Metacognition and Learning (2022) 17:443–476
https://doi.org/10.1007/s11409-021-09287-3

Association and dissociation between judgments of learning and memory: A Meta-analysis of the font size effect

Minyu Chang[1] · **Charles J. Brainerd**[1]

Received: 19 August 2021 / Accepted: 2 December 2021 / Published online: 28 February 2022

Abstract

The font size effect is a metamemory illusion in which larger-font items produce higher judgments of learning (JOLs) but not better memory, relative to smaller-font items. We conducted meta-analyses to determine what is currently known about how font size affects JOLs and memory accuracy. In addition, we implemented both univariate and multivariate meta-regressions to isolate the moderators of JOL effects and memory effects. The results revealed a small-to-moderate effect of font size on JOLs. There was also a small but significant effect of font size on memory. This suggests that JOLs and memory accuracy both increase with font size, rather than being completely dissociated. Moreover, JOL-memory dissociation only occurred when font size ranged between very small and intermediate. Our working explanation is that the memory effects of font size are tied to (dis)fluency, but its JOL effects are not. Some boundary conditions were identified for font size effects on both JOLs and memory. Specifically, larger font sizes only reliably increased both JOLs and memory accuracy (a) when font sizes ranged from intermediate to very large, (b) when study materials were unrelated word lists, (c) when JOLs were solicited immediately after encoding, and (d) when study time was relatively brief.

Keywords Font size · Judgments of learning · Metamemory · Memory · Meta-analysis

Judgments of learning (JOLs) are commonly used metamemory measures that ask people to predict their future memory performance for currently encoded items. Such judgments can provide important guidance for study time allocation and for subjective regulation of learning strategies (Dunlosky & Ariel, 2011; Kornell & Bjork, 2008). However, JOLs and actual memory performance have been found to react differently or even in opposite ways to certain manipulations, suggesting that JOLs are prone to illusions and biases (e.g., Besken & Mulligan, 2013; Rhodes & Castel, 2008). The *font size effect* is one of the most studied phenomena of this sort. This effect was first studied by Rhodes and Castel (2008), who collected JOLs for words presented in either 18 pt. or 48 pt., followed by free recall tests. They reported an effect in which average JOLs were higher for words presented in

✉ Minyu Chang
mc2674@cornell.edu

[1] Department of Psychology and Human Neuroscience Institute, Cornell University, G331 MVR Hall, Ithaca, NY 14853, USA

48 pt. than in 18 pt., but recall did not differ between the two sizes. Moreover, the font size effect proved robust even when participants were given warnings about the nature of the illusion or with the availability of more effective memory cues such as semantic relations between words. Subsequently, the font size effect has been replicated in numerous experiments (e.g., Blake & Castel, 2018; Bröder & Undorf, 2019; Hu et al., 2015; Kornell et al., 2011; Luna, Nogueira, & Albuquerque, 2019b; Mcdonough & Gallo, 2012; Mueller et al., 2014; Price & Harrison, 2017; Su et al., 2018; Susser et al., 2013; Tatz & Peynircioğlu, 2020; Tatz et al., 2020; Undorf et al., 2018).

Research on the font size effect has been motivated by some enduring theoretical and empirical questions. Theoretically, this effect offers a valuable opportunity to shed light on the underlying mechanisms of JOLs. According to Koriat's (1997) cue-utilization framework, JOLs are controlled by cues that are likely to be informative about future memory performance, rather than by a direct evaluation of the strength of memory traces. Thus, manipulations that create JOL-memory dissociation reveal important information about people's incorrect metacognitive beliefs and biases in the uses of cues. On the empirical side, the finding that font size affects metacognitive judgments but not memory accuracy is pertinent to one of the most wide-spread typographical practices in education and advertising—namely, highlighting particularly important information by selectively enlarging its visual presentation.

In this article, we will present meta-analyses for the accumulated data on the font size effect. For theoretical background, we first discuss the ongoing uncertainty about the contributions of fluency and of belief to the font size effect on JOLs. Then, we address an emerging controversy about whether font size solely affects JOLs, or whether it also influences memory accuracy, as some recent evidence suggests. Next, we outline the potential moderators that figure in our moderator analyses and explain how our analyses elaborate on prior articles in this line of research. Finally, we present our meta-analyses for the font size effect on JOLs and on memory and discuss the implications of the results.

Are the JOL effects of font size driven by fluency or beliefs?

Although the effects of font size on JOLs were well replicated, the underlying mechanism for such effects is still unclear. A number of experimental results support that processing fluency plays a causal role in the font size effect on JOLs, that is, larger-font items are rated with higher JOLs because they are processed more fluently (Rhodes & Castel, 2008; Susser et al., 2013; Undorf et al., 2017; Wang, Qu, & Zhang, 2020a; Wang, Yang, et al., 2020b; Yang et al., 2018). Processing fluency refers to the level of subjective ease that people feel as they encode an item, which is traditionally regarded as a perceptual phenomenon that accrues from prior integration of items' surface features (Jacoby, 1991; Mandler, 1980). In that connection, Yang et al. (2018; Experiment 1) and Wang et al. (2020b; Experiment 4) used reaction time in continuous identification as a direct measure of processing fluency. Consistent with the notion of size-driven increases in fluency, mean reaction times in a continuous identification task were faster for larger than for smaller words. This reaction time difference then proved to mediate the font size effect on JOLs, thereby supplying rather strong support for the fluency hypothesis.

However, Mueller et al. (2014) reported that there was no difference in fluency between larger and smaller words when fluency was measured via either reaction times in lexical decision tasks or self-paced study time, which argues against the fluency hypothesis. Consistent with Mueller et al.'s conclusion, Su et al. (2018; Experiment 2a) measured

processing fluency with self-paced study time and found that it did not moderate the effect of font size on JOLs. Nevertheless, Yang et al. (2018, 2021) argued that lexical decision tasks and self-paced study time may be neither sensitive nor valid measures of processing fluency. For example, evidence has suggested that lexical decision reaction time is not a clean measure of perceptual processes because it also depends heavily on conceptual processing (Chumbley & Balota, 1984). Moreover, treating lexical decision time and self-paced study time as measures of perceptual fluency overlooks the fact that these two measures have been found to be dissociated from each other (Witherby & Tauber, 2017).

Still, other findings are congruent with an alternative hypothesis that metacognitive beliefs are responsible for JOLs being higher for larger words than for smaller ones (Blake & Castel, 2018; Hu et al., 2015; Luna, Nogueira, & Albuquerque, 2019b; Mueller et al., 2014; Su et al., 2018). Hu et al. (2015; Experiment 2) collected data on participants' metacognitive beliefs about font size by asking them to predict the memorability of larger and of smaller words in advance of a metamemory experiment. Next, JOL and memory data were collected in a typical JOL-memory experiment with a font size manipulation. They reported that the larger-smaller JOL difference was predicted by the larger-smaller difference in pre-experimental predictions, pointing to preexisting beliefs as the basis for the effects of font size on JOLs. Similarly, Su et al. (2018; Experiment 2a) measured beliefs about font size prior to metamemory experiments, and they reported that pre-existing beliefs moderated the effects of font size on JOLs.

To sum up, although some studies suggest that the font size effect on JOLs is primarily driven by fluency differences between larger and smaller fonts, these results are still clouded by the lack of standardized fluency measures in the literature. Yet, other studies suggest the font size effect is dominated by metacognitive beliefs about the benefits of larger fonts.

Does font size affect memory?

Although the modal finding in the font size research has been that larger words produce higher JOLs but not better memory, Luna et al.'s (2018) meta-analysis of 28 font size experiments showed that memory was slightly better overall for larger words than for smaller words ($g = .08$), even though such a memory benefit did not reach statistical significance in any individual study. Meanwhile, Halamish (2018) presented words in three font sizes (5 pt., 18 pt., 48 pt) in a set of 12 experiments and conducted a mini meta-analysis, which provided evidence that memory was more accurate for 48-pt words and 5-pt words than for 18-pt words.

Undorf and Zimdahl (2019) expanded Halamish's (2018) design and administered a wide range of font sizes in their experiments. In their Experiment 1, they used 48 font sizes, ranging from 6 pt. to 500 pt., which were classified into three categories: very small, intermediate, and very large. This classification was based on the *fluent range of print size* (Legge & Bigelow, 2011). According to Legge and Bigelow's review, the fluent range is defined as the print size range with maximum reading fluency, which extends from an angular size of 0.2° to 2.0°. Accordingly, the fluent range for words viewed at a distance of 85 cm is approximately 17–161 pt. (Undorf & Zimdahl, 2019). Between 17 and 161 pt., fluency should be the highest and remain relatively stable, whereas below 17 pt., fluency should increase as font size increases, and above 161 pt. it should decrease as font size increases. Consistent with that, Undorf and Zimdahl reported that the processing fluency

was lower for very small (≤ 17 pt) and very large (≥ 161 pt) font sizes than for intermediate (17–161 pt) font sizes. Similar to Luna et al. (2018) and Halamish (2018), they reported that memory increased with font size, but to a much smaller extent than JOLs did. More importantly, although JOLs increased monotonically with font size, memory was overall better for very small and very large fonts than for intermediate ones. In the other three experiments, Undorf and Zimdahl detected the similar patterns using four different font sizes (9 pt., 29 pt., 93 pt., 294 pt).

In summary, recent evidence has challenged the classic finding that font size affects only JOLs but not memory, which in turn challenges the conventional view that the font size effect is merely a metacognitive illusion. Rather than dissociating memory from JOLs, it seems that memory may just be less sensitive to variations in the size of encoded words than JOLs are. Moreover, the nature of the size-memory relation seems to be more complex (nonmonotonic) than the nature of the size-JOL relation.

Potential moderators that affect the size-JOL and size-memory relations

To understand the causes of the font size effect, it is necessary to manipulate variables that may moderate it. The literature on potential moderators was examined in prior meta-analyses by Luna et al. (2018) and Halamish (2018). To facilitate comparisons between the current meta-analysis and previous ones, we included all the variables discussed in Luna et al.'s and Halamish's meta-analyses (font size comparison, JOL type, test format, study time) plus additional variables for which data have recently become available (stimulus type, belief instruction, experimental context, experimental design, publication bias). We define each moderator variable below and briefly sketch its theoretical or empirical background. Later, in the Method section, we discuss in more detail how each variable was coded for the current meta-analyses.

Font size comparison

Rhodes and Castel (2008) used 18 pt. versus 48 pt. in the initial font size experiments, and hence, many studies have implemented the same comparison. However, many other font size comparisons have been administered, too. Halamish (2018) is the only study that examined font size comparison as a moderator of the font size effect, but her study was limited to 5-pt versus 18-pt versus 48-pt fonts. In the current meta-analyses, we include a much broader range of font size comparisons, which were classified into three categories based on the fluent range of print size (17–161 pt.; Legge & Bigelow, 2011; Undorf & Zimdahl, 2019): (a) one intermediate font size (17–161 pt) compared with a very small font size (≤ 17 pt), where the intermediate font size is processed *more* fluently than the very small font size, (b) two font sizes within the intermediate range compared with each other, which are processed with *comparable* fluency, and (c) a very large font size (≥ 161 pt) compared with an intermediate font size, where the very large font size is processed *less* fluently than the intermediate font size (see Fig. 1 for a graphical illustration). By linking font size comparison to the fluent range of print size, we were able to examine the contribution of fluency to the font size effect.

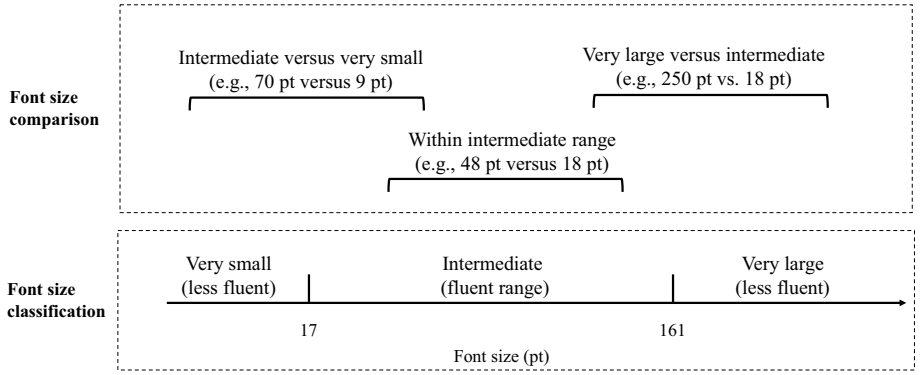

Fig. 1 Graphic illustration of font size comparisons classified based on the fluent range of print size (Legge & Bigelow, 2011; Undorf & Zimdahl, 2019)

JOL type

One of the most robust findings in the JOL literature is that delayed JOLs are better predictors of memory performance than immediate JOLs (Rhodes & Tauber, 2011). Therefore, it is possible that the dissociation between JOLs and memory induced by font size may be confined to immediate JOLs. Luna et al. (2018) tested this hypothesis and reported that the effects of font size on JOLs were significantly larger with immediate than with delayed JOLs, whereas the memory effects of font size did not vary with the JOL type.

