witherby and Tauber 2017a). For example, in Witherby and Tauber's (2017a) Experiment 6, participants were instructed to study 30 (15 concrete and 15 abstract) words and were later asked to freely recall as many of the words as possible. Participants then restudied the words which they failed to recall and then retook a free recall test. This study-test cycle repeated until the participant successfully recalled all words. Witherby and Tauber hypothesized that, if concrete words were processed more fluently, participants would require fewer trials (cycles) to remember them than abstract words. Their results revealed a null difference in the mean number of trials required to remember concrete and abstract words, and they hence proposed that their results were consistent with the analytic processing model (p. 649). Again, this proposal is problematic because the study trials method is likely to be an insensitive measure of processing fluency.

A significant issue with taking study trials as a measure of processing fluency is that it does not yield convergent results with other measures. Whereas Witherby and Tauber (2017a, Exp. 6) found no difference in the number of study trials required to learn concrete and abstract words, as mentioned above they did find a medium-sized difference ($d = 0.32$) in lexical decision (Exp. 4) and a small but significant effect ($d = 0.22$) in another putative measure of processing fluency, latency to generate a mental image of the word (Exp. 7).

Across Witherby and Tauber's (2017a) Experiments 2–7, they consistently found that concrete words were more memorable than abstract ones. This raises a paradox: why was no difference detected in the number of study-test cycles required to successfully recall concrete and abstract words? Regardless of whether the number of study trials is a measure of processing fluency, it is striking that Witherby and Tauber found no evidence that abstract words are less memorable than concrete ones and raises the possibility that their measure of processing fluency was insensitive or even invalid. The study that first adopted the study trials measure as an index of fluency (Koriat 2008) found a strong (inverse) correlation across words between number of study-test cycles required and final recall.

Another shortcoming of the study trials task is that it principally measures how memorable the study materials are, rather than providing a measure of the processing fluency experienced in a typical JOL task in which no repeated retrieval practice is involved (Dunlosky and Tauber 2016). For instance, even though large words are percieved more fluently than small ones (Yang et al. 2018a), it is reasonable to expect a null difference in study trials between large and small words because font size has minimal influence on memory (Rhodes and Castel 2008). Likewise, although identical word pairs (e.g., *dog-dog*) are associated with greater processing fluency than related ones (e.g., *cat-dog*), the study trials task is expected to detect that learners require more trials to remember identical than related pairs because related pairs are more memorable than identical ones (Mueller et al. 2016).

For the sake of brevity, we do not discuss all the studies that have claimed to find a null contribution of processing fluency to JOL formation. The above discussion is, we hope, sufficient to illustrate the problems in the experimental methods used to capture the contribution of processing fluency.

## Experimental methods for measuring beliefs

Besides the aforementioned problems in measuring processing fluency, there are also problems in some of the experimental methods used to measure the role of beliefs in the construction of JOLs. This section discusses the potential shortcomings of five experimental methods for assessing the role of beliefs in the construction of JOLs, including belief questionnaires, pre-study JOLs, the classic and revised learner-observer tasks, and the belief-manipulation task.

### Belief questionnaires

Many previous studies have measured people's metamemory beliefs through belief questionnaires (e.g., Li et al. 2016; Mueller et al. 2014; Witherby and Tauber 2017a; Yang et al. 2017b). For example, in Witherby and Tauber's (2017a) Experiment 1, learners' beliefs about the effect of concreteness on memory were investigated. In the questionnaire, participants were informed that some students had studied 20 concrete and 20 abstract words, and then all participants were asked to estimate how many concrete and abstract words they thought the students would remember on a later test. Participants predicted that students would remember more concrete than abstract words. Even though findings derived from direct questions like this can be informative about the presence or absence of metamemory beliefs, these results do not tell us whether people actually apply such beliefs in constructing their JOLs. It is important to emphasize that many studies have found that people do not always apply their beliefs when forming JOLs online (e.g., Koriat et al. 2004; Kornell and Hausman 2017; Kornell et al. 2011; Tauber et al. 2019).[2] Methods such as mediation analysis (discussed in detail below) are needed to validate the linkage between beliefs and JOLs.

### Pre-study JOLs

In an attempt to supplement or improve on belief questionnaires, some studies have employed a pre-study JOL task (e.g., Mueller et al. 2016; Mueller et al. 2014; Witherby and Tauber 2017a). In this procedure, before viewing each item, participants are informed about its type

[2] Tauber et al. (2019) provided a demonstration of this. In their study, they explored whether metamemory beliefs about "memory declin[ing] with aging across adulthood" contribute to online JOL formation. Tauber et al. first conducted a survey to verify the existence of metamemory beliefs about age-related memory decline. Student participants were offered two options (Yes/No) to answer the question "*Do you think aging influences memory? That is, does people's ability to learn new information decline as they become 65 years or older?*" Most (81%) answered "Yes". Then, across 7 experiments involving a variety of experimental manipulations, Tauber et al. investigated whether college students applied such beliefs to construct online JOLs. In these experiments, participants made item-by-item JOLs to study words to predict whether a younger (18–21 years old) and/or older (65+ years old) adult would be likely to remember them on a future test. A meta-analysis, integrating results from all 7 experiments, showed an overall null difference between JOLs made for younger and older adults, indicating that people do not apply their beliefs about age-related memory decline when they form online JOLs.

(e.g., whether it is a concrete or abstract word, or in large or small font, etc.) and then are asked to estimate the likelihood that it will be remembered. The logic of this procedure is that, because participants make a pre-study JOL before viewing the actual item, these judgments can only be based on metamemory beliefs rather than on subjective experience (processing fluency). For example, Witherby and Tauber's (2017a) Experiments 2 and 3 employed the pre-study JOL procedure to test whether participants applied beliefs about concreteness to form their JOLs. Before studying each word, participants were explicitly told that "The word you are about to study is concrete [abstract]" and were required to make a pre-study JOL. Then they studied the word. The results showed that participants gave significantly higher pre-study JOLs to concrete than to abstract words. Based on this finding, Witherby and Tauber (2017a) proposed that "participants used beliefs about the concreteness effect on memory to inform their JOLs on an item-by-item basis" (p. 643).

Here we propose that pre-study JOLs cannot be regarded as providing direct evidence to support the claim that beliefs contribute to JOL formation in the normal condition wherein JOLs are made after studying each item. (Interested readers can read Price and Harrison (2017) for further comparison between standard and pre-study JOLs.) In order to strongly support this claim, the results should demonstrate that people's beliefs about the effect of a given factor on memory significantly mediate or moderate that factor's effect on JOLs. In other words, it is not sufficient that pre-study and standard JOLs vary in the same direction between (say) concrete and abstract words, it is also necessary that beliefs (i.e., pre-study JOLs) statistically mediate the effect of concreteness on standard JOLs. The necessity of mediation analysis to identify the source(s) of a factor's effect on JOLs is discussed in the *Statistical issues* section below.

