**Experimental Psychology**

Do beliefs about font size affect retrospective metamemory judgments in addition to prospective judgments? A registered report

---Manuscript Draft---

<table>
<thead>
<tr>
<th>Manuscript Number:</th>
<th>MS-1796R4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Title:</td>
<td>Do beliefs about font size affect retrospective metamemory judgments in addition to prospective judgments? A registered report</td>
</tr>
<tr>
<td>Short Title:</td>
<td>Metamemory beliefs and confidence</td>
</tr>
<tr>
<td>Article Type:</td>
<td>Registered Report</td>
</tr>
<tr>
<td>Keywords:</td>
<td>metamemory; beliefs; confidence; font size</td>
</tr>
<tr>
<td>Abstract:</td>
<td>Beliefs about how memory works explain several effects on prospective metamemory judgments (e.g., the effect of font size on JOLs). Less is known about the effect of beliefs on retrospective judgments (i.e., confidence). Here, we tested whether font size also affects confidence ratings and whether beliefs play a similar role in confidence than in JOLs. In two experiments, participants studied words in small and large size, rated JOLs, and completed a font-size test in which they indicated the font size at study and a standard old/new recognition test. The results confirmed that font size affected both JOLs and confidence ratings. The presentation of the counter-belief that memory is better for words in small font size in Experiment 2 and the analyses of confidence for participants who did not believe that large fonts improved memory suggested that the effect of font size on confidence was based on beliefs. This research shows that the debate on theory-based and experience-based factors should not be limited to prospective metamemory judgments, but also encompass retrospective judgments.</td>
</tr>
<tr>
<td>Manuscript Classifications:</td>
<td>16: Memory</td>
</tr>
<tr>
<td>Funding Information:</td>
<td>Portuguese Foundation for Science and Technology and the Portuguese Ministry of Science, Technology and Higher Education through national funds and co-financed by FEDER through COMPETE2020 under the PT2020 Partnership Agreement (POCI-01-0145-FEDER-007653)</td>
</tr>
<tr>
<td>Corresponding Author:</td>
<td>Karlos Luna, Ph.D.</td>
</tr>
<tr>
<td>Universidad Nacional de Colombia</td>
<td></td>
</tr>
<tr>
<td>Bogotá, Bogotá COLOMBIA</td>
<td></td>
</tr>
<tr>
<td>Corresponding Author Secondary Information:</td>
<td></td>
</tr>
<tr>
<td>First Author:</td>
<td>Karlos Luna, Ph.D.</td>
</tr>
<tr>
<td>First Author Secondary Information:</td>
<td></td>
</tr>
<tr>
<td>Other Authors:</td>
<td>Pedro B Albuquerque, PhD</td>
</tr>
<tr>
<td>Author Comments:</td>
<td>This is the Stage 2 report of a previously accepted Registered Report-Stage 1.</td>
</tr>
<tr>
<td>Additional Information:</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has the manuscript or any component of it already been published or is currently under consideration by another journal?</td>
<td>No</td>
</tr>
<tr>
<td>If the paper has been written by more than one person, can the corresponding author attest that each author endorsed the manuscript in the form submitted,</td>
<td>Yes</td>
</tr>
<tr>
<td>Question</td>
<td>Answer</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>agreed to be listed as a coauthor, and has accepted the order of authorship?</td>
<td>Yes</td>
</tr>
<tr>
<td>Is there any conflict of interests concerning the content of this manuscript?</td>
<td>No</td>
</tr>
<tr>
<td>The manuscript comprehensively reports study design and data analysis, including all manipulations and all measures used; how sample size was determined; and which data, if any, were excluded from the analyses.</td>
<td>Yes</td>
</tr>
<tr>
<td>Does the length of the manuscript fall within the journal's word limits? [Preregistration of a single experiment: 5,000 words; Two and more experiments: 10,000 words – including title page, abstract, text, references, notes, appendices, tables, and figures]</td>
<td>Yes</td>
</tr>
<tr>
<td>Please state the word count. as follow-up to “Does the length of the manuscript fall within the journal's word limits? [Preregistration of a single experiment: 5,000 words; Two and more experiments: 10,000 words – including title page, abstract, text, references, notes, appendices, tables, and figures]”</td>
<td>A bit over 10000 words to accommodate Reviewers comments</td>
</tr>
<tr>
<td>Has the manuscript been prepared according to the Publication Manual of the American Psychological Association (7th ed.)?</td>
<td>Yes</td>
</tr>
<tr>
<td>Has this manuscript or another version of it been submitted to this journal previously?</td>
<td>No</td>
</tr>
<tr>
<td>Has the study that is being proposed in this manuscript already been conducted?</td>
<td>No</td>
</tr>
<tr>
<td>Before acceptance for publication, the manuscript must contain a permanent URL pointing to the raw data as well as any other required additional information (e.g., variable names, codebook, analysis script, etc.; please use open format, e.g., .csv, .txt format). Raw data must be deposited in a public or institutional repository that provides a permanent digital object identifier (DOI) for the electronic files (for a list of repositories see <a href="http://www.re3data.org">http://www.re3data.org</a>). Please provide a paragraph specifying the DOI at</td>
<td>The manuscript includes a DOI generated by a public data repository pointing to the raw data underlying the findings reported in the article.</td>
</tr>
</tbody>
</table>
the end of the manuscript after the reference section.