In addition, some font size experiments administered no JOL tasks during encoding (e.g., Bodner et al, 2020; Taikh & Bodner, 2016). Halamish (2018) compared the memory effects of font size when JOLs were solicited versus not solicited. She found that the memory advantages of 5 pt. and 48 pt. relative to 18 pt. were only significant in the no-JOL condition. Further, 48-pt words were easier to remember than 5-pt words when JOLs were solicited, but the opposite was true in the no-JOL condition. Accordingly, Halamish speculated that the memory advantage of very small fonts may only occur without the solicitation of JOLs. To sum up, the font size effect may be affected by whether immediate, delayed or no JOLs are administered.

Test format

Prior research has demonstrated that recognition is typically more sensitive to manipulations of perceptual characteristics (e.g., visual interference and blurring) than free recall (Hirshman & Mulligan, 1991; Rosner et al., 2015). As font size is a perceptual characteristic, its effect may also depend on test format, too. In that connection, Halamish (2018) used free recall in five of her 12 experiments and recognition in the other seven. She found that test format did not significantly moderate the effect of font size on either JOLs or memory. However, Luna et al.'s moderator analysis revealed that the tendency of font size to affect memory performance was significant in free recall but not in cued recall, although the difference between the two test formats did not reach statistical significance. In brief, currently there is no evidence that the effect of font size differs for free recall versus recognition, but there is weak evidence that it may differ for free versus cued recall. The richer

data contained in the current analysis will help to determine whether the prior patterns are due to low statistical power or statistical aberration.

Study time

The amount of time that is provided to encode items has been shown to modify the effects of perceptual manipulations. For instance, Yue et al. (2013) reported that memory for clear words was better than for blurred words at 2 s study times but not at 5 s. In addition, Geller (2017) found that cursive words were recalled better than normally-printed words at 2 s study times but not at 500 ms. Therefore, Halamish hypothesized that study time would moderate effects of font size in the same manner as other perceptual manipulations. However, her results show the opposite: Study time did not influence how font size affects either JOLs or memory accuracy. Nevertheless, it should be noted that Halamish only examined study times of 0.5 s versus 5 s, whereas most font size experiments have used study time ranging from 2 s to 5 s. Therefore, it remains uncertain whether the conclusion would be the same with those more commonly used study times.

Stimulus type

Although stimulus type has not been systematically investigated in font size experiments, several studies that used word pair or sentence stimuli have found smaller or even null effects of font size on JOLs, relative to canonical word list studies (e.g., Ball et al., 2014; Double, 2019; Luna, Albuquerque, & Martín-Luengo, 2019a; Price & Harrison, 2017). It is worth noting that the number of cues on which JOLs can be based increases as the complexity of the target items increases. For instance, word pairs and sentences provide more cues than single words. Moreover, participants may emphasize different types of cues when processing different stimulus types, such that they focus more on item-specific cues with word lists and relational cues with word pairs and sentences. Therefore, it is essential to examine whether stimulus type, which is associated with different numbers and types of cues, influences the font size effect. Moreover, it is noteworthy that test format is often confounded with stimulus type, such that single-word lists are usually followed by free recall tests, whereas word pairs are usually followed by cued recall. Accordingly, examining stimulus type as a moderator can potentially resolve uncertainties in the findings about test format.

Belief instruction

As fluency versus belief is an important point of theoretical contention in JOL research, we saw that some investigators have studied people's beliefs about font size. Their results showed that people believe that items presented in larger fonts are more fluently processed and easier to remember (Hu et al., 2015; Kornell et al., 2011; Mueller et al., 2014; Su et al., 2018). In light of such findings, subsequent studies have implemented explicit instructions that manipulate people's beliefs. When participants were given instructions that reinforced the belief that larger fonts are more fluent or more memorable, the typical font size effect on JOLs was detected. However, when the opposite instructions were given, such that

smaller fonts are more fluent or more memorable, this effect was sometimes reduced or even eliminated (Blake, 2018; Blake & Castel, 2018; Chen et al., 2019; Wang, Yang, et al., 2020b). These findings seem to tie the font size effect directly to people's beliefs, and the current meta-analysis provides the first systematic examination of these findings.

Experimental design

In most studies, font size was manipulated within subjects. In a few studies, however, font size was manipulated between subjects (e.g., Peynircioğlu & Tatz, 2019; Susser et al., 2013). According to the cue-utilization framework (Koriat, 1997), JOLs are inferential in nature, and hence, people rely heavily on variability in the perceived memorability of different items to make such judgments. In that regard, as JOLs may depend on certain comparative processes participants engage in, a natural hypothesis is that the font size effect should be more robust in within-subject designs than in between-subject designs, which we tested in the current analysis.

Experimental context

Font size experiments have been conducted in traditional laboratory settings and also in more informal online settings. First, in laboratory settings, the actual font sizes of study materials were under rigorous experimental control. However, in online settings, although the larger-to-smaller ratio was always controlled, items' actual sizes would depend on the sizes of participants' monitors and their browser settings. The latter fact means that there is an additional source of noise in online studies, compared to laboratory studies. Second, as the online setting became more prevalent during the Covid 19 pandemic, it seemed important to investigate whether laboratory and online studies yielded the same empirical patterns.

Publication status

To avoid publication bias, it has been recommended that unpublished papers, the so-called "gray literature," should be included in meta-analyses (Rhodes & Tauber, 2011; Rothstein et al., 2006), although this practice is not without criticism (Cook et al., 1993; Schmucker et al., 2017). In the present case, we have included unpublished theses or dissertations and preprints in addition to published journal articles. Therefore, we included publication status as a potential moderator to examine whether any of our conclusions would change as a function of whether data have been published pursuant to peer review.

The current meta-analyses

The objective of the following meta-analyses is to provide a comprehensive quantitative summary of the accumulated literature on the font size effect. We sought to identify moderator variables that would resolve current uncertainties in the literature, that would provide differential evidence on extant theoretical accounts of font size effects, and that

would provide some clear directions for future research. The present meta-analyses went beyond the earlier ones by Halamish (2018) and Luna et al. (2018) in three important ways. First, we covered a much larger number of studies. Halamish conducted a mini meta-analysis of her own 12 experiments, and Luna et al. (2018) covered data from 28 experiments. The current meta-analyses included more than twice as many studies as those two prior reviews. Specifically, we included 93 experiments in the meta-analysis of JOLs and 103 experiments in the meta-analysis of memory accuracy.

The second distinctive feature of our meta-analyses is that we analyzed a larger and more sophisticated set of moderators. As mentioned, prior meta-analyses have examined potential moderators including font size comparison, test format, study time, and JOL type. We re-examined all these variables with a larger database. Here, it is worth noting that we linked font size comparisons with the fluent range of print size (Legge & Bigelow, 2011; Undorf & Zimdahl, 2019), which provided a means of evaluating the fluency hypothesis, and we have included more commonly used study times in the font size literature, which allowed us to re-examine the moderating effect of study time. Further, we were able to include more potential moderators, such as stimulus type, belief instruction, experimental design, experimental context, and publication status. The stimulus type, experimental design, and belief instruction variables are of particular interest: The former two are potential boundary conditions for the font size effect that have not been systematically examined, and the latter one allowed us to directly test the belief hypothesis.

The third distinctive feature of our meta-analyses is that we implemented both univariate and multivariate approaches in our moderator analyses. Both of the prior meta-analyses were restricted to univariate moderator analyses. The advantage of additional multivariate analyses is that we can examine each moderator's effect while controlling for the effects of other moderators. This allows us to simultaneously evaluate the relations among multiple moderators and identify the ones with the strongest effects (Black et al., 2016). Although such an approach has been recommended by various researchers (Harrer et al., 2019; Thompson & Higgins, 2002; Tipton et al., 2019), it is far from common practice in the literature, and for that reason, we also conducted a traditional univariate analysis.

Method

Experiment selection

The current meta-analyses included studies in which font size was manipulated, and either JOLs or memory tests or both were administered. Because Rhodes and Castel (2008) was the first study to investigate the font size effect, we began our literature search by identifying articles that cited this paper in both Google Scholar and Web of Science. We refined our search using the keyword "font size." We also conducted additional searches by retrieving relevant articles from the identified articles' reference lists. The step-by-step procedure is depicted in Fig. 2.

To be eligible for our meta-analyses, an experiment had to meet three criteria: (a) the font size of study materials had been manipulated; (b) either JOL tasks or memory tests or both had been administered; and (c) sufficient data had been reported to calculate effect sizes. If an experiment did not meet the last criterion, we contacted the authors to request the necessary data. When that was failed, we either removed the experiments or estimated

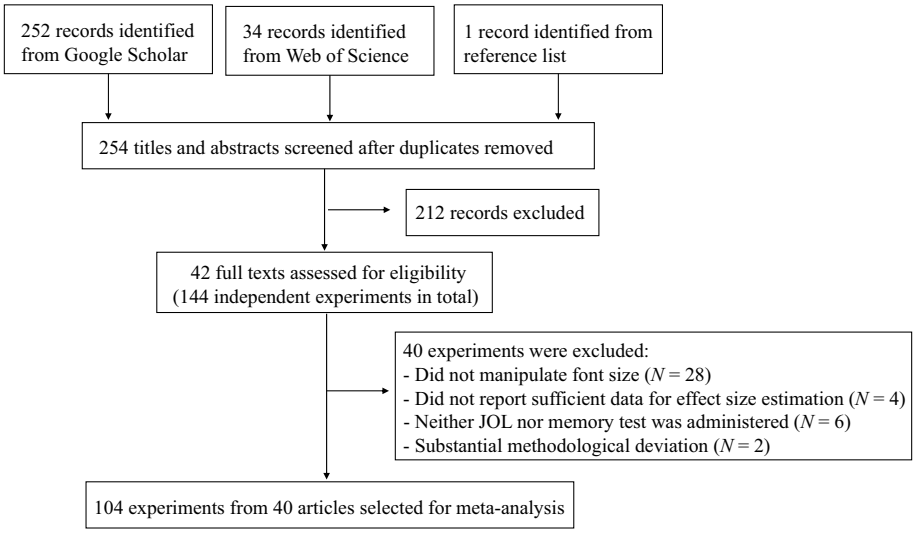

Fig. 2 Flow chart for the experiment selection procedure

the effect sizes using conservative methods based on the partial data reported (see Appendix for more details).

In the end, we included a total of 115 effect sizes from 93 experiments in the first meta-analysis of font size effects on JOLs. Next, we included a total of 132 effect sizes from 103 experiments in the second meta-analysis, for font size effects on memory. The effect sizes and the main methodological details for all the studies included in the meta-analyses can be found in the Supplementary Materials.

Statistical procedures

All analyses were conducted using the "metafor" packages in R (Viechtbauer, 2010). We followed the meta-analytic procedure recommended by Harrer et al. (2019). First, we calculated the effect sizes of all experiments using Hedges's *g* (Hedges, 1981). Second, we pooled the effect sizes using a random-effects model. Next, we checked the between-study heterogeneity to identify and examine outliers and influential cases. Then, we reran the random-effects model with influential outliers removed. After that, we used a mixed-effects model to conduct both univariate and multivariate meta-regressions for the potential moderators described above. Finally, we conducted publication bias analyses.

Effect size calculation

For experiments included in Luna et al.'s (2018) and Halamish's (2018) meta-analyses, we used the Hedges's *g* values reported in those papers. For all the remaining experiments, we first calculated Cohen's *d* (Cumming, 2013; Ellis, 2010; Goulet-Pelletier & Cousineau, 2018) and then converted it to Hedges's *g* (Hedges, 1981). The formulas we used to calculate Cohen's *d* and Hedges's *g*, as well as the other statistical details in the effect size calculations, have been relegated to the Appendix.

Identification of outliers and influential cases

We followed Viechtbauer and Cheung's (2010) recommendations in identifying outliers and influential cases. The specific statistical criteria we implemented can be found in the Appendix. Meta-analyses were conducted with and without influential outliers, but subsequent moderator analyses were conducted with outlier-free data.

Coding of potential moderator variables

Font size comparison Using Undorf and Zimdahl's (2019) classification of font sizes, we classified font sizes ≤17 pt. as very small, font sizes between 17 pt. and 161 pt. as intermediate, and font sizes ≥161 pt. as very large.[1] Accordingly, we coded font size comparison as a categorical variable with three levels: very small versus intermediate, within intermediate range, and intermediate versus very large (see Fig. 1).