### The classic learner-observer task

Some studies have employed a classic learner-observer task to assess the role of beliefs in the formation of JOLs (e.g., Matvey et al. 2001; Undorf et al. 2017). In those experiments, a study group performed a study task in which they viewed stimuli sequentially and made item-by-item JOLs. By contrast, an observation group viewed another participant's study trials and made item-by-item JOLs attempting to predict the likelihood that the other participant would remember each item later. For the observation group, all stimuli were presented in the same format (e.g., the same study duration or font size) as the other participant experienced, but the actual stimuli were replaced by meaningless letter strings (e.g., *abcde*; Yang et al. 2018a) or a black rectangle (e.g., Mueller et al. 2014). Therefore, in the observation group, participants did not view the true stimuli and could only form their JOLs based on their own beliefs. Again, such a design makes it impossible to ascertain whether the beliefs measured from the observation group actually contribute to JOLs measured from the study group.

It should be noted that the classic learner-observer task can be used to measure beliefs and JOLs when the study and observation tasks are successively performed by the same participants (e.g., Koriat and Ackerman 2010; Undorf and Erdfelder 2011). For instance, Undorf and Erdfelder (2011) first instructed participants to complete a study task in a self-paced procedure and make item-by-item JOLs. Next, the same participants performed an observation task, in which they viewed how much time another participant spent studying each item, and made item-by-item JOLs to predict the likelihood that the other participant would remember each item in a later test. Asking participants to successively perform the study and observation tasks allows beliefs and JOLs to be collected from the same participants. However, because participants observe another participant's study trials in the observation task, it is still difficult to conduct mediation analyses to quantify the contribution of beliefs to JOLs.

## The revised learner-observer task

To solve the above-discussed limitations, Yang et al. (2018a) recently revised the classic learner-observer task. In their Experiment 3, Yang et al. (2018a) explored whether beliefs contribute to the font size effect on JOLs. Participants first performed a learning task in which they identified words displayed in either large or small fonts, and made item-by-item JOLs. Next, they performed an observation task, in which they were told to view another participant's identification trials and made item-by-item JOLs to predict that participant's recall likelihood. In reality, all participants were shown their own identification trials, but all words were replaced by meaningless letter strings (i.e., *abcde*) presented in the same font size and for the same duration as the words in the learning task. Yang et al. recorded each item's font size, processing fluency (identification RT in the learning task), JOLs (in the learning task), and beliefs (JOLs in the observation task) from the same participants.

To test whether beliefs play a role in the font size effect on JOLs, Yang et al. (2018a) conducted a multilevel mediation analysis with font size as the independent variable, beliefs (JOLs in the observation task) as the mediator, and JOLs (in the learning task) as the dependent variable. The results showed that beliefs only mediated the font size effect on JOLs to a small and non-significant degree (for detailed results, see Yang et al. 2018a, p. 107). Importantly, this weak mediation does not categorically refute the hypothesis that beliefs may contribute to the font size effect on JOLs. It is, after all, a null result and hence, as will be discussed later, must be interpreted in terms of the experiment's power to detect a small but real effect. But the key point is that techniques for evaluating the mediating effect of beliefs on JOLs do exist.

Several studies have measured participants' beliefs using pre-learning questionnaires (that is, before initiating the main experiment, they administered questionnaires to assess participants' beliefs about a factor's effect on memory; e.g., Frank and Kuhlmann 2016; Hu et al. 2015). Such a design also allows beliefs and JOLs to be collected from the same participants and permits multilevel mediation analyses (using the 1–2–1 model; see Zhang et al. 2008, for details) to be conducted to quantify the contribution of beliefs. Given that questionnaires measure beliefs through only one or two questions, the measured "beliefs" may contain substantial measurement error. By comparison, the revised learner-observer task measures beliefs repeatedly across trials, which is likely to reduce measurement error and enhance measurement stability. In addition, pre-learning questionnaires can only measure *a priori* beliefs but not beliefs that gradually develop across the learning task (see Undorf and Erdfelder 2011, for an illustration of such a belief change). By contrast, the learner-observer task can readily detect newly developed beliefs because the observation task is administered following the completion of the study task.

## Belief-manipulation

Several recent studies investigated the role of beliefs by directly manipulating them (e.g., Blake and Castel 2018; Chen et al. 2019; Yang et al. 2017b). We term this the *belief-manipulation* method. In this procedure, before the main part of the experiment, participants are directly and explicitly given instructions designed to attenuate (e.g., "prior research proved no relationship between font size and memory") or even reverse (e.g., "prior research proved that small words were easier to remember than large ones") their a priori beliefs about a factor's effect on memory (e.g., "large words are easier to remember"). Through direct manipulation, the aim is to determine whether the updated beliefs (induced by instructions) can reduce, eliminate, or even reverse that factor's effect on JOLs. For instance, to investigate the role of beliefs in the font size effect on JOLs, Blake and Castel (2018, Experiment 1B) informed their participants that "Research has shown that, for college-age participants, words in smaller fonts are easier to recall than words in larger fonts", and then instructed them to study large and small words and make item-by-item JOLs. Blake and Castel observed that their belief manipulation reduced (but did not eliminate) the font size effect on JOLs.

The belief-manipulation method suffers from at least two major disadvantages. The first is the high risk of demand characteristics. Given that the belief-manipulation instructions directly and explicitly inform participants that one type of to-be-studied material is easier to remember than another, and those instructions also typically conflict with their own beliefs, participants may assume that the researchers want them to make memory predictions in alignment with the instructions. Accordingly they may simply provide JOLs to satisfy such an assumed "task requirement", regardless of whether those instructions truly alter their beliefs,[3] and no matter whether they really apply the updated beliefs to form their JOLs. The second disadvantage is that this method can only provide answers about whether explicitly manipulated beliefs influence JOL formation under these specific (artificial) conditions, but not whether natural beliefs actually contribute to JOL construction under "normal" conditions wherein they are not directly manipulated. That is to say, belief-manipulation is likely to be an invalid method to explore whether beliefs are responsible for a given JOL effect in the natural non-manipulated condition.

## Statistical issues

Thus far, we have summarized some problems besetting the use of common experimental methods in metacognition research. In this section, we focus on statistical issues. Specifically, we first list several examples to explain how low statistical power can lead to false negative results regarding the role of processing fluency in JOLs. Next, the double-standard analytic treatment of processing fluency and beliefs in some previous studies is discussed. Lastly, the benefits and pitfalls of three mediation analysis methods are explained.