<table>
<thead>
<tr>
<th>Please provide the URL here as well. as follow-up to “Before acceptance for publication, the manuscript must contain a permanent URL pointing to the raw data as well as any other required additional information (e.g., variable names, codebook, analysis script, etc.; please use open format, e.g., .csv, .txt format). Raw data must be deposited in a public or institutional repository that provides a permanent digital object identifier (DOI) for the electronic files (for a list of repositories see <a href="http://www.re3data.org">http://www.re3data.org</a>). Please provide a paragraph specifying the DOI at the end of the manuscript after the reference section.”</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://osf.io/z62gk/?view_only=6497c613a6604d97a8aae05e2470cdf8">https://osf.io/z62gk/?view_only=6497c613a6604d97a8aae05e2470cdf8</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>If your original materials (including all stimuli, instructions where relevant, item wordings where relevant, and precise references to materials taken from other sources, such as tests or stimuli databases) are available from a stable online data repository, your manuscript may be eligible for the open materials badge. Please provide a URL pointing to the data repository at the end of the manuscript after the reference section.</th>
</tr>
</thead>
<tbody>
<tr>
<td>To the best of my ability and knowledge, I have provided all original materials and clear references to all other materials via a stable online repository.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Please provide the URL here as well. as follow-up to “If your original materials (including all stimuli, instructions where relevant, item wordings where relevant, and precise references to materials taken from other sources, such as tests or stimuli databases) are available from a stable online data repository, your manuscript may be eligible for the open materials badge. Please provide a URL pointing to the data repository at the end of the manuscript after the reference section.”</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://osf.io/z62gk/?view_only=6497c613a6604d97a8aae05e2470cdf8">https://osf.io/z62gk/?view_only=6497c613a6604d97a8aae05e2470cdf8</a></td>
</tr>
</tbody>
</table>

Future estimations of memory retrieval (JOLs) are affected by beliefs about the factors influencing memory. In this research, we showed that beliefs also affect participants’ confidence that they selected the correct answer. In particular, beliefs about the effect of font size on memory influenced confidence ratings.
understand the significance your work. If your article is accepted for publication, the plain language summary may be used for marketing and social media purposes by Hogrefe Publishing. Feel free to use the summary for your own purposes as well. If you do not wish to provide a summary, please put a placeholder in the box below and continue with your submission.