JOL type In the meta-analysis for JOLs, we coded JOL type as a categorical variable with two levels: immediate JOL and delayed JOL. In the meta-analysis for memory performance, we coded JOL type as a categorical variable with four levels: immediate JOL, delayed JOL, pre-study JOL and no JOL. Pre-study JOL corresponds to Mueller et al. (2014; Experiment 4), in which JOLs were administered before each item with a prompt indicating the item would be presented in larger or smaller font. We did not include this study in the meta-analysis for JOLs, as we think pre-study JOL is conceptually not comparable to either immediate or delayed JOLs.

Test format Test format was coded as a categorical variable with four levels: free recall, cued recall, recognition, and other. Free recall is the most commonly used test format since Rhodes and Castel (2008), followed by cued recall and then recognition. One exception is that Mcdonough and Gallo (2012; Experiment 3) used a criterion recollection test and a two-forced-alternative-choice (2AFC) test, which was not comparable to any other studies. Therefore, we did not include memory data from this experiment in the meta-analysis for memory, and we grouped this experiment under the level "other" in the meta-analysis for JOLs.

Study time The variable study time was coded as a categorical variable with eight levels: .5 s, 2 s, 3 s, 4 s, 5 s, 8 s, self-paced, and other. The level of "other" corresponds to Luna et al. (2019a; Experiment 2 & 3), who presented related or unrelated sentences at 400 ms/word, so that the presentation time was not uniform across all items.

[1] For studies that administered more than two font sizes, we only included data for two font sizes in our analyses. In this case, we selected two font sizes that crossed the boundaries of the fluent range (17 pt., 161 pt) whenever possible, and we selected font sizes that are as comparable to other studies as possible. Here, we should acknowledge that the 17–161 pt. fluent range was established with a fixed viewing distance of 85 cm (Undorf & Zimdahl, 2019), whereas viewing distance was not standardized across studies. However, this should not threaten the validity of our classification for two reasons. First, Undorf and Zimdahl (2019; Experiment 2) found a similar U-shaped relation between lexical decision times and font sizes with and without chin-rests, suggesting that the quadratic font size-fluency relation remains robust even when viewing distance is not fixed. Second, most of the font sizes included in our meta-analyses fall intuitively into a certain category as they are far from the upper and lower limits of the fluent range (see Supplementary Materials for more details). Still, some font sizes fall near the boundaries of the fluent range, such as 18 pt. and 160 pt. In the former case, Undorf et al. (2017) demonstrated that 18 pt. and 48 pt. belong to the fluent range even when viewing distance varies between 25 and 95 cm. In the latter case, we grouped such font sizes under both categories across the boundary (e.g., either intermediate or very large), and we found no change in the results.

Stimulus type The variable stimulus type was coded as a categorical variable with six levels: unrelated word list, related word pairs, unrelated word pairs, mixed word pairs (including both related and unrelated pairs), related sentences, and unrelated sentences.

Belief instruction We coded belief instruction as a categorical variable with three levels: (a) congruent, in which participants were given instructions that were consistent with their natural beliefs that larger items are more fluent or easier to remember; (b) incongruent, in which participants were given instructions that were opposite to their natural beliefs; and (c) control, in which participants were given no instructions about font sizes.

Experimental design The experimental design variable was coded as a categorical variable with two levels: between-subject and within-subject.

Experimental context The experimental context variable was coded as a categorical variable with two levels: laboratory and online.

Publication status Publication status was coded as a categorical variable with two levels: published and unpublished.

Univariate and multivariate meta-regressions

We first followed the traditional univariate approach of examining one moderator at a time. For each moderator, we ran a mixed-effects meta-regression. Next, we proceeded to the multivariate meta-regressions. Although the multivariate approach has obvious benefits, it poses certain validity threats, such as over-fitting and multicollinearity (Berlin & Antman, 1992; Harrer et al., 2019; Higgins & Thompson, 2004). To avoid the over-fitting problem, we minimized the number of predictor variables by only including moderators that were statistically significant in the univariate meta-regressions (Black et al., 2016; Harrer et al., 2019). As for the multicollinearity problem, given that there is no gold-standard method of diagnosing multicollinearity with categorical variables, we considered multiple sources of information, including the generalized variance inflation factor (GVIF; Fox & Monette, 1992), χ^2 tests of independence, and the sensitivity of the model coefficients to specific moderator variables. Variables that met the following criteria were inspected and removed when necessary: (a) GVIF >10; (b) correlated with multiple other moderators in the χ^2 tests; and (c) produced substantial changes in model coefficients when removed.

Examination of publication Bias

Publication bias, also called the "file-drawer" problem, is the hypothesis that studies that report larger effect sizes or achieved statistical significance are more likely to be published than those reporting smaller effect sizes or failing to achieve statistical significance (Rothstein et al., 2006). According to this hypothesis, published studies may not be representative of all studies on a target topic. Fortunately, diagnostic analyses for publication bias have been developed, such as the funnel plot and Egger's regression

test of funnel plot asymmetry.[2] Our aim of the publication bias analyses is to determine whether the font size studies included in our meta-analyses are representative of the literature on this topic. Thus, we included both published and unpublished studies in the publication bias analyses and applied the aforementioned diagnostic methods to the data.

Results

For both JOLs and memory performance, we first report the meta-analysis of all studies to examine whether the pooled effect sizes for font size were significantly different from zero. Then, we investigate the effects of the potential moderators with both univariate and multivariate meta-regressions. Last, we report the diagnostic results for publication bias.

Meta-analysis for JOLs

An effect size from Soderstrom (2012) was identified as an influential outlier and was excluded from the analysis. The meta-analysis for JOLs revealed a small-to-moderate effect of font size, $g = .38$, $SE = .02$, 95% CI = [.33, .43], $p < .001$. As shown in Fig. 3, items presented in larger fonts received higher JOL ratings than items presented in smaller fonts. In addition, there was relatively large heterogeneity in the sample, $Q(113) = 392.56$, $p < .001$, prediction interval [−.01, .77]. This means that there was substantial variability in the effects of font size on JOLs among the included studies. We also conducted the meta-analysis with the outlier retained, and the results were virtually identical to those results with outlier removed, $g = .39$, $SE = .03$, 95% CI = [.34, .44], $p < .001$.

Moderator analyses for JOLs

Univariate meta-regression

A summary of the moderator analyses is displayed in Table 1, with data columns 1–6 specifying the number of effect sizes in each level of the moderator (k), the mean weighted effect size (g), the standard error (SE), the 95% confidence interval, and the between-group heterogeneity statistics (Q_{BU}), which indicate whether there is a significant difference among the levels of the categorical variable.

First, across the three types of font size comparisons, larger-font items were always given higher JOL ratings than smaller-font items, as the gs were all positive and significant. In addition, the effect size was significantly larger for the comparison between very large and intermediate fonts ($g = .64$, $p < .001$) than for the comparison between intermediate and

[2] A funnel plot is a scatter plot with effect size estimates of individual studies plotted on the abscissa and some measure of size or precision (typically standard error) for the corresponding studies plotted on the ordinate (Sterne et al., 2011). Studies with larger standard errors are placed at the bottom of the ordinate and those with smaller standard errors are placed in the top. If there is no publication bias, studies with larger effect sizes should be less dispersed compared to studies with smaller effect sizes, making the distribution look like a symmetrical funnel. In addition, we have also created contour-enhanced funnel plots, which takes statistical significance into consideration (Peters et al., 2008). Apart from visual inspection, the asymmetry of funnel plots can be statistically evaluated with the Egger's test (Egger et al., 1997).

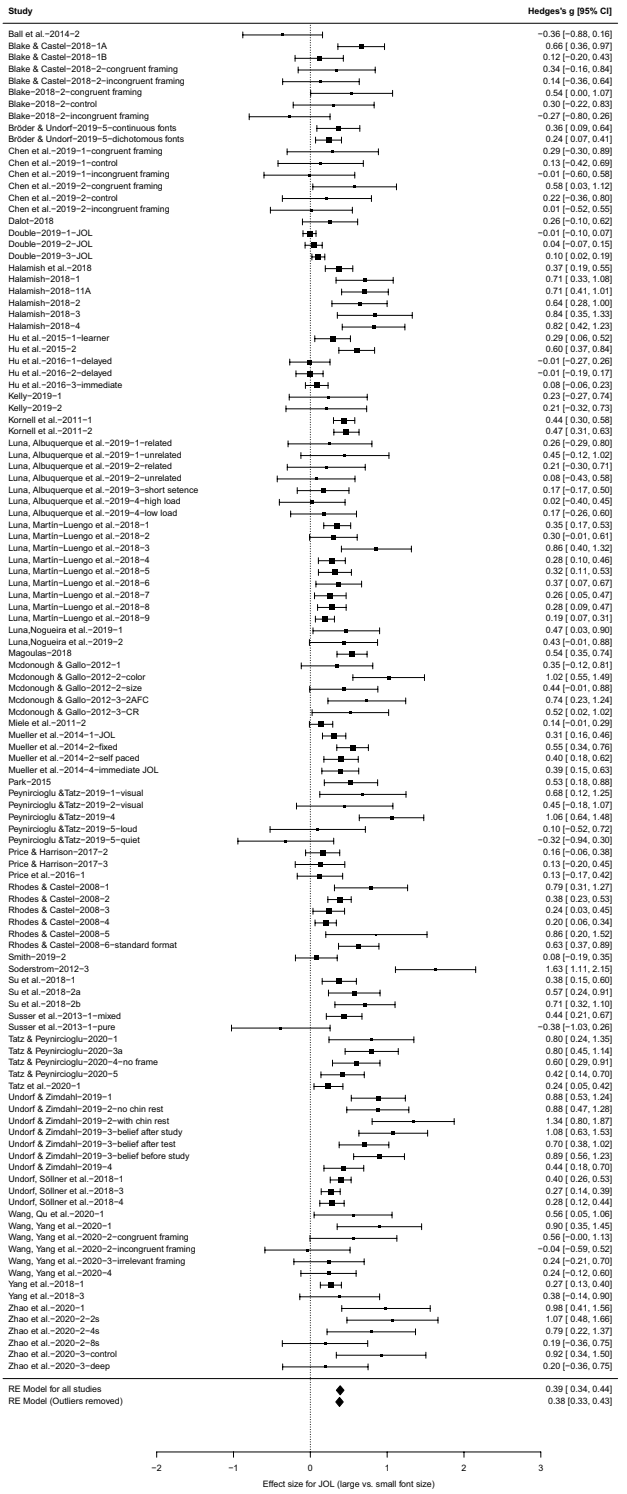

Fig. 3 Forest plot with effect sizes and 95% confidence intervals for the comparison between larger and smaller fonts in JOLs. Positive effect sizes indicate that items presented in larger fonts were rated with higher JOLs than items presented in smaller fonts. The sizes of the squares in the forest plot are proportional to the weights of the experiments, which are calculated as the inverse sampling variances. RE Model = random-effects model

Table 1 Summary of moderator analyses for the effects of font size on judgments of learning

Variable				95% CI			
	k	*g*	*SE*	lower	upper	Q_{BU}	Q_{BM}
Font size comparison						12.49**	7.46*
Intermediate vs. Very small	28	.39***	.05	.29	.48		
Within intermediate range	75	.34***	.03	.28	.40		
Very large vs. Intermediate	11	.65***	.08	.49	.82		
JOL type						5.29*	6.47*
Immediate	108	.40***	.03	.35	.45		
Delayed	6	.18	.09	−.002	.36		
Test format						21.53***	–
Free recall	81	.41***	.03	.36	.47		
Recognition	7	.59***	.10	.40	.79		
Cued recall	24	.17**	.05	.06	.28		
Other	2	.63**	.22	.19	1.07		
Stimuli type						42.94***	26.96***
Unrelated word list	85	.45***	.02	.40	.49		
Related word pair	3	.11	.10	−.09	.30		
Unrelated word pair	15	.14*	.06	.03	.25		
Mixed word pair	3	.18	.11	−.04	.40		
Related sentence	3	.03	.18	−.31	.38		
Unrelated sentence	5	.16	.12	−.07	.40		
Study time						42.25***	–
.5 s	3	.78***	.15	.49	1.06		
2 s	17	.65***	.06	.53	.77		
3 s	13	.30***	.06	.19	.40		
4 s	16	.28***	.06	.17	.40		
5 s	45	.36***	.04	.29	.43		
8 s	4	.13	.09	−.06	.31		
Other	3	.15	.16	−.16	.46		
Self-paced	13	.37***	.06	.25	.49		
Belief instruction						7.41*	7.30*
Congruent	5	.46**	.15	.16	.76		
Incongruent	6	.02	.14	−.24	.29		
Control	103	.39***	.03	.34	.44		
Experimental design						4.90*	6.94**
Between-subject	4	−.04	.19	−.41	.34		
Within-subject	110	.39***	.03	.34	.44		
Experimental context						7.64**	2.59
Laboratory	103	.40***	.02	.35	.45		
Online	11	.19**	.07	.06	.33		
Publication status						7.14**	1.79
Published	105	.40***	.02	.35	.44		
Unpublished	9	.19**	.07	.05	.33		
Memory effect size	112	–	–	–	–	11.81***	.22

Table 1 (continued)

Note. k = number of effect sizes; g = Hedges's g; SE = standard error; CI = confidence interval; Q_{BU} = heterogeneity between subgroups in the univariate meta-regression, which indicates whether there was a significant difference between the different levels of the categorical variable; Q_{BM} = heterogeneity between subgroups in the multivariate meta-regression; Very small = font sizes equal to or smaller than 17 pt.; Intermediate = font sizes between 17 pt. and 161 pt.; Very large = font sizes equal to or larger than 161 pt.; JOL = judgment of learning; Mixed word pair = a mixture of related and unrelated word pairs; Congruent = participants were given instructions that are congruent with their general beliefs, such that larger words are more fluent or easier to remember or both; Incongruent = participants were given instructions that are incongruent with their general beliefs, such that larger words are less fluent or harder to remember or both; Control = no instructions about the effects of font size were given. Memory effect size = the sizes of the effects of font size on memory performance

* = $p < .05$; ** = $p < .01$; *** = $p < .001$

very small fonts ($g = .39$, $p < .001$) and for the comparison between two font sizes within the intermediate range ($g = .34$, $p < .001$), $Q_{BU} = 12.49$, $p = .002$. The difference between the latter two effect sizes was not reliable.