### Low statistical power for mediation analysis

To establish a role of processing fluency in JOLs, research findings in metamemory studies must meet at least three requirements. First, a given factor must significantly affect JOLs. Secondly, that factor must significantly affect a measure of processing fluency. Thirdly, that factor's effect on processing fluency (e.g., the difference in processing fluency between different types of materials) must significantly mediate its effect on JOLs. As an illustration of the latter, Dunlosky and Mueller (2016) re-analyzed data from Magreehan et al.'s (2016) Experiment 4 to illustrate why mediation analysis is vital for identifying potential mechanisms underlying a factor's effect on metacognitive judgments.

[3] Several studies administered a post-task questionnaire at the very end of their experiments to check whether participants trusted the instructions or not (e.g., Blake and Castel 2018).

In this experiment, Magreehan et al. investigated the perceptual-degradation effect on JOLs. Magreehan et al. observed that participants spent less time encoding and made higher JOLs to bold word pairs presented in black against a white background (e.g., **CORN – PLANET**) than they did to italicized pairs in light grey (e.g., *HARNESS – SNAKE*). Dunlosky and Mueller (2016) noted that a seductive inference from this experiment was that perceptual degradation affects JOLs through processing fluency.[4] To illustrate why this tempting inference was problematic (and putting aside our earlier remarks about the study time allocation measure), Dunlosky and Mueller conducted a mediation analysis on Magreehan et al.'s data and found that study duration did not significantly mediate the perceptual-degradation effect. Hence, Dunlosky and Mueller proposed that their mediation analysis results "are inconsistent with the hypothesis that differential processing fluency is responsible for the impact of font on JOLs and hence suggest some other factor is responsible" (p. 126).[5] Based on this evidence, Dunlosky and Mueller (2016, p. 127) recommended that "after one establishes that a manipulation (e.g., kind of font) influences a person's judgments, learning, or reasoning, then further empirical work [such as mediation analysis] may be needed to reveal the source of the influence."

It is well-known that mediation analysis requires large sample sizes and many trials (Fritz and MacKinnon 2007), and underpowered studies can frequently lead to false negative findings (Type II error; see Vadillo et al. 2016, for a detailed discussion). Unfortunately, sample sizes and numbers of trials were relatively small in some previous studies, which might be insufficient to detect a mediating role of processing fluency.

Jia et al. (2015), for instance, only presented 10 high- and 10 low-frequency words to 30 participants to measure the role of processing fluency in the word frequency effect on JOLs. P. Li et al. (2016) employed 12 animate and 12 inanimate words and 28 participants to explore the role of processing fluency in the animacy effect on JOLs. The number of participants (67) and trials (32 in each of the fluent and disfluent font conditions) were somewhat greater in Magreehan et al.'s (2016) experiment described above, but the key result of Dunlosky and Mueller's (2016) mediation analysis was still a null result with uncertain precision. In their Experiment 7, Witherby and Tauber (2017a) presented 15 concrete and 15 abstract words to 40 participants to explore the role of processing fluency in the concreteness effect on JOLs. In this latter example, the indirect effect of concreteness on JOLs through processing fluency was 1.88, 95% confidence interval (CI) [−0.17, 5.69]. Thus, the mediation results are compatible with a true mediation effect in excess of 5.69 (the upper bound of the CI) on a 100-point scale, which by most standards is a large effect. Although the indirect effect approached significance (as revealed by the lower bound of the CI being close to 0) and the CI includes a large mediation effect,[6] Witherby and Tauber (2017a) proposed that "image latency did not mediate the relationship between concreteness and JOLs" (p. 648). Indeed, Witherby and Tauber acknowledged that their Experiment 7 was underpowered (p. 649).[7] Lack of power might be

[4] We note that neither Magreehan et al. nor Dunlosky and Mueller claimed that perceptual degradation affects JOLs through processing fluency.

[5] Although their mediation results were non-significant, Dunlosky and Mueller did not reject a potential role of fluency.

[6] Put differently, the study was underpowered to detect a mediation effect of 5 points on the 0–100 JOL scale, as both 0 and 5 fell inside the CI.

[7] Regarding their Experiment 7, Witherby and Tauber found by simulation that approximately 5000 participants are required to observe a significant mediation effect of fluency at 0.8 power. Based on this, they proposed that "image latency will not be a primary factor driving the concreteness effect on JOLs" (p. 649). It is important to note that all their mediation and simulation analyses were conducted using clustered data (i.e., mean JOLs and median RTs), which suffer from significant analytic pitfalls, as discussed in a later section.

a fundamental source of some of the null results regarding the mediating role of processing fluency found in the aforementioned studies.

### Double-standard analytic treatment of processing fluency and beliefs

Another fundamental statistical issue is that many previous studies have treated the roles of processing fluency and beliefs to different levels of analytical rigour, only subjecting processing fluency but not beliefs to mediation analysis (e.g., Jia et al. 2015; Li et al. 2016; Mueller and Dunlosky 2017; Mueller et al. 2013, 2014, 2016; Susser and Mulligan 2015; Witherby and Tauber 2017a).

As demonstrated by Dunlosky and Mueller (2016) in the analysis described above, mediation analysis is an effective tool for identifying the potential sources of a given factor's effect on metacognitive judgments. Unfortunately, the majority of previous studies did not conduct mediation analyses to validate the role of beliefs in the construction of JOLs (e.g., Jia et al. 2015; Mueller et al. 2013, 2014, 2016; Susser and Mulligan 2015; Witherby and Tauber 2017a). Furthermore, processing fluency and beliefs have frequently been treated to different standards of analytical rigour. As an illustration, consider the study by Mueller et al. (2016) which explored why, despite the fact that semantically related word pairs (e.g., *dog-cat*) are better recalled on average than identical word pairs (e.g., *dog-dog*) in a later test, people give higher JOLs to identical than to related pairs – the identity effect on JOLs.

In Muller et al.'s Experiment 1, they instructed participants to spend as much time as they wanted to study each word pair and make item-by-item JOLs. The results showed that participants spent less time studying identical than related pairs and gave higher JOLs to identical pairs. However, a partial correlation analysis (see below for a detailed discussion of this analytic method) showed that the extent to which processing fluency (indexed by study duration) mediated the identity effect on JOLs was not statistically significant. Mueller et al. (p. 787) proposed that their Experiment 1 "disconfirmed one version of the fluency hypothesis for the identity effect—namely, the fluency of processing (as measured by study time) did not statistically mediate the relationship between pair type (identical *vs.* related pairs) and JOLs". Then in their Experiment 3, Mueller et al. employed the pre-study JOL paradigm to explore the role of beliefs in the identity effect on JOLs. The results showed that participants gave higher pre-study JOLs to identical pairs than to related ones. Without conducting a mediation analysis, Mueller et al. (p. 790) concluded that "beliefs about the type of word pair contribute significantly to JOLs". Other studies have similarly treated processing fluency and beliefs to different standards of analytical evaluation (e.g., Undorf and Erdfelder 2014; Witherby and Tauber 2017a).