Example 1: The layperson's view equates aging with cognitive decline. However, research suggests that there are both gains and losses associated with aging. We propose a framework for studying how particular ecologies - niches - can lead to failure, maintenance, or improvements in performance with increased age.

Example 2: People who weigh the pros and cons of a potential goal differ in their thinking from people who plan their course of action to reach a goal. The present article shows that this difference in thinking carries over to how they see and take risks. People who plan when, where, and how to act for a goal are more eager to take risks in an entirely unrelated risk game.

By providing this text, you consent to it being used for social media and marketing purposes. Please see our privacy policy here.

Suggested Reviewers:

Monika Undorf, PhD
Universität Mannheim
undorf@psychologie.uni-mannheim.de
Expert in the effect of theory-based and experience-based factors in metamemory judgments.

John Dunlosky, PhD
Kent State University
jdunlosk@kent.edu
Expert in metamemory.

Opposed Reviewers:
Dear Dr James R Schmidt, Action Editor,

Please find attached a revised version of the Registered Report MS-1796R3 entitled "Do beliefs about font size affect retrospective metamemory judgments in addition to prospective judgments? A registered report”.

Thank you very much for your helpful and constructive reviews of the report. We found the comments very useful and made a number of changes accordingly. Specific answers to the comments are presented in the Response to comments file. The Reviewers’ comments are presented in upper-case and our responses in lower-case to easily discriminate between the two.

We hope that these revisions and responses satisfactorily address yours and the Reviewers’ concerns.

Sincerely,

The Authors
REVIEWERS’ COMMENTS:

REVIEWER #2: I AM EXCITED TO SEE THE COMPLETION OF THIS STUDY AND CONTINUE TO APPLAUD THE AUTHORS FOR CONDUCTING A REGISTERED REPORT ON THIS SUBJECT MATTER! BELOW ARE MY COMMENTS ON THE RESULTS AND DISCUSSION.

RESULTS

1. EXPERIMENT 1 - EXPLORATORY ANALYSIS - IS THIS ANALYSIS MEANINGFUL?
FIRST, THE SAMPLE SIZE DID NOT REACH THE REQUISITE SUGGESTED IN THE POWER ANALYSIS, SO ANY RESULTS MAY NOT BE MEANINGFUL. SECOND, IDEALLY ONE WOULD ONLY TEST THE SUBSAMPLE WITH SMALL > LARGE BELIEFS AND SHOW THE OPPOSITE EFFECTS IN CONFIDENCE. INSTEAD, THE AUTHORS REPORT NULL RESULTS, WHICH IS INCONSISTENT WITH A BELIEF HYPOTHESIS. AGAIN THOUGH, GIVEN THE SMALL SAMPLE SIZE, THE ANALYSIS SIMPLY MIGHT NOT BE INTERPRETABLE. GIVEN THAT THIS ANALYSIS ALSO WAS CONDUCTED IN E2 (AND ALSO WITH LESS POWER THAN REQUIRED, I BELIEVE), I WONDER IF ONLY INCLUDING THE SMALL > LARGE BELIEFS GROUP COLLAPSED ACROSS BOTH SAMPLES WOULD YIELD A) A LARGE ENOUGH SAMPLE SIZE AND B) SIGNIFICANT CONFIDENCE EFFECTS IN THE DIRECTION OF SMALL > LARGE. GIVEN THE STRONG CONCLUSIONS BEING MADE BY THE AUTHORS (P. 28) THAT THESE EXPLORATORY ANALYSES SHOWED "CLEAR RESULTS," MAKING SURE THEY ARE WELL-POWERED AND MAKING SURE THAT THEY ARE IN A CLEAR DIRECTION OF A BELIEF-BASED HYPOTHESIS IS NECESSARY.

We thank the Reviewer for this comment. We tried to be careful in interpreting the results of the exploratory analyses (see, for example, the Discussion of Experiment 1), but we missed the reference to “clear results”. We have removed that clause and are now more cautious in the interpretation of the exploratory results.