Meanwhile, the analysis for the moderator effect of JOL type showed that the effect of font size varied significantly for immediate versus delayed JOLs, $Q_{BU} = 5.29$, $p = .022$. Specifically, the effect size was significant for immediate JOLs ($g = .40$, $p < .001$), but it was reduced and only marginally significant for delayed JOLs ($g = .18$, $p = .053$).

The moderator analysis for test format showed that the effect size was larger for free recall ($g = .41$, $p < .001$) or recognition tests ($g = .59$, $p < .001$) than for cued recall tests ($g = .17$, $p = .002$), $Q_{BU} = 21.53$, $p < .001$. Here, it is important to remind ourselves that memory tests were not administered until after JOLs. Accordingly, it is more likely that the test format effects are due to differences in the effects of encoding individual words (in free recall and recognition experiments) versus encoding word pairs (in cued recall experiments) during the study phase.

Indeed, we found that there was a significant moderator effect of stimulus type, $Q_{BU} = 42.94$, $p < .001$. The effect sizes were reliable for unrelated word lists ($g = .45$, $p < .001$) and for unrelated word pairs ($g = .14$, $p = .011$) but not for other types of materials (gs range from .03 to .18, ps > .102). The effect size was significantly larger for unrelated word lists than for all the remaining types of stimuli, whereas no significant difference was found among the remaining types of stimuli.

Next, we found a reliable moderator effect of study time. The effect sizes were significantly larger when study times were shorter (.5 s and 2 s; gs = .78 and .65) than when they were longer (3 s, 4 s, 5 s, 8 s; gs = .30, .28, .36, and .13) or when they were self-paced ($g = .37$), $Q_{BU} = 42.25$, $p < .001$. Except for 8 s, the effect sizes were reliable for all of the other study times (ps < .001).

Belief instruction also proved to be a significant moderator of the font size effect on JOLs, $Q_{BU} = 7.41$, $p = .025$. The effect size was significant when participants were informed that larger words were more fluently processed or more memorable ($g = .46$, $p = .002$) and when they were not given any instruction about font size ($g = .39$, $p < .001$). These two effect sizes did not differ reliably. However, when participants were informed that larger words were less fluently processed or less memorable, the effect size was far smaller and unreliable ($g = .02$, $p = .868$), indicating that larger words no longer produce higher JOLs than smaller words.

In addition, there was a reliable moderator effect for experimental context, $Q_{BU} = 4.90$, $p = .027$. The effect size was significant when font size was manipulated within subjects ($g = .39$, $p < .001$) but not when it was manipulated between subjects ($g = -.04$, $p = .845$).

Interestingly, the moderator effects of experimental design and of publication status were also reliable. The effect size was larger in laboratory settings ($g = .40$, $p < .001$) than in online settings ($g = .19$, $p = .005$), $Q_{BU} = 7.64$, $p = .006$, and it was also larger in published ($g = .40$, $p < .001$) than in unpublished papers ($g = .19$, $p = .008$), $Q_{BU} = 7.14$, $p = .008$.

Last, we ran an additional univariate meta-regression using the effect size of font size on memory as a continuous moderator.[3] This was meant to examine the effects of font size on JOLs *relative to* memory. Two memory effect sizes from Mcdonough and Gallo (2012; Experiment 3) were removed from this analysis because their memory tests were not comparable to any other studies. The results showed that the moderator effect of memory effect size was significant, $Q_{BU} = 11.81$, $p < .001$. In the meta-regression model, the intercept is .34, $p < .001$, and the regression coefficient for the memory effect size is .50, $p < .001$. This suggests that the sizes of the JOL effects were reliably higher compared to the memory effect size, and the JOL effects increased as the memory effects increased.

Multivariate meta-regression

All potential moderators were found to have significant effects in the univariate meta-regression. Therefore, we included all of them in the multivariate meta-regression. In addition, we included the memory effect size as a continuous moderator in the multivariate meta-regression, so as to examine the pure JOL effects of the moderators while controlling for memory effects.[4] In a preliminary analysis, three variables, study time, stimulus type and test format, were identified as variables that caused multicollinearity among the moderators. All three variables had GVIFs >20 and produced substantial changes in model coefficients when removed. In addition, χ^2 tests revealed that study time was correlated with five other moderators, whereas stimulus type and test format were mainly correlated with each other. The close relationship between stimulus type and test format is not surprising inasmuch as word lists were always followed by free recall or recognition, whereas word pairs were always followed by cued recall. Consequently, we removed study time and test format from the multivariate model. We retained stimulus type instead of test format because participants were not necessarily informed of the test format when making JOLs, but they were always aware of the stimulus type.

We reported the between-group heterogeneity statistics (Q_{BM}) in the 7th column of Table 1, which is simply the multivariate counterpart of Q_{BU}. As can be seen there, the moderator effect of the font size comparison ($Q_{BM} = 7.46$, $p = .024$) remained reliable when controlling for the other variables, as did the effects of JOL type ($Q_{BM} = 6.47$, $p = .011$), stimulus type ($Q_{BM} = 26.96$, $p < .001$), belief instruction ($Q_{BM} = 7.30$, $p = .026$) and experimental design ($Q_{BM} = 6.94$, $p = .008$). However, the effects of experimental context ($Q_{BM} = 2.59$, $p = .107$) and publication status ($Q_{BM} = 1.79$, $p = .181$) were no longer reliable, indicating that the univariate effects of online versus laboratory experiments and of published versus unpublished experiments could be accounted for by other moderator variables. Similarly, the moderator effect of memory effect size was no longer significant ($Q_{BM} = .22$, $p = .642$), suggesting that the JOL-memory relation was constrained by the other moderator variables. In other words, JOL effects of font size only vary as a function of memory effects within certain levels of those moderator variables.

[3] We thank an anonymous reviewer for this suggestion.

[4] The qualitative pattern in the multivariate meta-regression results remained the same with or without the continuous moderator of memory effect size.

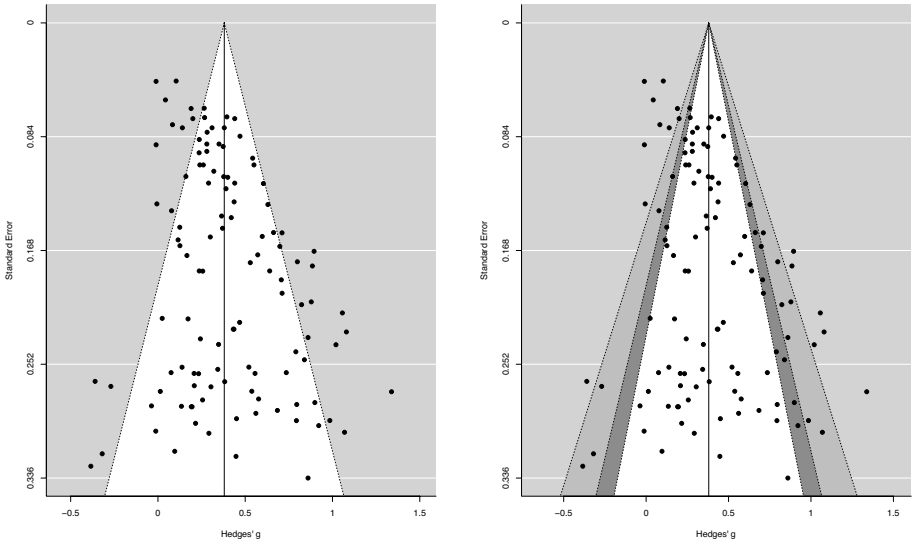

Fig. 4 Standard funnel plot (left panel) and contour-enhanced funnel plot (right panel) for the meta-analysis of the font size effect on JOLs. In both funnel plots, the effect estimates (Hedges's *g*) were plotted against standard errors from the same studies. In the contour-enhanced funnel plot, the white area corresponds to $p > .10$, the dark gray-shaded area corresponds to $.05 < p < .10$, the medium gray-shaded area corresponds to $.01 < p < .05$, and the area outside the funnel corresponds to $p < .01$

Publication bias analysis for JOLs

Funnel plots were generated for data from studies included in the meta-analysis for JOLs (Fig. 4). As can be seen in the left panel, the funnel plot appeared to be asymmetrical, which was confirmed by Egger's test, $z = 3.57$, $p < .001$. However, when we examined the contour-enhanced funnel plot (Fig. 4, right panel), we saw that the studies were evenly distributed between the regions of statistical significance versus statistical insignificance. This indicates that experiments that are readily available were not more likely to produce significant font size effects than experiments that are not readily available. Here, it is important to note that funnel plot asymmetry can result from factors other than publication bias, such as heterogeneity (Sterne et al., 2011). Considering that between-experiment heterogeneity was quite substantial, $Q(113) = 393.38$, $p < .001$, it is very likely that the asymmetry in the left panel of Fig. 4 was due to high heterogeneity rather than publication bias. As a follow up, we ran an additional fixed-effects meta-analysis for JOLs, a procedure that was recommended by Sterne et al. (2011). There was still a robust font size effect, $g = .29$, $SE = .01$, 95% CI = [.37, .41], $p < .001$, although the effect size was smaller than in the random-effects model. Given that the fixed-effects model did not yield a different conclusion than the random-effects model, we focus on the random-effects model results in the remainder of the paper.

Meta-analysis of memory performance

Three effect sizes from Undorf et al. (2018; Experiment 4), Double (2019; Experiment 3, no-JOL condition), and Tatz and Peynircioğlu (2020; Experiment 4) were identified as influential outliers and were excluded from the meta-analysis of memory performance. As shown in Fig. 5, this meta-analysis indicated that memory for larger-font items was slightly better than for smaller-font items, $g = .05$, $SE = .02$, 95% CI = [.02, .08], $p < .001$. The heterogeneity in the sample was quite low, $Q(128) = 124.92$, $p = .561$, prediction interval [.02, .08], suggesting that the effect sizes are relatively homogeneous among the studies. When the three influential outliers were not removed from the meta-analysis, the results were almost the same, $g = .06$, $SE = .02$, 95% CI = [.03, .10], $p < .001$.

Moderator analyses for memory performance

Univariate meta-regression

The univariate meta-regression results for memory performance are displayed in columns 2–6 of Table 2. Similar to the results for JOLs, the effect of font size was moderated by font size comparison, $Q_{BU} = 14.24$, $p < .001$. The effect size was largest when very large fonts were compared to intermediate fonts ($g = .22$, $p < .001$), and it was smaller but still reliable when the two font sizes being compared were both within the intermediate range ($g = .06$, $p = .002$). In contrast to the results for JOLs, the effect size for the comparison of intermediate versus very small fonts was minuscule and not reliable ($g = -.002$, $p = .942$).

JOL type was also a reliable moderator of the font size effect on memory, $Q_{BU} = 12.92$, $p = .005$. The effect sizes did not differ reliably between immediate-JOL ($g = .08$, $p < .001$) and delayed-JOL conditions ($g = .06$, $p = .248$), although only the former effect size was reliable. Interestingly, the font size effect was reversed (i.e., better memory for smaller words) when no JOLs were elicited ($g = -.09$, $p = .034$).