### Issues in mediation analysis methods

Finally, this section discusses some important methodological issues in mediation analysis. Because the majority of previous studies did not conduct mediation analyses to explore the roles of beliefs in JOL formation, this section focuses on their use in quantifying and drawing theoretical conclusions about the role of processing fluency. Below we first briefly describe shortcomings of partial correlation analysis, and then employ a case study (Witherby and Tauber 2017a) to compare two other mediation analysis methods: clustered vs. multilevel mediation.

The most widely-used mediation analysis method is partial correlation (e.g., Hertzog et al. 2003; Mueller et al. 2013, 2016; Susser et al. 2016; Susser and Mulligan 2015). Using this method, researchers first compute a zero-order correlation between item type (e.g., words in small vs. large fonts) and JOLs, obtaining a correlation value $r_{0i}$ for each participant ($i$ denotes each participant). They then calculate a first-order correlation ($r_{1i}$) between item type and JOLs with processing fluency (e.g., lexical decision RTs) controlled, and finally conduct a paired $t$-test between $r_{0i}$ and $r_{1i}$. If, across the $i$ (= 1…$N$) participants, $r_{0i}$ is significantly greater than $r_{1i}$ (i.e., controlling processing fluency weakens the correlation between item type and JOLs), processing fluency is inferred to contribute to JOLs. Unfortunately partial correlation is acknowledged as problematic because this method often increases Type I errors (false positives) in certain circumstances (for detailed discussion, see MacKinnon et al. 2002; Montoya and Hayes 2017; Murayama et al. 2014).

Considering the shortcomings of the partial correlation approach, some researchers have recently applied other mediation analysis methods, such as clustered (e.g., Dunlosky et al. 2014; Witherby and Tauber 2017a) and multilevel mediation (e.g., Undorf et al. 2017; Yang et al. 2018a, b). For example, Witherby and Tauber (2017a) adopted a clustered mediation analysis to explore the role of processing fluency in the concreteness effect on JOLs. Using a lexical decision task in their Experiment 4, Witherby and Tauber (2017a) found that participants responded faster to concrete than to abstract words and that they gave higher JOLs to concrete than to abstract words. To explore whether processing fluency underlies the concreteness effect on JOLs (i.e., whether processing fluency mediates the concreteness effect on JOLs), Witherby and Tauber (2017a) conducted a path-analytic mediation analysis using the *SPSS MEMORE* package (Montoya and Hayes 2017).

Specifically, they calculated a mean JOL for concrete and abstract words, and a median RT for concrete and abstract words for each participant. They then inserted these data into the *MEMORE* program to run a mediation analysis. The logic of Witherby and Tauber's analysis was that, if processing fluency contributes to the concreteness effect on JOLs, the difference in median RTs between concrete and abstract words should predict the difference in mean JOLs across participants. Their results showed that processing fluency failed to significantly mediate the concreteness effect on JOLs: although concreteness was correlated with both JOLs and lexical decision RTs, there was no reliable correlation between the differences in median RTs and the differences in mean JOLs, and hence no statistical mediation by processing fluency (RTs) of the concreteness-JOLs association.

As we can see, by using a path-analytic mediation analysis method and the MEMORE package, Witherby and Tauber (2017a) analyzed the mediation effect using the clustered mediation analysis method (i.e., their mediation analysis was based on the median RTs and mean JOLs for each participant). Here, we propose that it is problematic to test a mediation effect using participant clustering. Since it is of particular interest whether RTs measured by the lexical decision task mediate the concreteness effect on JOLs within each participant, it is inappropriate to draw inferences at the lower level (i.e., item level within each participant) based on data from a higher level (i.e., mean JOLs and median RTs at the participant level) (Snijders 2011). Numerous studies have shown that the relationship between two variables at a higher level can differ from the relationship between the same variables at a lower level (Piantadosi et al. 1988; Robinson 1950).[8] For example, in Witherby and Tauber's study,

[8] See Simpson's paradox and the "UC Berkeley gender bias" affair (available at https://en.wikipedia.org/wiki/Simpson%27s_paradox) for further illustrations of the pitfalls of clustered data analysis.

although participants who had shorter median RTs in the lexical decision task did not give higher mean JOLs (at the participant level), items with relatively short RTs might still receive relatively higher JOLs within each participant (at the item level). In other words, the RT-JOL relationship might not be detectable through participant-level analyses, despite its existence at the item level.

We reanalyzed the data from Witherby and Tauber's (2017a) Experiment 4 to illustrate why it is inappropriate to explore the association between RTs and JOLs using clustered data. Since the fluency effect on JOLs has been demonstrated in numerous studies, we accordingly expect an inverse relationship between RTs and JOLs in Witherby and Tauber's (2017a) Experiment 4. To test this hypothesis, we conducted a multilevel regression analysis, using the R *lme4* package (Bates et al. 2015), to measure the relationship between RTs and JOLs across items. Multilevel regression simultaneously takes account of variation at both item and participant levels, and allows us to directly examine the effect at the item level (Snijders 2011). After removing all cases of non-words and words which were erroneously judged as non-words, we regressed JOLs onto RTs in a multilevel linear regression model with a fixed effect for RTs and random slopes and intercepts across participants. The results show an inverse relationship between RTs and JOLs, $b = -6.11$, 95% CI [−9.30, −3.11], $p < .001$, indicating that every decrease of 1 s in RTs increased JOLs by 6.11 points on a 0–100 scale – the classic fluency effect on JOLs. Then we computed the relationship between RTs and JOLs at the participant level (i.e., using clustered data: mean JOLs and median RTs). For each participant, we calculated a mean JOL and a median RT for all correctly judged words. In contrast to the multilevel regression analysis above, this clustered regression analysis found no relationship between median RTs and mean JOLs, $b = 0.42$, $p = .98$. These results clearly reveal that it is inappropriate to assess the relationship between RTs and JOLs using mean JOLs and median RTs, because such clustering loses sight of the RT-JOL relationship at the item level.