In Experiments 1 and 2, we could not test participants with small > large beliefs only because there were very few participants with that belief (3 in Experiment 1 and 10 in total in Experiment 2; collapsing participants in both experiments does not seem helpful either). Instead, we collapsed them with the participants who did not hold any belief. However, as a result of this comment and point 4 below, we replaced the exploratory analyses from Experiments 1 and 2, and now we compare participants believing and not believing that large words would be better remembered. Thus, we used the full sample of participants. In both experiments, for font-size and recognition tests, the results show that font size affected confidence only for participants in the belief group. As most of the participants in the no belief group believed that font size did not affect memory, it is reasonable that there was no font-size effect. We see this result as consistent with the beliefs explanation.

2. EXPERIMENT 2 - QUESTIONS ABOUT BELIEFS: TO SHOW THAT THEIR MANIPULATION ALTERED BELIEFS, IT WOULD BE NICE TO CALCULATE A CHI-SQUARE TEST ON THE PROPORTIONS TO SEE IF THEY SIGNIFICANTLY DIFFER BETWEEN THE TWO GROUPS. OTHERWISE, THE DIFFERENCES COULD BE DUE TO CHANCE, ESPECIALLY GIVEN THAT THE DIFFERENCES IN BELIEFS BETWEEN THOSE TWO GROUPS WERE ACTUALLY SMALLER THAN THE DIFFERENCE BETWEEN THE "LARGER IS BETTER" GROUP AND THE NO-INSTRUCTION GROUP IN EXPERIMENT 1. IN GENERAL, THE GROUP DIFFERENCES IN BELIEFS FROM E1 TO E2 ARE QUITE LARGE AND SURPRISING. DO THE AUTHORS HAVE ANY EXPLANATION? WERE
OVERLAPPING STUDENTS TESTED IN EACH EXPERIENCE AND SO MAY HAVE BECOME "WISER" IN EXPERIMENT 2?

In our lab, we use the SONA system software for data collection. SONA software registers the experiments a given student has completed, and thus we can prevent the same student from participating in related or similar experiments. As a result, we are certain that there were no overlapping participants between the two experiments. We agree with the Reviewer that there is a surprising difference in beliefs between Experiments 1 and 2. In line with a suggestion from Reviewer 4, we think that the instructions on the effect of font size on memory caused a sort of reactivity effect (now mentioned in the Discussion of Experiment 2). That reactivity may explain the differences in beliefs between Experiment 1 and 2 and the small effect of our beliefs manipulation. A chi-squared showed no significant differences for beliefs depending on the instruction, chi2(2) = 1.92, p = 0.38. However, other results suggest that there was, indeed, an effect of the instructions in confidence. The measure of beliefs was collected at the end of the experiment and that, as pointed previously by a Reviewer, may not have been optimal. For example, experience with the items remembered may have influenced the answer to the beliefs questions. Despite that, we preferred to keep the beliefs question at the end of the experiment to avoid interference with the main measures. In sum, the results of the beliefs question do not show an effect of the manipulation, but other results from the experiment suggest that there was, indeed, a small effect.

3. EXPERIMENT 2 - I'M SURPRISED AT THE DIFFERENT CONCLUSIONS FROM THE NHST AND BAYES FACTOR. SOMETIMES THEY CAN DIFFER IF THE DISTRIBUTIONS OF VALUES IS HIGHLY SKewed OR IF OUTLIERS EXIST. DID THE AUTHORS INVESTIGATE WHETHER EITHER COULD BE DRIVING THE SLIGHTLY DIVERGING CONCLUSIONS?

We did not search for reasons for the divergent results beforehand, but we explored our data after the Reviewer’s comment.