In addition, we found that the effect sizes varied significantly across the stimulus types ($Q_{BU} = 12.70$, $p = .026$). Specifically, the effect size was reliable when participants studied unrelated word lists ($g = .07$, $p < .001$), but not with any other stimulus type (gs range from −.07 to .10, ps > .060). This was broadly consistent with the earlier stimulus type results for JOLs.

Finally, there was a significant moderator effect for study time, $Q_{BU} = 18.24$, $p = .011$. The effect sizes were only reliable with relatively short study times (2 s and 3 s; gs = .18 and .14, ps < .002), and the effect sizes for those study times were significantly larger than when study time was self-paced ($g = .02$, $p = .569$). This is also congruent with the earlier results for JOLs, in which the effect sizes were largest with relatively short study times.

Multivariate meta-regression

We included all four moderators that survived the univariate meta-regression in the multivariate model. As before, study time had a GVIF >10, correlated with all the other variables in χ^2 tests, and produced dramatic changes in the model coefficients when deleted. All the results pointed to the conclusion that study time was causing the multicollinearity

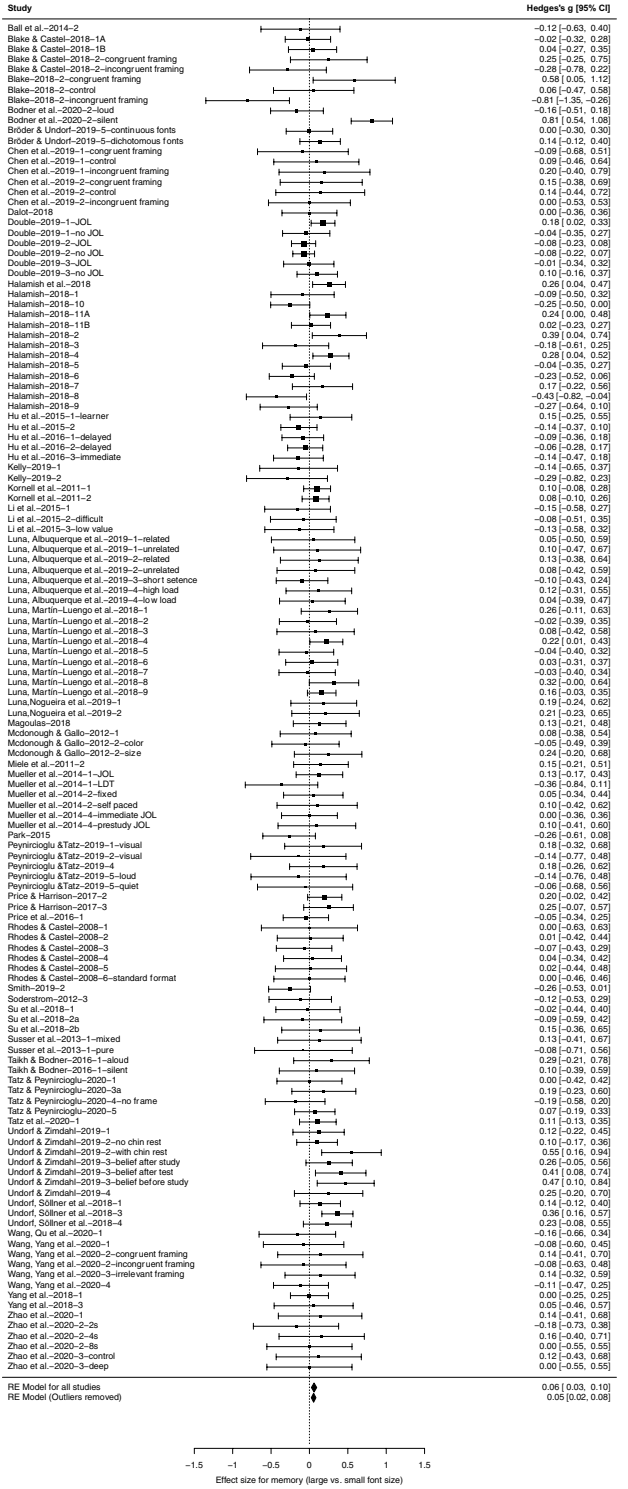

Fig. 5 Forest plot with effect sizes and 95% confidence intervals for the comparison between larger and smaller fonts in memory performance. Positive effect sizes indicate that items presented in larger fonts were memorized better than items presented in smaller fonts. The sizes of the squares in the forest plot are proportional to the weights of the experiments, which are calculated as the inverse sampling variances. RE Model = random-effects model

Table 2 Summary of moderator analyses for the effects of font size on memory performance

Variable				95% CI			
	k	*g*	*SE*	lower	upper	Q_{BU}	Q_{BM}
Font size comparison						14.24***	7.82*
Intermediate vs. Very small	37	−.002	.03	−.05	.05		
Within intermediate range	81	.06**	.02	.02	.10		
Very large vs. Intermediate	11	.22***	.05	.11	.32		
JOL type						12.92**	10.54*
Immediate	106	.08***	.02	.04	.11		
Delayed	6	.06	.05	−.04	.17		
Pre-study	1	.10	.26	−.41	.60		
None	16	−.09*	.04	−.18	−.01		
Test format						3.47	–
Free recall	87	.07***	.02	.03	.11		
Recognition	15	.05	.04	−.03	.14		
Cued recall	27	−.01	.04	−.08	.07		
Stimuli type						12.70*	13.49*
Unrelated word list	95	.07***	.02	.04	.11		
Related word pair	5	.10	.06	−.02	.21		
Unrelated word pair	18	−.07	.04	−.15	.01		
Mixed word pair	3	.16	.08	−.01	.32		
Related sentence	3	.02	.15	−.28	.32		
Unrelated sentence	5	.02	.10	−.17	.22		
Study time						18.24*	–
.5 s	8	.01	.05	−.09	.12		
2 s	18	.18***	.04	.09	.26		
3 s	12	.14**	.04	.05	.23		
4 s	16	−.04	.05	−.13	.05		
5 s	50	.02	.03	−.03	.08		
8 s	5	.08	.05	−.01	.17		
Other	3	−.001	.13	−.25	.24		
Self-paced	17	.02	.04	−.06	.10		
Belief instruction						4.74	–
Congruent	5	.22	.12	−.02	.46		
Incongruent	6	−.13	.11	−.35	.08		
Control	118	.06***	.02	.02	.09		
Experimental design						1.02	–
Between-subject	4	−.11	.16	−.42	.21		
Within-subject	125	.06***	.02	.03	.09		
Experimental context						.005	–
Laboratory	116	.05**	.02	.02	.09		
Online	13	.05	.04	−.02	.12		
Publication status						3.15	–
Published	117	.07***	.02	.03	.10		
Unpublished	12	−.01	.04	−.09	.07		

Table 2 (continued)

Note. k = number of effect sizes; g = Hedges's g; SE = standard error; CI = confidence interval; Q_{BU} = heterogeneity between subgroups in the univariate meta-regression, which indicates whether there was a significant difference between the different levels of the categorical variable; Q_{BM} = heterogeneity between subgroups in the multivariate meta-regression; Very small = font sizes equal to or smaller than 17 pt.; Intermediate = font sizes between 17 pt. and 161 pt.; Very large = font sizes equal to or larger than 161 pt.; JOL = judgment of learning; Mixed word pair = a mixture of related and unrelated word pairs; Congruent = participants were given instructions that are congruent with their general beliefs, such that larger words are more fluent or easier to remember or both; Incongruent = participants were given instructions that are incongruent with their general beliefs, such that larger words are less fluent or harder to remember or both; Control = no instructions about the effects of font size were given

* = $p < .05$; ** = $p < .01$; *** = $p < .001$

problem, and thus, we removed this variable from the model. The multivariate meta-regression results are displayed in the 7th column of Table 2. All of the three remaining variables, font size comparison ($Q_{BM} = 7.82$, $p = .020$), JOL type ($Q_{BM} = 10.54$, $p = .015$), and stimulus type ($Q_{BM} = 13.49$, $p = .019$), were reliable moderators of the font size effect on memory when the effects of other variables were controlled.

Publication bias analysis for memory performance

Visual inspection of the standard and contour-enhanced funnel plots for the meta-analysis of memory performance revealed no asymmetry (Fig. 6). The Egger's test confirmed that there is no significant asymmetry in the funnel plot, $z = -1.62$, $p = .106$, indicating that the studies included in our studies should be representative of the literature on the font size effect. For readers' interests, we still ran a fixed-effects meta-analysis for the font size effect on memory, and the results were identical to those of the random-effects analysis, $g = .05$, $SE = .02$, 95% CI = [.02, .08], $p < .001$.

Additional analyses

The moderator effects of font size comparison suggest that the JOL effects of font size were greatest in the very large versus intermediate comparison. Why? A plausible explanation is that the mean absolute font size difference was greater in this comparison than in the other two (e.g., the absolute font size difference in 250 pt. versus 18 pt. was larger than that in 48 pt. versus 18 pt. or in 70 pt. versus 9 pt.; see Fig. 1). Specifically, the mean absolute font size difference was 55.6 pt. for very small versus intermediate, 35.3 pt. within the intermediate range, and 259 pt. for very large versus intermediate. As the absolute font size difference was strongly correlated with the font size comparison, it is inappropriate to include it as an additional moderator in multivariate meta-regressions. Therefore, we converted it to a categorical variable with three levels: small, medium, and large difference (see Supplementary Materials for more details). Then, we conducted two additional analyses: (a) a univariate meta-regression with absolute font size difference as the moderator; and (b) a mediation analysis to test whether absolute font size difference mediated the effect of font size comparison on the JOL effect sizes. Concerning (a), the moderator effect of absolute font size difference was significant, $Q_{BU} = 21.83$, $p < .001$. The effect of font size on JOLs was largest when the absolute font size difference was large ($g = .64$, $p < .001$), followed by when it was medium ($g = .44$, $p < .001$),

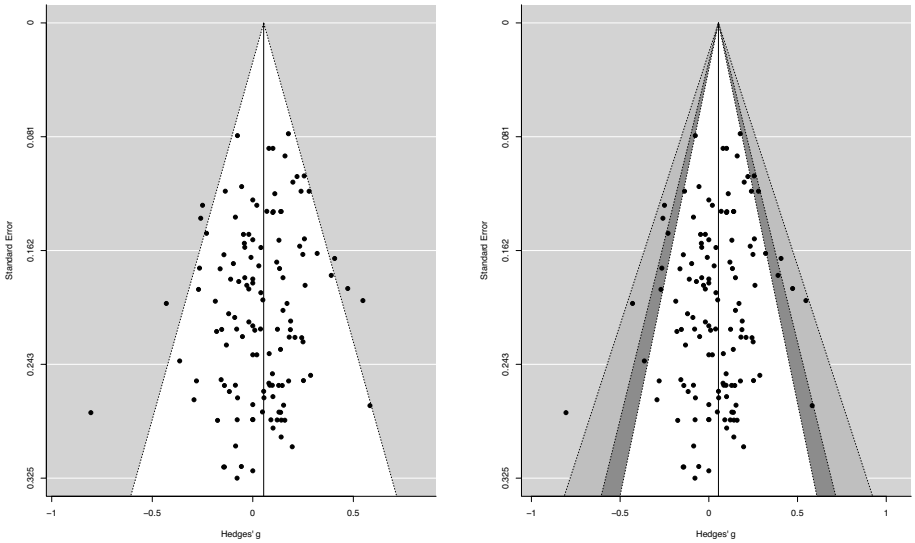

Fig. 6 Standard funnel plot (left panel) and contour-enhanced funnel plot (right panel) for the meta-analysis for font size effect on memory performance. In both funnel plots, the effect estimates (Hedges's *g*) were plotted against standard error from the same studies. In the contour-enhanced funnel plot, the white area corresponds to $p > .10$, the dark gray-shaded area corresponds to $.05 < p < .10$, the medium gray-shaded area corresponds to $.01 < p < .05$, and the area outside the funnel corresponds to $p < .01$

and when it was small ($g = .31$, $p < .001$). The difference in effect sizes was significant in all pairwise comparisons. As for (b), we found that absolute font size difference *fully* mediated the effect of font size comparison: The standardized indirect effect via absolute font size difference was reliable, $\beta = .10$, $p < .001$, but the direct effect was not, $\beta = .02$, $p = .624$ (see Fig. 7 Panel A). Together, these results indicated that the font size effect on JOLs increased as the absolute difference between font sizes increased, regardless of whether the larger fonts were processed more or less fluently than the smaller fonts.