Finally, we reanalyzed the data from Witherby and Tauber's (2017a) Experiment 4 to test whether lexical decision RT mediates the concreteness effect on JOLs using a multilevel mediation model via the R *bmlm* package (Vuorre 2017), which allows us to assess the mediation effect of RTs at the item level (Zhang et al. 2008). The *bmlm* package provides a Bayesian estimation of multilevel mediation models (Vuorre 2017). The analysis was conducted with word concreteness (represented as a dichotomous variable) serving as the independent variable, JOLs as the dependent variable, and RTs as the mediator. The mediation effect was estimated with 4 Markov Chain Monte Carlo (MCMC) chains and 10,000 iterations for each chain.

The mediation results are shown in Table 2. The total effect of concreteness on JOLs is 3.15, 95% CI [1.40, 4.91] and the direct effect of concreteness on JOLs is 2.50, 95% CI [0.77, 4.22]. This direct effect is significant and indicates that processing fluency (RTs) cannot fully explain the concreteness effect on JOLs. Crucially, the indirect effect of concreteness on JOLs through RTs is 0.65, 95% CI [0.23, 1.20], and the proportion of the total effect of concreteness on JOLs mediated by RTs is 22%, 95% CI [7%, 49%]. These results reveal that RTs significantly mediated the concreteness effect on JOLs. We also estimated the mediation effect for each participant (see Fig. 1). Thirty-nine out of 40 participants showed a simulated mediation parameter greater than 0, $\chi^2(1) = 36.1$, $p < .001$. Overall, these results clearly support the claim that processing fluency contributes to the concreteness effect on JOLs. However, appropriate analytic methods are required to reach this conclusion. To our knowledge, this analysis is the first to provide evidence to support the mediating role of processing fluency in the concreteness effect on JOLs.

Overall, the above example clearly shows that multilevel mediation, compared with clustered mediation, is more appropriate for capturing a given factor's mediating role at the trial level. Accordingly, we advise that future JOL research should consider employing multilevel (rather than clustered) mediation to assess the role of processing fluency in the construction of JOLs. Reassuringly, more and more researchers are acknowledging the merits of multilevel mediation, and this statistical approach is gradually gaining popularity in the JOL field (Frank and Kuhlmann 2016; Hu et al. 2020; Yang et al. 2018a).

## Summary of problems and suggested remedies

This section briefly summarizes potential shortcomings of the research methods discussed above, and then provides some suggestions about how to solve or at least mitigate those problems in future research.

### Employing sensitive and valid tasks to measure processing fluency

Different tasks (e.g., CID vs. lexical decision) are sensitive to different types of processing fluency (e.g., perceptual fluency), and employing an appropriate experimental task to measure a given type of processing fluency is a prerequisite for exploring its role in the formation of JOLs. Future research should develop and employ more appropriate experimental methods to measure processing fluency. In addition, the sensitivity and validity of a given task to measure a given type of processing fluency should be further examined, especially the self-regulated study time allocation and study trials tasks (for illustrations regarding how to verify task sensitivity, see Grainger and Segui 1990; Yang et al. 2018a).

Although the above discussion demonstrates significant problems in several experimental methods, we strongly emphasize that we are not recommending that future research abandons these methods, nor are we concluding that processing fluency cannot be measured. Instead, we propose that some measures may lack sensitivity (at least to some types of processing fluency) or even be invalid, and encourage researchers to take these problems into account and, more importantly, to explore possible approaches to solve (or at least minimize) these problems.

Future research should also consider employing sets of, rather than only one or two, tasks to measure the contribution of processing fluency. In a rare instance, Witherby and Tauber (2017a) used four tasks to evaluate whether processing fluency is a contributor to the concreteness effect on JOLs. A further suggestion is that even when the results from all tasks are convergent on disproving a contribution of processing fluency, the limitations of the tasks should not be neglected.

**Table 2** Multilevel mediation analysis of Witherby and Tauber's (2017a) Experiment 4

| Effects | *b* | *SE* | 95% CI |
|---|---|---|---|
| Total effect of concreteness on JOLs | 3.15 | 0.90 | [1.40, 4.91] |
| Direct effect of concreteness on JOLs | 2.50 | 0.87 | [0.77, 4.22] |
| Indirect effect of concreteness on JOLs through RTs | 0.65 | 0.25 | [0.23, 1.20] |
| Proportion of the concreteness effect on JOLs mediated by RTs | 22% | 13% | [7%, 49%] |

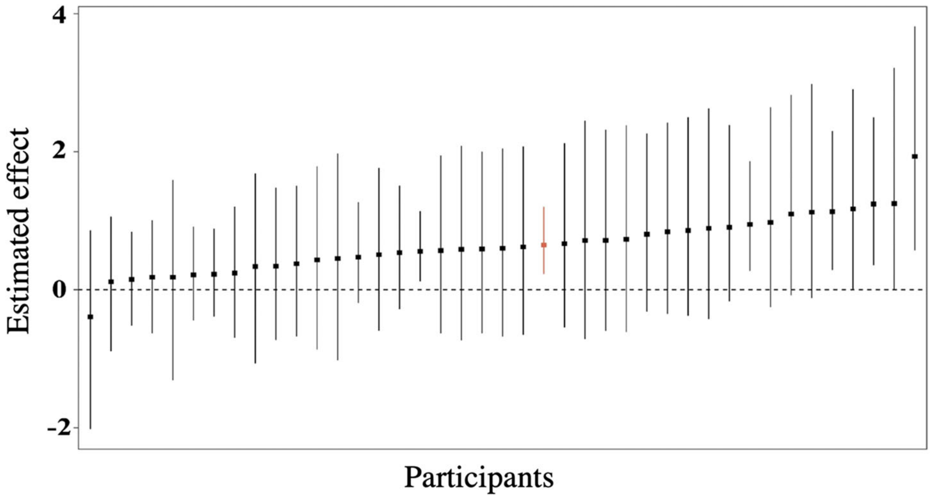


**Fig. 1** Bayesian estimation of the mediation effect of RTs in the concreteness effect on JOLs for each participant in a reanalysis of the data from Witherby and Tauber's (2017a) Experiment 4. Each black point represents a participant and the red point represents the fixed mediation effect across all participants. Error bars represent 95% CI

## Collecting beliefs and JOLs from the same participants

Results derived from belief questionnaires and pre-study JOLs can yield implications about the existence or absence of metamemory beliefs, but cannot be utilized to quantify their mediational or moderating roles, because beliefs and JOLs are measured from different groups of participants. Even though the classic learner-observer paradigm allows beliefs and JOLs to be collected from the same individuals when they are instructed to successively perform the learner and observer tasks, participants actually view another participant's learning trials in the observer task, thus making it impossible to run mediation analysis to quantify the contribution of beliefs to that individual's JOLs. Directly manipulating beliefs through instructions may fail to reveal whether participants apply their updated beliefs to form JOLs because the task requirement might be too overt. In addition, this task cannot reveal whether natural/non-manipulated beliefs are responsible for a given factor's effect on JOLs.