First, we are not sure of the usefulness of the outlier concept in metamemory judgments as those collected here. For example, consider a participant in a given condition with confidence M = 5 (in a scale 1-7), SD = 1. If we apply the popular and straightforward criterion of plus minus 2*SD to identify outliers, it would mean that judgments in the range 3-7 would be acceptable and answers 1-2 would be outliers. However, we do not think these responses should be identified as outliers and removed. In our view, responses 1-2 would be perfectly acceptable answers indicating that the participant believes that the answer is very likely incorrect. The logic of outliers applies well to reaction times and other measures, but not so well to the metamemory judgments used here. There is no good reason we can come with to consider an answer 1 (even in a participant who only answered 7 for all the other items) as an outlier that should be removed. Despite this, in our experience the high variability of metamemory measures prevents the existence of outliers. As per distribution, in our experience distributions of confidence ratings depend heavily on the difficulty of the materials. With easy materials, distributions tend to be negatively skewed, and with difficult materials distributions tend to be positively skewed, with higher skewness with more extreme materials (either easy or difficult). Thus, we are not certain of what we can learn from skewness in a confidence rating, but we acknowledge that the effects of skewness in, for example, Bayesian analyses are beyond our knowledge. Our confidence distributions were slightly negatively skewed. For example, in Experiment 2, font-size test, participants answered with confidence 1 8% of times, with confidence 2 13%, 3 16%, 4 15%, 5 16%, 6 14%, and confidence 7 19% of the times. Our interpretation of the divergence in results between NHST and Bayes Factor lies in the high variability of p-values, as shown by Cumming in the p-value dance (e.g., https://www.youtube.com/watch?v=5OL1RqHrZQ8). A p = .10, considering the variability of p-values, may not be totally inconsistent with a BF = 4.
DISCUSSION

4. GIVEN THE RELATIVELY WEAK EFFECTS AND MANY NON-SIGNIFICANT INTERACTIONS IN EXPERIMENT 2, I SUPPOSE I AM NOT ENTIRELY CONVINCED THAT ONE CAN SAY BELIEFS ARE A DRIVER OF FONT SIZE EFFECTS. IT COULD STILL BE THE CASE THAT A THIRD VARIABLE (E.G., FLUENCY) IS AFFECTING CONFIDENCE JUDGMENTS BOTH AT ENCODING AND AT RETRIEVAL. TO EXPLAIN WHY DIFFERING BELIEFS IN THE EXPLORATORY ANALYSIS ALTERED THE EFFECTS, ONE WOULD FIRST NEED TO CONDUCT A TEST TO SHOW A SIGNIFICANT INTERACTION (WHICH WAS NOT DONE) AND ONE WOULD HAVE TO RULE OUT THE POSSIBILITY THAT THE TWO GROUPS WERE SIMPLY NOT DIFFERENTLY PROCESSING THE STIMULI (E.G., DIFFERED ON PERCEPTUAL THRESHOLDS OR INTERPRETATION OF FLUENCY EFFECTS, ETC.). THAT IS THE WEAKNESS OF BETWEEN SUBJECT ANALYSES.

In Experiments 1 and 2, we now report analyses 2 (belief group: yes, no) x 2 (small, large) for the font-size and the recognition tests. In both experiments and tests, the interaction was significant and showed that the effect of font size on confidence was present in the belief group and absent in the no-belief group. We interpret these results as that beliefs are mostly responsible for the observed effect of font size on confidence.

5. THE DISCUSSION ON P. 30-31 ON THE RATIONALE FOR THE FONT SIZE EFFECT FOR SMALLER WORDS SOUNDS VERY SIMILAR TO A DISTINCTIVENESS HEURISTIC ARGUED BY MCDONOUGH AND GALLO (2012) AND THIS MIGHT WANT TO BE ACKNOWLEDGED IN THAT SECTION.

We thank the Reviewer for this comment. We now cite McDonough and Gallo (2012) in that section. Space constraints and the fact that memory results of the font-size test are largely secondary to our research prevented us from developing further the idea of the distinctiveness heuristic.