Likewise, we also considered (a) whether absolute font size differences moderate the font size effect on memory, and (b) whether the effect of font size comparison on memory was mediated by absolute font size differences. Regarding (a), there was a significant moderator effect of absolute font size difference on memory, $Q_{BU} = 13.46$, $p = .001$. The effect sizes were positive and significant with large absolute font size difference ($g = .17$, $p < .001$) and with small absolute font size difference ($g = .06$, $p = .001$), but the effect size was negative and insignificant with medium absolute font size difference ($g = -.02$, $p = .473$). The difference in effect size was significant in all pairwise comparisons. Second, we found that absolute font size difference only *partially* mediated the effects of font size comparison. The standard indirect effect via absolute font size difference was significant, $\beta = .04$, $p = .152$, and so was the direct effect, $\beta = .05$, $p = .039$ (see Fig. 7 Panel B). Thus, the memory effect of font size cannot be fully accounted for by the absolute size difference. Unlike JOL effects, the memory benefits of larger relative to smaller fonts did not increase monotonically with the absolute size discrepancy between the two fonts.

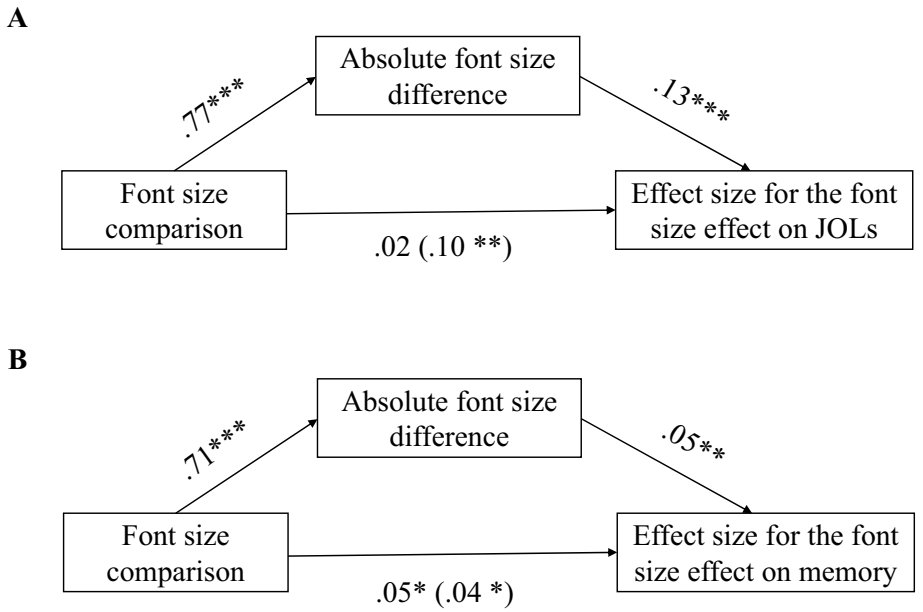

Fig. 7 Panel A = Standardized regression coefficients for the relationship between font size comparison and effect size for the font size effect on JOLs as mediated by absolute font size difference; Panel B = Standardized regression coefficients for the relationship between font size comparison and effect size for the font size effect on memory performance as mediated by absolute font size difference. Indirect effects were indicated in the parentheses. * = $p < .05$, ** = $p < .01$, *** = $p < .001$

Discussion

Our meta-analyses provide a detailed quantitative summary of the effects of font size on JOLs and on memory. When effect sizes were pooled over all studies, the meta-analysis for JOLs revealed a small-to-moderate font size effect, which firmly establishes that people judge larger words as being easier to remember than smaller ones ($g = .38$). Another overriding outcome is that people's tendency to judge that larger words are more memorable than smaller ones is not an illusion—larger words are in fact easier to remember overall. The meta-analysis for memory performance revealed that there was a small but highly reliable memory benefit for larger words relative to smaller words ($g = .05$). Although this finding is at odds with the canonical findings that font size affects only JOLs but not memory, it is not unprecedented (e.g., Halamish, 2018; Luna et al., 2018; Undorf & Zimdahl, 2019). For instance, Luna et al. (2018) conducted a meta-analysis of a much smaller set of studies and reported that although none of the studies had yielded a reliable memory effect of font size, memory was slightly more accurate overall for larger words ($g = .08$). Luna et al. conjectured that there may simply be a font size calibration difference between memory and JOLs such that JOLs are more sensitive to font size than memory is. Several of our results are consistent with that hypothesis, including (a) the font size effects for both JOLs and memory were highly reliable; (b) the JOL effect increased as the memory effect increased; and (c) the JOL effect was consistently much larger than the memory effect.

When it comes to explaining the font size effect, the moderator analyses supply critical results that generated considerable grist for the explanatory mill, as several variables were found to moderate font size effects on JOLs and on memory (see Tables 1 and 2). Below, we first consider the theoretical implications of the moderator results of font size comparison and belief instruction, which produced instructive findings about the contribution of fluency to the JOL and memory effects of font size and the contribution of belief to the JOL effects. Then, we outline some key boundary conditions for those effects, along with their theoretical implications, after which we note some broader implications and recommendations for future research.

The JOL effects of font size are not constrained by fluency, but the memory effects are

To recap, the font sizes that were administered in individual experiments were grouped into very small, intermediate, and very large using an external criterion: the fluent range of print size (Legge & Bigelow, 2011; Undorf & Zimdahl, 2019). Accordingly, font size comparisons across the studies were classified into three levels: intermediate versus very small, within the intermediate range, and very large versus intermediate (see Fig. 1). Remember that very small font sizes are below the fluent range, intermediate font sizes are within the fluent range, and very large font sizes are above the fluent range. Therefore, in the three font size comparisons, the larger font was processed with *more*, *comparable*, and *less* fluency than the smaller font, respectively.

If the effects of font size on JOLs are primarily driven by fluency, which means that larger-font items receive higher JOLs because their surface forms are more fluently encoded, one prediction is that items printed in very large fonts should not receive higher JOLs than those printed in intermediate fonts, because they are processed *less* fluently than the latter. However, this prediction was not supported by the moderator results. When very large fonts were compared to intermediate fonts, the JOL effect of font size was not only robust but *larger* than the other two font size comparisons. This result was conceptually consistent with Undorf and Zimdahl's (2019) findings that although the fluency hypothesis predicts an inverted-U relation between JOLs and the font size categories (a low ➔ high ➔ low JOL trend for very small, intermediate and very large fonts), the actual relation was monotonic-increasing. Moreover, in our additional analyses, we demonstrated that the effects of font size comparison can be completely mediated by the absolute larger-smaller font size difference. The JOL effect sizes of font size increased monotonically as the absolute larger-smaller font size difference increased, even when the larger font size was less fluent than the smaller one. Thus, there results suggest that higher fluency is not necessary for larger fonts to receive higher JOLs.

When it comes to the font size effect on memory, we saw that the memory effect was largest for the comparison between very large and intermediate fonts, followed by the comparison within the intermediate range, which aligned with the prior JOL results. We saw that the effect sizes were positive and significant in these two font size comparisons, suggesting that larger fonts improved memory performance. This again supports the view that the font size effect is not a pure illusion, at least when font size comparison involves intermediate to very large font sizes.

However, when it comes to the intermediate versus very small comparison, there was no longer a memory advantage for larger fonts, and hence, there was a JOL-memory dissociation. We should note that in the prior two font size comparisons, larger

fonts were processed with *comparable* or *less* fluency than smaller fonts, but in the current one, larger fonts (intermediate) were processed *more* fluently than smaller fonts (very small). Thus, it seems that although JOLs were always higher for larger fonts than smaller fonts, regardless of which ones involve more fluent encoding, the memory benefits of larger fonts are dependent on (dis)fluency in that they only occur when larger fonts are processed equally or less fluently than smaller fonts. This finding is in general agreement with the findings from perceptual disfluency research, which demonstrate that reducing the processing fluency of study materials (e.g., via blurring, hard-to-read fonts, backward masking) can sometimes enhance encoding and subsequent retrieval (Geller et al., 2018; Magreehan et al., 2016; Mulligan, 2000; Rosner et al., 2015). A possible explanation is that disfluency serves as a signal that the current materials have not been mastered yet, and hence, participants are more likely to engage in deeper and more effortful processing (Diemand-Yauman et al., 2011).

In our additional analyses, we showed that the font size effects on memory were positive with large and small absolute font size differences, but it was negative with medium absolute font size differences. Notably, font size pairs with medium absolute difference were dominated by very small versus intermediate font size comparisons, where the larger fonts were processed *more* fluently. Meanwhile, we also showed that the effects of font size comparison cannot be fully accounted for by the absolute font size difference, suggesting that memory effect sizes, different from JOL effect sizes, did not increase monotonically with the absolute font size difference. Here, it seems that the key boundary condition is that when larger fonts are more fluently encoded than smaller font (i.e., intermediate versus very small), larger fonts no longer produced memory advantages.

In summary, although JOLs and memory accuracy both increased with font size when font size ranged from intermediate to very large, only JOLs did so when size ranged from very small to intermediate. Thus, JOLs and memory were only dissociated when very small and intermediate font sizes were compared, where the larger fonts were more fluently encoded than the smaller fonts. This suggests that fluency plays a key role in JOL-memory dissociation. That is, while the JOL effect of font size is not constrained by fluency, the memory effect seems to be closely tied to (dis)fluency.

The JOL effects of font size are constrained by belief

Whereas the results for the font size comparison moderator ran counter to the fluency hypothesis for the JOL effects of font size, the results for the belief instruction moderator were consistent with the belief hypothesis. Here, we found that when participants were told that larger items are more memorable/more fluently processed or when they received no instructions, larger fonts provoked higher JOLs, and the effect sizes did not differ between conditions. However, when participants were told that larger font size impairs memory or reduces fluency, the font size effect on JOLs vanished. Therefore, fluency was not a constraining factor in the effects of font size on JOLs, but belief was. Together, our results support the belief hypothesis more than the fluency hypothesis.

However, this does not necessarily mean that the font size effect is solely driven by belief. First, it will be recalled that some belief instructions were fluency-based (e.g., that larger fonts are more fluently processed). Thus, the data do not rule out the possibility that fluency may moderate how font size affects memorability indirectly via

beliefs (see also Chen et al., 2019; Price & Harrison, 2017; Wang, Yang, et al., 2020b). Second, it has been argued that performance under belief instructions may not be representative of normal encoding, in which participants follow their natural beliefs. In particular, participants' performance could be subject to desirability bias if they perceive the belief instructions to be experimenter expectations (Yang et al., 2021).

Boundary conditions for font size effects on JOLs

Like Luna et al. (2018), we found that compared to immediate JOLs, delayed JOLs were much less sensitive to font size. The JOL effect of font size was large and significant in immediate JOLs but small and insignificant in delayed JOLs. This finding is reminiscent of the discovery that delayed JOLs track actual memory performance more closely than immediate JOLs do (Rhodes & Tauber, 2011). There are multiple possible reasons why the font size effect shrank and became unreliable in delayed JOLs. For instance, verbatim memories of items' surface forms become rapidly inaccessible with delay, and font size is a prototypical surface detail. Thus, delayed JOLs may therefore be more reliant on other cues that are crucial for memory performance, such as semantic content (Brainerd et al., 2009). Alternatively, this could also be a belief effect, in which participants *assume* that surface features are less important to memory as time passes, owing to time-dependent reductions in their ability to retrieve such details (Luna et al., 2018).

Meanwhile, we found that the font size effect was only reliable with unrelated word lists and unrelated word pairs. Moreover, the effect size was much larger and more robust with unrelated word lists compared to unrelated word pairs. This is a novel finding that has not been previously reported in the font size literature: It suggests that stimulus complexity may restrict the font size effect. A possible explanation for this is that more cues and more strategies become available as the stimuli become more complex, which dilutes the effects of any single cue (font size in this case). Taking word pairs as an illustration, relative to a list of unrelated words, participants may base their JOLs on additional cues such as semantic, phonological, orthographical, or self-generated connections between cue and target words, and encoding strategies such as interactive imagery (Wilton, 2006) and verbal elaboration (Jensen & Rohwer, 1963) can be activated to generate those connections. Another possible explanation lies in the higher cognitive loads imposed by more complex stimuli. Obviously, encoding two words or a sentence consumes more cognitive resources than encoding a single word. In that connection, Luna, Albuquerque, and Martín-Luengo (2019a) reported that the font size effect on JOLs was reduced when the length of sentences increased, and it was erased when sentence length was manipulated within subjects. Luna, Albuquerque, et al. reasoned that high cognitive load may limit participants' abilities to process multiple cues concurrently or constrain their access to beliefs about the diagnosticity of cues.

In addition, the moderator effect of experimental design showed that JOLs were only affected by font size when they were manipulated within subjects. This result is broadly consistent with the cue-utilization account (Koriat, 1997), which specifies that JOLs depend on a comparative process in which target items are perceived to differ in cues that presumably affect their relative memorability. Obviously, as Susser et al. (2013) discussed, differences in font size are salient when font size is manipulated within subjects, but not when it is manipulated between subjects. The literature contains other examples of surface features for which between- versus within-subject manipulation is a boundary condition for JOL effects. For instance, Magreehan et al. (2016) found that

perceptual fluency (manipulated by size and background contrast) only affected JOLs reliably in within-subject designs.