Going beyound the aforementioned methods, the revised learner-observer task allows researchers to measure beliefs and JOLs from the same individuals and trials, and also permits the statistical quantification of the contribution of beliefs. In addition, using the revised learner-observer task, future research can measure processing fluency, beliefs, and JOLs from the same participants, and then quantitatively compare the contribution of processing fluency and beliefs, which can be potentially used to assess the dual-basis and analytic processing models. However, as acknowledged by Yang et al. (2018a, p. 108), the revised learner-observer paradigm measures processing fluency and JOLs concurrently in the study task but beliefs are measured in the observation task, which may reduce the procedure's power to detect a role of beliefs. Future research should aim to develop more elegant procedures to simultaneously measure processing fluency, beliefs, and JOLs.

### Increasing statistical power, avoiding double-standard analytic treatment, and employing an appropriate mediation model

Many previous studies have subjected processing fluency to mediation analysis to evaluate its role in JOL formation, but unfortunately the sample sizes used in some studies were insufficient to detect a significant mediating effect (e.g., Jia et al. 2015). Future research should be wary of underpowered samples, and pre-planned sample sizes should be estimated (for how to estimate the required sample sizes for mediation analysis, see Faul et al. 2007; Fritz and MacKinnon 2007; Kadam and Bhalerao 2010).

Many previous studies have also subjected processing fluency but not metamemory beliefs to mediation analysis (e.g., Mueller et al. 2013, 2014, 2016; Witherby and Tauber 2017a). This double-standard analytic treatment is problematic since the higher standard set for establishing an influence of processing fluency means that the resulting findings can fallaciously support the analytic processing model (because beliefs are more likely to survive a weak test than processing fluency is to survive a stringent one). Future research should measure beliefs and JOLs from the same participants, and directly test, via mediation or moderation analyses (for differences between these two statistical methods, see Hu et al. 2020), whether beliefs contribute to JOL formation. The revised learner-observer paradigm is available as a method to achieve this aim.

Future research should also be cautious about mediation analysis methods, as inappropriate methods can lead to incorrect conclusions. For instance, partial correlation may increase Type I errors, and clustered mediation cannot appropriately be used to explore the role of processing fluency in the formation of JOLs because the association between JOLs and processing fluency mainly exists at the item level and clustered data eliminate variance associated with the item-level relationship. Going beyond these two methods, we suggest that multilevel mediation is more appropriate. Multilevel mediation is becoming increasingly popular in cognitive psychology, and more and more software packages for conducting such analyses are available to researchers (e.g., Muthén and Muthén 1998-2010; Preacher et al. 2010; Vuorre 2017; Zhang et al. 2008). Future research should consider using multilevel instead of clustered mediation, following several precedents (e.g., Frank and Kuhlmann 2016; Hu et al. 2020; Yang et al. 2018a).

## Concluding remarks

Bearing both theoretical and practical importance, the underlying mechanisms whereby metacognitive judgments of learning are constructed have received considerable attention in recent years. An emerging body of studies has focused on the roles of processing fluency and metamemory beliefs about how memory operates, but the findings are inconsistent. The measurement tools and analytic models suffer from a variety of limitations. Although many of these tools and models are imperfect, the current review does not intend to suggest that they should be abandoned. Instead, it aims to encourage researchers to explore potential approaches to solve (or at least minimize) those pitfalls and develop more elegant techniques. These important issues also motivate a call for re-evaluation of previous findings and the production of new and more robust data.

**Acknowledgments** We thank Amber E. Witherby and Sarah K. Tauber for sharing and allowing us to reanalyze their data (Witherby and Tauber 2017a). This study was supported by the Natural Science Foundation of China (32000742; 31671130), the Fundamental Research Funds for the Central Universities (2019NTSS28), and the United Kingdom Economic and Social Research Council (ES/P009522/1).

## Compliance with ethical standards

**Conflict of interest** The authors declare no competing interests. The current review does not involve new data collection, and hence ethical approval and formal informed consent are not required.

## Appendix: How does fluency affect JOLs?

Even though it is clear that beliefs contribute to the formation of JOLs in an analytical way, how processing fluency might affect JOLs has not been clearly elucidated. There are several possibilities (Dunlosky et al. 2014; Matvey et al. 2001; Mueller and Dunlosky 2017; Yang et al. 2018a). The first is that it affects JOLs in a direct way, in which processing fluency may produce a subjective feeling-of-knowing and this subjective feeling directly acts as a basis for higher JOL ratings (Koriat 1997; Matvey et al. 2001; Schwarz and Reber 1999; Yang et al. 2018a). The second possibility is that fluency affects JOLs indirectly through beliefs about fluency, that is, believing that more fluently processed items are more memorable (Mueller and Dunlosky 2017).

A third possibility is that whether and how experienced fluency affects JOLs is dependent on an unconscious interpretive process that attributes processing fluency to a possible source (Jacoby et al. 1989; Kelley and Jacoby 1990). Based on Jacoby's source attribution approach (also see Whittlesea 1993), when processing fluency is attributed to memory, it drives learners to make higher JOLs. By contrast, when processing fluency is attributed to other situational or external factors (such as font size) that are not directly related to memory, it exerts little influence on JOLs. For instance, the reason why perceptual fluency affects JOLs may be that learners unconsciously and mistakenly attribute the fluency feelings induced by perceptual features to ones predictive of memorability.

Overall, processing fluency may affect JOLs in a variety of ways. Only a few studies have explored the possible underlying mechanisms and the research findings are inconsistent to date (e.g., Matvey et al. 2001; Su et al. 2018; Susser et al. 2017; Undorf and Erdfelder 2011; Yang et al. 2018a). Further research is needed to shed additional light on this important issue.

## References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*, 219–235. https://doi.org/10.1177/1088868309341564.

Ariel, R., Hines, J. C., & Hertzog, C. (2014). Test framing generates a stability bias for predictions of learning by causing people to discount their learning beliefs. *Journal of Memory and Language, 75*, 181–198. https://doi.org/10.1016/j.jml.2014.06.003.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 340–357. https://doi.org/10.1037/0096-1523.10.3.340.

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

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*, 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*, 897–903. https://doi.org/10.3758/s13421-013-0307-8.

Blake, A. B., & Castel, A. D. (2018). On belief and fluency in the construction of judgments of learning: assessing and altering the direct effects of belief. *Acta Psychologica, 186*, 27–38. https://doi.org/10.1016/j.actpsy.2018.04.004.