REVIEWER #4: SUMMARY
THIS STUDY SOUGHT TO EXAMINE WHETHER BELIEFS ABOUT FONT SIZE AND MEMORY INFLUENCE CONFIDENCE JUDGMENTS. EXPERIMENT 1 TESTED WHETHER BELIEFS IN FONT SIZE EFFECTS INFLUENCED CONFIDENCE IN RECOGNITION MEMORY AND/OR SOURCE MEMORY. THE STUDY FOUND SMALL BUT SIGNIFICANT EFFECTS, IN CONTRAST TO PRIOR RESEARCH. EXPERIMENT 2 EXTENDED EXPERIMENT 1 BUT USED A MANIPULATION WHERE PARTICIPANTS WERE TOLD THAT FONT SIZE DID OR DID NOT INFLUENCE MEMORY. AGAIN IT WAS FOUND THAT FONT SIZE INFLUENCED CONFIDENCE. FURTHERMORE, THE EFFECT OF FONT SIZE ON CONFIDENCE WAS LARGEST FOR THOSE WHO ESPoused A BELIEF THAT LARGE FONT WORDS WOULD BE REMEMBERED BETTER.

EVALUATION
OVERALL THE STUDIES WERE WELL DESIGNED AND THE MANUSCRIPT WAS WELL WRITTEN. I ESPECIALLY APPRECIATE PREREGISTRATION OF ALL ANALYSES, INCLUSION OF BOTH NHST AND BAYSIAN STATISTICS, AND CAREFUL INTERPRETATION OF RESULTS.

I HAVE GROUPED MY RECOMMENDATIONS AND CRITIQUES INTO "MAJOR" AND "MINOR" ISSUES, THOUGH I DON'T KNOW THAT ANY OF THEM ARE TRULY "MAJOR." I FEEL THAT ALL OF THEM COULD BE EASILY ADDRESSED IN A RESUBMISSION.
"MAJOR" CONCERNS/SUGGESTIONS

THE PROPORTION OF PEOPLE BELIEVING LARGE FONT WORDS WOULD BE REMEMBERED BETTER, WHETHER THEY WERE TOLD THEY WOULD OR WOULD NOT, WAS SURPRISINGLY LOW IN EXPERIMENT 2. INDEED, THE RATES WERE HIGHER WHEN THEY WERE TOLD NOTHING (EXPERIMENT 1 AND PREVIOUS STUDIES) THAN WHEN THEY WERE TOLD LARGE FONT WOULD BE BETTER. I THINK THIS RAISES SERIOUS QUESTIONS ABOUT THE REACTIVITY OF PRIMING SUCH BELIEFS. CONTRARY TO THE EXPECTED "EXPERIMENTER EXPECTANCY EFFECTS" THE PRIME SEEMS ALMOST TO HAVE AROUSED SUSPICION. SOME DISCUSSION OF THIS IS WARRANTED.

We thank the Reviewer for this suggestion. We added a paragraph in the Discussion of Experiment 2 commenting this issue.

ALONG THE SAME LINE, IT MIGHT HELP TO DISCUSS HOW THE PROPORTION BELIEVING THAT LARGE FONT IMPROVES MEMORY COMPARES TO OTHER STUDIES. THIS WAS COMMENTED ON FOR JOLS BUT NOT BELIEFS.

To test beliefs directly, a popular way is to present a hypothetical experiment and ask participants about the likely results. For example, researchers may say that in a past experiment 40 words were presented, 20 in large font and 20 in small font, and request an estimation of how many words of each type participants in that experiment remembered. That procedure shows how many participants report higher estimates for large than small font, i.e., believe that large words would be better remembered. However, researchers use to report the estimations of the number of words remembered per condition, and the number of participants with better-large or better-small beliefs is not reported (e.g., see Hu et al., 2015; Kornell et al., 2011; or Su et al., 2017). We only found the number of participants holding better-large and better-small beliefs in Mueller et al. (2014, Experiment 3b): 88% believed better-large and 10% believed better-small. Bearing in mind the differences in the measure of beliefs in Mueller et al. (2014) and here, we preferred not to mention it in the text.