Another novel outcome of the moderator analysis is that study time influences JOLs' sensitivity to font size differences. In an earlier meta-analysis of fewer studies, Halamish (2018) found that the effect sizes did not differ between .5 s and 5 s of study time. However, our results suggest that effect sizes are significantly larger with relatively short study times (.5 s and 2 s) than with relatively long ones (3 s, 4 s, 5 s, 8 s) or with self-paced study time. This pattern is consistent with the results that were recently reported by Zhao et al.'s (2020; Experiment 2), who found that JOLs' sensitivity to font size decreased as study time increased from 2 s or 4 s to 8 s. Zhao et al. argued that longer study times allow participants to focus on other cues that are more diagnostic of memory accuracy than font size is. Nevertheless, it is worth mentioning that another possibility is that participants may not make full use of additional study time but instead become inattentive or begin mind-wandering, which could harm short-term memory for font sizes and reduce the effects of font size on JOLs. Meanwhile, caution needs to be made that Zhao et al.'s encoding procedure required participants to pronounce each word aloud as it was presented, a method that was not used in most font size experiments. Therefore, it waits to be replicated whether their study-time effect would hold for more traditional presentation methods.

It is worth mentioning that although the moderator effect of experiemental context was significant in the univariate meta-regression, we later found that this effect vanished when controlling for the other moderators in the multivariate meta-regression. The implication is that it is not the online format itself that altered the JOL effects, but rather, it was the specific characteristics of online studies (e.g., font size comparison, stimuli type etc.) that lowered the JOL effects.

To sum up, in addition to the aforementioned font size comparison and belief instruction, we have identified several other boundary conditions for the effect of font size on JOLs, including JOL type, stimulus type, and experimental design. Larger font sizes only reliably elevated JOLs when immediate JOLs rather than delayed JOLs were solicited, when unrelated word lists or unrelated word pairs were encoded, and when font size was manipulated within subjects. In addition, the magnitude of the font size effect on JOLs decreased with study time: It was stronger when study time was relatively short than when study time was relatively long or self-paced. Last, there was no evidence the additional noise introduced in online studies modified the JOL effects of font size.

Boundary conditions for font size effects on memory

In the moderator analysis for JOL type, we observed an interesting reversal in which memory was better for larger fonts than smaller fonts when JOL tasks were administered immediately after study, but the pattern was reversed when no JOL tasks were administered. Thus, it seems that whether or not JOLs are solicited is a key determinant of whether items encoded in larger or smaller fonts are easier to remember. These patterns agree with recent research on *JOL reactivity*, which refers to the tendency of JOLs to produce changes in subsequent memory performance relative to no-JOL control conditions (for reviews, see Double et al., 2018; Double & Birney, 2019). It is still unclear why JOL elicitation reverses the memory effects of font size, but there are some possibilities that are worth considering. First, it is possible that item-by-item JOLs cause participants to engage in additional metacognitive processing about the memory-enhancing qualities of larger fonts, which were

not spontaneously processed in the absence of JOLs. Second, JOLs and very small fonts may both lead participants to engage in deeper processing for individual items, and hence, they may have overlapping effects on memory. Accordingly, the potential desirability of the very small font could be masked when JOLs are solicited.

Similar to JOLs, the memory effect of font size was influenced by stimulus type: Larger words were only remembered better when participants encoded lists of unrelated words. This was analogous to JOLs, inasmuch as the effect was larger for word lists than for word pairs or sentences. As we discussed, there are multiple possible explanations, such as the richness of available cues and strategies and the amounts of cognitive loads. Considering the similarity between the JOL data and the memory data, those same mechanisms may also account for the way that stimulus type moderates memory performance. Interestingly, in contrast with Luna et al. (2018), we found that the moderator effect of test format was not even marginally significant. Considering the dependence of test format on stimulus type, the prior evidence for the moderator effect of test format is very likely due to stimulus type.

The moderator effect of study time was another point of similarity between the JOL data and the memory data. Larger words were only remembered better when study times were relatively brief (2 s and 3 s). A plausible explanation is that processing is more heavily dependent on salient surface features, such as font size, when study times are insufficient for deeper forms of processing. According to that account, processing shifts away from surface features toward semantic content as study time lengthens. An alternative explanation, however, is that participants just became less attentive when study time increased, which undermined the effects of font size on memory.

In summary, the boundary conditions for font size effects on memory mostly overlapped with those on JOLs: Larger fonts only improved memory when immediate JOLs were solicited, when unrelated items were encoded, and when study time was relatively brief. It will be recalled from the prior section that a fourth boundary condition for the memory effect of font size is font size comparison: Larger fonts only improved memory when very large and intermediate fonts were compared, and when two font sizes within the intermediate range were compared.

Implications and recommendations for future research

In closing, we highlight some crucial theoretical and empirical questions that need to be resolved in future font size studies and metacognition research in general. The first is concerned with the method of measuring processing fluency. We saw that there has been considerable variability in the measures that have been used to date, which have ranged from lexical decision time to self-paced study time to continuous identification time. Some authors have pointed out in this connection that the current fluency measures may be sensitive to different types of fluency, when only perceptual fluency is of interest for the font size effect. For example, Yang et al. (2018, 2021) demonstrated that continuous identification is more sensitive to the perceptual components of fluency than lexical decision. Thus, further research is required to establish the validity of these methods for measuring perceptual fluency. Further, due to the contradictory findings produced by these different fluency measures (e.g., Mueller et al., 2014; Yang et al., 2018), as well as the scarcity of fluency data, we implemented an indirect test of the fluency hypothesis by linking font size comparison to the fluent range of print size. Still, we emphasize that it is crucial to secure a

valid and consensual fluency measure that allows for direct tests of the contribution of fluency to JOLs (Undorf, 2019; Yang et al., 2021).

A second question that is in need of resolution is concerned with stimulus type moderators. Regarding stimulus type, it will be remembered that prior font size experiments have used unrelated word lists, related and unrelated word pairs, and related and unrelated sentences as stimuli, and both the JOL and memory effects of font size varied significantly across the different stimuli types. One possible explanation is that the available cues/strategies increase with stimulus complexity, which may limit people's abilities to process multiple cues simultaneously. This has important implications for the cue integration approach (e.g., Peynircioğlu & Tatz, 2019; Undorf et al., 2018), which posits that multiple cues can be integrated in making metacognitive judgments—namely, certain characteristics of increasing stimulus complexity, such as richer relational information, may divert people from incorporating surface cues such as font size into metacognitive judgments. Here, there is a missing type of stimulus that would provide incisive information—namely, related word lists. Studying the font size effect with related word lists would be very instructive, as the classic principle of item versus relational processing (e.g., Hunt, 2003) suggests that the influence of the properties of individual items (e.g., size, position, type face) trade off with the influence of relations among items (e.g., semantic concepts, emotional qualities). That is, the influence of item properties wane as relational properties become more numerous, and conversely. Hence, it is possible that font size effects would be weaker for lists of related words than for lists of unrelated words.

Third, we found that the memory effect of font size changed dramatically as a consequence of whether or not JOLs were solicited, which is consistent with the literature on JOL reactivity (for reviews, see Double et al., 2018; Double & Birney, 2019). The implication is that surface characteristics such as font size may not have unconditional effects on memory, but rather, their effects may depend on using such information to make JOLs. However, that hypothesis requires focused experimentation with designs that include no-JOL control groups. To date, such designs have not been common practice in JOL research (Double et al., 2018). Thus, future research that focuses on memory or metamemory effects of perceptual features, such as perceptual disfluency, should consider including a no-JOL control group and compare performance on memory tests that are preceded versus not preceded by JOLs. It is also important to explore alternative metamemory measures that are less reactive, such as think-aloud verbal reports (Ericsson & Simon, 1980).

Finally, we observed that the JOL effects of font size were larger than the corresponding memory effects, suggesting that the large font size of learning materials can produce overconfidence. Such findings have obvious educational implications, considering that key information is often presented in larger font in textbooks or lecture slides. Although this may indeed create memory benefits, it also leads to overconfidence in mastery of the materials. Consequently, students' allocation of study time may be misguided by such overconfidence. Therefore, the underlying mechanism of such overconfidence and methods to reduce it merit further investigation. Because only a small portion of the studies included in our meta-analyses reported indexes of metamemory accuracy (i.e., the correspondence between JOLs and memory), such as resolution and calibration, we were not able to conduct any meaningful moderator analyses on that. Thus, we encourage researchers to include such analyses in future studies and isolate the factors that moderate the font size effect on metamemory accuracy.

Conclusions

We found that JOLs and memory accuracy were both higher overall for larger-font items than for smaller-font ones, suggesting that the font size effect was not completely a metacognitive illusion. This demonstrates that participants' JOLs were to some extent correct in predicting memory benefits for larger fonts. However, the effect sizes for font size were much smaller for memory than for JOLs, suggesting that JOLs are much more sensitive to changes in font size than memory is. At a more fine-grained level, we found that the JOL effect and memory effect were in broad alignment when the font size comparison involved very large versus intermediate fonts or two fonts from the intermediate range, but the two effects were dissociated when the font size comparison involved very small versus intermediate fonts. This is most likely because the memory effect was influenced by (dis)fluency whereas the JOL effect was not. Last, we found that both the JOL and memory effects of font size were moderated by font size comparison, JOL type, stimulus type, and study time, which provide theoretical and empirical implications for future research on the font size effect and on metamemory in general.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11409-021-09287-3.

Acknowledgments Thanks to Glen Bodner, Kit Double, Vered Halamish, Xiao Hu, Nate Kornell, Nikolaos Magoulas, Neil Mulligan, Matthew Rhodes, Ningxin Su, Joshua Tatz, Monika Undorf, and Chunliang Yang for sharing their data.

Code availability Available upon request to the first author.

Author's contribution Conceptualization: Minyu Chang; Methodology: Minyu Chang; Formal analysis and investigation: Minyu Chang; Writing - original draft preparation: Minyu Chang; Writing - review and editing: Minyu Chang, Charles. J. Brainerd; Supervision: Charles. J. Brainerd.

Data availability All relevant data are within the paper and the Supplementary Materials.

Declarations

Conflict of interest Not applicable.

References

References preceded by a superscript of J provided data for the meta-analysis for JOLs, and references preceded by a superscript of M provided data for the meta-analysis for memory performance.

[JM]Ball, B. H., Klein, K. N., & Brewer, G. A. (2014). Processing fluency mediates the influence of perceptual information on monitoring learning of educationally relevant materials. *Journal of Experimental Psychology: Applied, 20*(4), 336–348. https://doi.org/10.1037/xap0000023.

Berlin, J. A., & Antman, E. M. (1992). Advantages and limitations of meta-analytic regressions of clinical trials data. *Controlled Clinical Trials, 13*(5), 422. https://doi.org/10.1016/0197-2456(92)90151-O

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

Black, N., Mullan, B., & Sharpe, L. (2016). Computer-delivered interventions for reducing alcohol consumption: Meta-analysis and meta-regression using behaviour change techniques and theory. *Health Psychology Review, 10*(3), 341–357. https://doi.org/10.1080/17437199.2016.1168268

[JM]Blake, A. B. (2018). *Factors that influence metacognitive judgments: Effects at encoding, in the presence of diagnostic cues, and after incidental encoding* [Doctoral dissertation, The University of California, Los Angeles]. https://escholarship.org/uc/item/7d71z5kj. Accessed 7 July 2020

[JM]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.

[M]Bodner, G. E., Huff, M. J., & Taikh, A. (2020). Pure-list production improves item recognition and sometimes also improves source memory. *Memory & Cognition, 48*(7), 1281–1294. https://doi.org/10.3758/s13421-020-01044-2.

Brainerd, C. J., Reyna, V. F., & Howe, M. L. (2009). Trichotomous processes in early memory development, aging, and neurocognitive impairment: A unified theory. *Psychological Review, 116*(4), 783–832. https://doi.org/10.1037/a0016963

[JM]Bröder, A., & Undorf, M. (2019). Metamemory viewed through the judgment lens. *Acta Psychologica, 197*, 153–165. https://doi.org/10.1016/j.actpsy.2019.04.011.

[JM]Chen, Y., Li, F., & Li, W. (2019). The influence of learner's beliefs about processing fluency on font-size effect. *Acta Psychologica Sinica, 15*(2), 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*(6), 590–606. https://doi.org/10.3758/BF03213348

Cook, D. J., Guyatt, G. H., Ryan, G., Clifton, J., Buckingham, L., Willan, A., McIlroy, W., & Oxman, A. D. (1993). Should unpublished data be included in meta-analyses? Current convictions and controversies. *Journal of the American Medical Association, 269*(21), 2749–2753.