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review, 20*, 1350–1356. https://doi.org/10.3758/s13423-013-0442-z.

Castel, A. D. (2007). The adaptive and strategic use of memory by older adults: evaluative processing and value-directed remembering. *Psychology of Learning and Motivation, 48*, 225–270. https://doi.org/10.1016/s0079-7421(07)48006-9.

Chen, Y., Li, F., & Li, W. (2019). The influence of learner's beliefs about processing fluency on font-size effect. *Acta Psychologica Sinica, 51*, 154–162. https://doi.org/10.3724/sp.J.1041.2019.00154.

Chumbley, J. I., & Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Memory & Cognition, 12*, 590–606. https://doi.org/10.3758/BF03213348.

DeLozier, S., & Dunlosky, J. (2015). How do students improve their value-based learning with task experience? *Memory, 23*, 928–942. https://doi.org/10.1080/09658211.2014.938083.

Dunlosky, J., & Hertzog, C. (2000). Updating knowledge about encoding strategies: a componential analysis of learning about strategy effectiveness from task experience. *Psychology & Aging, 15*, 462–474. https://doi.org/10.1037//0882-7974.15.3.462.

Dunlosky, J., & Mueller, M. (2016). Recommendations for exploring the disfluency hypothesis for establishing whether perceptually degrading materials impacts performance. *Metacognition and Learning, 11*, 123–131. https://doi.org/10.1007/s11409-016-9155-9.

Dunlosky, J., & Tauber, S. K. (2016). *The Oxford handbook of metamemory*. Oxford University Press.

Dunlosky, J., Mueller, M., & Tauber, S. K. (2014). The contribution of processing fluency (and beliefs) to people's judgments of learning. In D. S. Lindsay, C. M. Kelley, A. P. Yonelinas, & H. L. Roediger, III (Eds.), *Remembering: attributions, processes, and control in human memory. Papers in honour of Larry L. Jacoby* (pp. 46–64). New York: Psychology Press.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. https://doi.org/10.3758/BF03193146.

Ferrand, L., Brysbaert, M., Keuleers, E., New, B., Bonin, P., Meot, A., et al. (2011). Comparing word processing times in naming, lexical decision, and progressive demasking: evidence from chronolex. *Frontiers in Psychology, 2*, 306. https://doi.org/10.3389/fpsyg.2011.00306.

Frank, D. J., & Kuhlmann, B. G. (2016). More than just beliefs: experience and beliefs jointly contribute to volume effects on metacognitive judgments. *Journal of Experimental Psychology: Learning, Memory & Cognition, 43*, 680–693. https://doi.org/10.1037/xlm0000332.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18*, 233–239. https://doi.org/10.1111/j.1467-9280.2007.01882.x.

Froger, C., Sacher, M., Gaudouen, M. S., Isingrini, M., & Taconnat, L. (2011). Metamemory judgments and study time allocation in young and older adults: dissociative effects of a generation task. *Canadian Journal of Experimental Psychology, 65*, 269–276. https://doi.org/10.1037/a0022429.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language, 29*, 228–244. https://doi.org/10.1016/0749-596X(90)90074-A.

Grainger, J., & Segui, J. (1990). Neighborhood frequency effects in visual word recognition: a comparison of lexical decision and masked identification latencies. *Attention, Perception, & Psychophysics, 47*, 191–198. https://doi.org/10.3758/BF03205983.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 22–34. https://doi.org/10.1037/0278-7393.29.1.22.

Hu, X., Li, T., Zheng, J., Su, N., Liu, Z., & Luo, L. (2015). How much do metamemory beliefs contribute to the font-size effect in judgments of learning? *PLoS One, 10*, e0142351. https://doi.org/10.1371/journal.pone.0142351.

Hu, X., Zheng, J., Fan, T., Su, N., Yang, C., & Luo, L. (2020). Using multilevel mediation model to measure the contribution of beliefs to judgments of learning. *Frontiers in Psychology, 11*, 637. https://doi.org/10.3389/fpsyg.2020.00637.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: essays in honour of Endel Tulving* (pp. 391–422). East Sussex: Psychology Press.

Jia, X., Li, P., Li, X., Zhang, Y., Cao, W., Cao, L., & Li, W. (2015). The effect of word frequency on judgments of learning: contributions of beliefs and processing fluency. *Frontiers in Psychology, 6*, 1995. https://doi.org/10.3389/fpsyg.2015.01995.

Kadam, P., & Bhalerao, S. (2010). Sample size calculation. *International Journal of Ayurveda Research, 1*, 55–57. https://doi.org/10.4103/0974-7788.59946.

Kelley, C. M., & Jacoby, L. L. (1990). The construction of subjective experience: memory attributions. *Mind & Language, 5*, 49–68. https://doi.org/10.1111/j.1468-0017.1990.tb00152.x.

Kinder, A., Shanks, D. R., Cock, J., & Tunney, R. J. (2003). Recollection, fluency, and the explicit/implicit distinction in artificial grammar learning. *Journal of Experimental Psychology: General, 132*, 551–565. https://doi.org/10.1037/0096-3445.132.4.551.

Koriat, A. (1997). Monitoring one's own knowledge during study: a cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349–370. https://doi.org/10.1037/0096-3445.126.4.349.

Koriat, A. (2008). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition, 36*, 416–428. https://doi.org/10.3758/MC.36.2.416.

Koriat, A. (2012). The relationships between monitoring, regulation and performance. Learning and Instruction, 22, 296–298. https://doi.org/10.1016/j.learninstruc.2012.01.002

Koriat, A., & Ackerman, R. (2010). Metacognition and mindreading: judgments of learning for self and other during self-paced study. *Consciousness and Cognition, 19*, 251–264. https://doi.org/10.1016/j.concog.2009.12.010.

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 Psycholology: General, 133*, 643–656. https://doi.org/10.1037/0096-3445.133.4.643.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36–69. https://doi.org/10.1037/0096-3445.135.1.36.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: the benefits—and costs—of dropping flashcards. *Memory, 16*, 125–136. https://doi.org/10.1080/09658210701763899.

Kornell, N., & Hausman, H. (2017). Performance bias: why judgments of learning are not affected by learning. *Memory & Cognition, 45*, 1270–1280. https://doi.org/10.3758/s13421-017-0740-1.

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*, 787–794. https://doi.org/10.1177/0956797611407929.

Li, F., Xie, R., Li, X., & Li, W. (2015). The influence of perceptual information on control processes involved in self-regulated learning: evidence from item selection. *Psychonomic Bulletin & Review, 22*, 1007–1013. https://doi.org/10.3758/s13423-014-0762-7.

Li, P., Jia, X., Li, X., & Li, W. (2016). The effect of animacy on metamemory. *Memory & Cognition, 44*, 696–705. https://doi.org/10.3758/s13421-016-0598-7.

Liu, Y., & Reichle, E. D. (2017). Eye-movement evidence for object-based attention in Chinese reading. *Psychological Science, 29*, 278–287. https://doi.org/10.1177/0956797617734827.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83–104. https://doi.org/10.1037/1082-989X.7.1.83.

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*, 35–56. https://doi.org/10.1007/s11409-015-9147-1.

Matvey, G., Dunlosky, J., & Guttentag, R. (2001). Fluency of retrieval at study affects judgments of learning (JOLs): an analytic or nonanalytic basis for JOLs? *Memory & Cognition, 29*, 222–233. https://doi.org/10.3758/BF03194916.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*, 463–477. https://doi.org/10.1016/j.jml.2004.12.001.

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods, 22*, 6–27. https://doi.org/10.1037/met0000086.

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., 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*, 378–384. https://doi.org/10.3758/s13423-012-0343-6.

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*, 1–12. https://doi.org/10.1016/j.jml.2013.09.007.

Mueller, M. L., Dunlosky, J., & Tauber, S. K. (2016). The effect of identical word pairs on people's metamemory judgments: what are the contributions of processing fluency and beliefs about memory? *The Quarterly Journal of Experimental Psychology, 69*, 781–799. https://doi.org/10.1080/17470218.2015.1058404.

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*, 1287–1306. https://doi.org/10.1037/a0036914.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus user's guide (Version 6.)*. Los Angeles: Muthén & Muthén.

Piantadosi, S., Byar, D. P., & Green, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology, 127*, 893–904.

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General, 143*, 644–667. https://doi.org/10.1037/a0033194.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*, 209–213. https://doi.org/10.1037/a0020141.supp.

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.

Reber, R., & Zupanek, N. (2002). Effects of processing fluency on estimates of probability and frequency. In P. Sedlmeier & T. Betsch (Eds.), *Etc. frequency processing and cognition*: Oxford Scholarship Online.

Reber, R., Schwartz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8*, 364–382. https://doi.org/10.1207/s15327957pspr0804_3.

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.

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: effects on monitoring and control. *Psychonomic Bulletin & Review, 16*, 550–554. https://doi.org/10.3758/PBR.16.3.550.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351–357. https://doi.org/10.2307/2087176.

Schilling, H. E., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: word frequency effects and individual differences. *Memory & Cognition, 26*, 1270–1281. https://doi.org/10.3758/BF03201199.

Schwarz, N., & Reber, R. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8*, 338–342. https://doi.org/10.1006/ccog.1999.0386.

Snijders, T. A. B. (2011). Multilevel analysis. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 879–882). Berlin, Heidelberg: Springer Berlin Heidelberg.

Su, N., Li, T., Zheng, J., Hu, X., Fan, T., & Luo, L. (2018). How font size affects judgments of learning: simultaneous mediating effect of item-specific beliefs about fluency and moderating effect of beliefs about font size and memory. *PLoS One, 13*, e0200888. https://doi.org/10.1371/journal.pone.0200888.

Susser, J. A., & Mulligan, N. W. (2015). The effect of motoric fluency on metamemory. *Psychonomic Bulletin & Review, 22*, 1014–1019. https://doi.org/10.3758/s13423-014-0768-1.

Susser, J. A., Jin, A., & Mulligan, N. W. (2016). Identity priming consistently affects perceptual fluency but only affects metamemory when primes are obvious. *Journal of Experimental Psychology: Learning, Memory and Cognition, 42*, 657–662. https://doi.org/10.1037/xlm0000189.

Susser, J. A., Panitz, J., Buchin, Z., & Mulligan, N. W. (2017). The motoric fluency effect on metamemory. *Journal of Memory and Language, 95*, 116–123. https://doi.org/10.1016/j.jml.2017.03.002.

Tauber, S. K., Witherby, A. E., & Dunlosky, J. (2019). Beliefs about memory decline in aging do not impact judgments of learning (JOLs): a challenge for belief-based explanations of JOLs. *Memory & Cognition, 47*, 1102–1119. https://doi.org/10.3758/s13421-019-00919-3.

Undorf, M., & Ackerman, R. (2017). The puzzle of study time allocation for the most challenging items. *Psychonomic Bulletin & Review, 24*, 2003–2011. https://doi.org/10.3758/s13423-017-1261-4.

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1264–1269. https://doi.org/10.1037/a0023719.

Undorf, M., & Erdfelder, E. (2014). The relatedness effect on judgments of learning: a closer look at the contribution of processing fluency. *Memory & Cognition, 43*, 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*, 1217–1224. https://doi.org/10.3758/s13423-016-1189-0.

Undorf, M., & Zimdahl, M. F. (2019). Metamemory and memory for a wide range of font sizes: what is the contribution of perceptual fluency? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*, 97–109. https://doi.org/10.1037/xlm0000571.

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.

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review, 23*, 87–102. https://doi.org/10.3758/s13423-015-0892-6.

Vuorre, M. (2017). Bmlm: Bayesian multilevel mediation. R package version 1.3.0. Retrieved from https://cran.r-project.org/package=bmlm

Whittlesea, B. W. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1235–1253.

Witherby, A. E., & Tauber, S. K. (2017a). The concreteness effect on judgments of learning: evaluating the contributions of fluency and beliefs. *Memory & Cognition, 45*, 639–650. https://doi.org/10.3758/s13421-016-0681-0.

Witherby, A. E., & Tauber, S. K. (2017b). Monitoring of learning for emotional faces: how do fine-grained categories of emotion influence participants' judgments of learning and beliefs about memory? *Cognition & Emotion*, 1–7. https://doi.org/10.1080/02699931.2017.1360252.

Yang, C., Potts, R., & Shanks, D. R. (2017a). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied, 23*, 263–277. https://doi.org/10.1037/xap0000122.

Yang, C., Potts, R., & Shanks, D. R. (2017b). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 1073–1092. https://doi.org/10.1037/xlm0000363.

Yang, C., Huang, T. S. T., & Shanks, D. R. (2018a). Perceptual fluency affects judgments of learning: the font size effect. *Journal of Memory and Language, 99*, 99–110. https://doi.org/10.1016/j.jml.2017.11.005.

Yang, C., Sun, B., & Shanks, D. R. (2018b). The anchoring effect in metamemory monitoring. *Memory & Cognition, 46*, 384–397. https://doi.org/10.3758/s13421-017-0772-6.

Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2008). Testing multilevel mediation using hierarchical linear models. *Organizational Research Methods, 12*, 695–719. https://doi.org/10.1177/1094428108327450.