Cumming, G. (2013). *Understanding the new satistics: Effect sizes, confidence intervals, and meta-analysis* (1st ed.). https://doi.org/10.4324/9780203807002

[JM]Dalot, R. G. (2018). *Effect of font-size, real-size, and mental imagery on JOLs and memory* [Master's Thesis, University of Minho]. http://repositorium.sdum.uminho.pt/. Accessed 2 Feb 2021

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the (): Effects of disfluency on educational outcomes. *Cognition, 118*(1), 111–115. https://doi.org/10.1016/j.cognition.2010.09.012

[JM]Double, K. S. (2019). Do judgments of learning impair recall when uninformative cues are salient? *PsyArXiv*. https://doi.org/10.31234/osf.io/a5bxw.

Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. *Frontiers in Psychology, 10*, Article 2755. https://doi.org/10.3389/fpsyg.2019.02755

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory, 26*(6), 741–750. https://doi.org/10.1080/09658211.2017.1404111

Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 54, pp. 103–140). Academic Press. https://doi.org/10.1016/B978-0-12-385527-5.00004-8

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629–634.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association, 87*(417), 178–183. https://doi.org/10.2307/2290467

Geller, J. (2017). *Would disfluency by any other name still be disfluent? Examining the boundary conditions of the disfluency effect* [doctoral dissertation, Iowa State University]. https://lib.dr.iastate.edu/etd/15520/. Accessed 10 Sept 2020

Geller, J., Still, M. L., Dark, V. J., & Carpenter, S. K. (2018). Would disfluency by any other name still be disfluent? Examining the disfluency effect with cursive handwriting. *Memory & Cognition, 46*(7), 1109–1126. https://doi.org/10.3758/s13421-018-0824-6

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, part I: The Cohen's d family. *The Quantitative Methods for Psychology, 14*(4), 242–265. https://doi.org/10.20982/tqmp.14.4.p242

[JM]Halamish, V. (2018). Can very small font size enhance memory? *Memory & Cognition, 46*(6), 979–993. https://doi.org/10.3758/s13421-018-0816-6.

[JM]Halamish, V., Nachman, H., & Katzir, T. (2018). The effect of font size on Children's memory and Metamemory. *Frontiers in Psychology, 9*, 1577. https://doi.org/10.3389/fpsyg.2018.01577.

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). Doing meta-analysis in R: A hands-on guide. PROTECT Lab Erlangen. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/. Accessed 28 Feb 2021

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related rstimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Higgins, J. P. T., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine, 23*(11), 1663–1682. https://doi.org/10.1002/sim.1752

Hirshman, E., & Mulligan, N. (1991). Perceptual interference improves explicit memory but does not enhance data-driven processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(3), 507–513. https://doi.org/10.1037/0278-7393.17.3.507

[JM]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*(11), e0142351. https://doi.org/10.1371/journal.pone.0142351.

[JM]Hu, X., Liu, Z., Li, T., & Luo, L. (2016). Influence of cue word perceptual information on metamemory accuracy in judgement of learning. *Memory, 24*(3), 383–398. https://doi.org/10.1080/09658211.2015.1009470.

Hunt, R. R. (2003). Two contributions of distinctive processing to accurate memory. *Journal of Memory and Language, 48*(4), 811–825. https://doi.org/10.1016/S0749-596X(03)00018-4

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513–541. https://doi.org/10.1016/0749-596X(91)90025-F

Jensen, A. R., & Rohwer, W. D. (1963). Verbal mediation in paired-associate and serial learning. *Journal of Verbal Learning and Verbal Behavior, 1*(5), 346–352. https://doi.org/10.1016/S0022-5371(63)80015-8

[JM]Kelly, E. B. (2019). *Belief mediates the font-size effect on judgements of learning* [Bachelor's thesis, University of North Carolina at Chapel Hill]. https://cdr.lib.unc.edu/concern/honors_theses/xw42nd030. Accessed 2 Feb 2021

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A. (2015). Metacognition: Decision making processes in self-monitoring and self-regulation. In K. Gideon & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 356–379). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118468333.ch12

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*(4), 643–656. https://doi.org/10.1037/0096-3445.133.4.643

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—And costs—Of dropping flashcards. *Memory, 16*(2), 125–136. https://doi.org/10.1080/09658210701763899

[JM]Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science, 22*(6), 787–794. https://doi.org/10.1177/0956797611407929.

Legge, G. E., & Bigelow, C. A. (2011). Does print size matter for reading? A review of findings from vision science and typography. *Journal of Vision, 11*(5), 8–8. https://doi.org/10.1167/11.5.8

[M]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*(4), 1007–1013. https://doi.org/10.3758/s13423-014-0762-7.

[JM]Luna, K., Martín-Luengo, B., & Albuquerque, P. B. (2018). Do delayed judgements of learning reduce metamemory illusions? A meta-analysis. *Quarterly Journal of Experimental Psychology, 71*(7), 1626–1636. https://doi.org/10.1080/17470218.2017.1343362.

[JM]Luna, K., Albuquerque, P. B., & Martín-Luengo, B. (2019a). Cognitive load eliminates the effect of perceptual information on judgments of learning with sentences. *Memory & Cognition, 47*(1), 106–116. https://doi.org/10.3758/s13421-018-0853-1.

[JM]Luna, K., Nogueira, M., & Albuquerque, P. B. (2019b). Words in larger font are perceived as more important: Explaining the belief that font size affects memory. *Memory, 27*(4), 555–560. https://doi.org/10.1080/09658211.2018.1529797.

[JM]Magoulas, N. (2018). *Perceptual fluency, metamnemonic estimates and actual memory performance: An interlinguistic approach* [Bachelor's thesis, the American College of Greece]. https://doi.org/10.13140/RG.2.2.29052.18568.

Magreehan, D. A., Serra, M. J., Schwartz, N. H., & Narciss, S. (2016). Further boundary conditions for the effects of perceptual disfluency on judgments of learning. *Metacognition and Learning, 11*(1), 35–56. https://doi.org/10.1007/s11409-015-9147-1

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*(3), 252–271. https://doi.org/10.1037/0033-295X.87.3.252

[J]Mcdonough, I., & Gallo, D. (2012). Illusory expectations can affect retrieval-monitoring accuracy. Journal of Experimental Psychology. *Learning, Memory, and Cognition, 38*(2), 391–404. https://doi.org/10.1037/a0025548.

[JM]Miele, D. B., Finn, B., & Molden, D. C. (2011). Does easily learned mean easily remembered?: It depends on your beliefs about intelligence. *Psychological Science, 22*(3), 320–324. https://doi.org/10.1177/0956797610397954.

[JM]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.

[JM]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? *Quarterly Journal of Experimental Psychology, 69*(4), 781–799. https://doi.org/10.1080/17470218.2015.1058404.

Mulligan, N. W. (2000). Perceptual interference at encoding enhances item-specific encoding and disrupts relational encoding: Evidence from multiple recall tests. *Memory & Cognition, 28*(4), 539–546. https://doi.org/10.3758/BF03201244

[JM]Park, K. M. (2015). *Boundary conditions of font size effects* [Master's Thesis, The University of Alabama in Huntsville]. https://search.proquest.com/openview/743d096003d21109e12ba7269617571e/1?pq-origsite=gscholar&cbl=18750&diss=y. Accessed 2 Feb 2021

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology, 61*(10), 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010

[JM]Peynircioğlu, Z. F., & Tatz, J. R. (2019). Intensifying the intensity illusion in judgments of learning: Modality and cue combinations. *Memory & Cognition, 47*(3), 412–419. https://doi.org/10.3758/s13421-018-0875-8.

[JM]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.

[JM]Price, J., McElroy, K., & Martin, N. J. (2016). The role of font size and font style in younger and older adults' predicted and actual recall performance. *Aging, Neuropsychology, and Cognition, 23*(3), 366–388. https://doi.org/10.1080/13825585.2015.1102194.

[JM]Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*(4), 615–625. https://doi.org/10.1037/a0013684.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*(1), 131–148. https://doi.org/10.1037/a0021705

Rosner, T. M., Davis, H., & Milliken, B. (2015). Perceptual blurring and recognition memory: A desirable difficulty effect revealed. *Acta Psychologica, 160*, 11–22. https://doi.org/10.1016/j.actpsy.2015.06.006

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2006). *Publication bias in meta-analysis*. John Wiley & Sons.

Schmucker, C. M., Blümle, A., Schell, L. K., Schwarzer, G., Oeller, P., Cabrera, L., von Elm, E., Briel, M., & Meerpohl, J. J. (2017). Systematic review finds that study data not published in full text articles have unclear impact on meta-analyses results in medical research. *PLoS One, 12*(4), e0176210. https://doi.org/10.1371/journal.pone.0176210

[JM]Smith, W. G. (2019). *When to Move on to New Learning: Meta-cognition's Role on Updating with Incorporation* [Master's Thesis, University of North Carolina at Greensboro]. https://libres.uncg.edu/ir/uncg/f/Smith_uncg_0154M_12411.pdf. Accessed 4 Feb 2021

[JM]Soderstrom, N. C. (2012). *An investigation of the basis of judgments of remembering and knowing (JORKS)* [Doctoral dissertation, Colorado State University]. https://mountainscholar.org/handle/10217/67949. Accessed 16 Feb 2021

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal, 343*, d4002. https://doi.org/10.1136/bmj.d4002

[JM]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*(7), e0200888. https://doi.org/10.1371/journal.pone.0200888.

[JM]Susser, J. A., Mulligan, N. W., & Besken, M. (2013). The effects of list composition and perceptual fluency on judgments of learning (JOLs). *Memory & Cognition, 41*(7), 1000–1011. https://doi.org/10.3758/s13421-013-0323-8.

[M]Taikh, A., & Bodner, G. (2016). Evaluating the basis of the between-group production effect in recognition. *Canadian Journal of Experimental Psychology, 70*(2), 186–194. https://doi.org/10.1037/cep0000083.

[JM]Tatz, J. R., & Peynircioğlu, Z. F. (2020). Judgments of learning in context: Backgrounds can both reduce and produce metamemory illusions. *Memory & Cognition, 48*(4), 581–595. https://doi.org/10.3758/s13421-019-00991-9.

[JM]Tatz, J. R., Undorf, M., & Peynircioglu, Z. (2020). Effect of impoverished information on multisensory integration in judgments of learning. *Journal of Experimental Psychology Learning Memory and Cognition, 47(3), 481–497.*. https://doi.org/10.1037/xlm0000953.

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*(11), 1559–1573. https://doi.org/10.1002/sim.1187

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods, 10*(2), 161–179. https://doi.org/10.1002/jrsm.1338

[J]Undorf, M. (2019). Fluency illusions in metamemory. In: *Memory quirks: The study of odd phenomena in memory* (150-174). Routledge. https://www.madoc.bib.uni-mannheim.de/52584/. Accessed 28 Sept 2020

[JM]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*(1), 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

Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition, 46*(4), 507–519. https://doi.org/10.3758/s13421-017-0780-6.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. https://doi.org/10.1002/jrsm.11

[JM]Wang, J., Qu, K., & Zhang, H. (2020a). The effect of processing fluency on the font size effect of judgments of learning. *Journal of Psychological Science, 43*(1), 17–22.

[JM]Wang, Z., Yang, C., Zhao, W., & Jiang, Y. (2020b). Perceptual fluency affects judgments of learning nonanalytically and analytically through beliefs about how perceptual fluency affects memory. *Frontiers in Psychology, 11*, 552824. https://doi.org/10.3389/fpsyg.2020.552824.

Wilton, R. N. (2006). Interactive imagery and colour in paired-associate learning. *Acta Psychologica, 121*(1), 21–40. https://doi.org/10.1016/j.actpsy.2005.05.006

Witherby, A. E., & Tauber, S. K. (2017). The concreteness effect on judgments of learning: Evaluating the contributions of fluency and beliefs. *Memory & Cognition, 45*(4), 639–650. https://doi.org/10.3758/s13421-016-0681-0

Yang, C., Huang, T. S.-T., & Shanks, D. R. (2018). 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

[JM]Yang, C., Yu, R., Hu, X., Luo, L., Huang, T. S.-T., & Shanks, D. R. (2021). How to assess the contributions of processing fluency and beliefs to the formation of judgments of learning: Methods and pitfalls. *Metacognition and Learning*. https://doi.org/10.1007/s11409-020-09254-4.

Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—And is not—A desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition, 41*(2), 229–241. https://doi.org/10.3758/s13421-012-0255-8

[JM]Zhao, W., Jiang, Y., Wang, Z., & Jingyuan, H. (2020). Influence of encoding strength on the font size effect. *Acta Psychologica Sinica, 52*(10), 1156–1167. https://doi.org/10.3724/SP.J.1041.2020.01156.