MINOR CONCERNS/SUGGESTIONS

SUBHEADINGS FOR THE ACCURACY, CONFIDENCE, HIT RATE, ETC. ANALYSES WOULD HELP.

We thank the Reviewer for this suggestion, but we think that within each test for which we present accuracy and confidence analyses (i.e., font-size and recognition), the results are not so long as to require more subheadings.

ON A STYLISTIC NOTE, I THINK GRAPHS GENERALLY MAKE PATTERNS EASIER TO SEE THAN TABLES. BUT THIS IS JUST MY PERSONAL PREFERENCE.

We added Figure 2 with confidence in the font-size test, which is the most relevant data for the objectives of this research. We also removed the corresponding data from Table 2 to avoid repetition.
Registered Report – Stage 2

Do Beliefs About Font Size Affect Retrospective Metamemory Judgments in Addition to Prospective Judgments? A Registered Report

Karlos Luna¹ and Pedro B. Albuquerque²

¹ Department of Psychology, Universidad Nacional de Colombia (Colombia)
² School of Psychology, University of Minho (Portugal)

Authors’ note

Karlos Luna [ORCID] https://orcid.org/0000-0001-8592-1172

Pedro B. Albuquerque [ORCID] https://orcid.org/0000-0002-5874-4497

The approved registered report, data, and code to run the experiments are available at https://osf.io/z62gk/?view_only=6497c613a6604d97a8aae05e2470cdf8.

The authors have no conflicts of interest to disclose. Thanks to Nuno Fernandes for help with data collection.

Correspondence concerning this article should be addressed to Karlos Luna, Carrera 30 # 45-03, Edif. 212 Departamento de Psicología, Universidad Nacional de Colombia, Ciudad Universitaria, Bogotá, 111321, Colombia. E-mail: klunao@unal.edu.co
Abstract

Beliefs about how memory works explain several effects on prospective metamemory judgments (e.g., the effect of font size on JOLs). Less is known about the effect of beliefs on retrospective judgments (i.e., confidence). Here, we tested whether font size also affects confidence ratings and whether beliefs play a similar role in confidence than in JOLs. In two experiments, participants studied words in small and large size, rated JOLs, and completed a font-size test in which they indicated the font size at study and a standard old/new recognition test. The results confirmed that font size affected both JOLs and confidence ratings. The presentation of the counter-belief that memory is better for words in small font size in Experiment 2 and the analyses of confidence for participants who did not believe that large fonts improved memory suggested that the effect of font size on confidence was based on beliefs. This research shows that the debate on theory-based and experience-based factors should not be limited to prospective metamemory judgments, but also encompass retrospective judgments.

Keywords: metamemory, beliefs, confidence, font size
Do Beliefs About Font Size Affect Retrospective Metamemory Judgments in Addition to Prospective Judgments? A Registered Report

Currently, there is debate on whether metamemory judgments are based on theory-based or experience-based factors. Theory-based factors refer to the use of knowledge and beliefs about how memory works when making metamemory judgments. Experience-based factors are metamemory cues that derive from the direct experience with the items or with the task, such as retrieval fluency or retrieval speed. That debate has been informed mainly by research with prospective metamemory judgments, such as judgments of learning (JOLs, e.g., Ball et al., 2014; Besken, 2016; Besken et al., 2019; Frank & Kuhlmann, 2017; Undorf & Zimdahl, 2019; Witherby & Tauber, 2017; Yan et al., 2016) or ease-of-learning judgments (EOLs, e.g., Jemstedt et al., 2018), or with judgments about the source in which an object was presented (judgments of source or JOSs, e.g., Schaper et al., 2019).

Our aim here is to extend that debate to retrospective metamemory judgments, namely confidence (for research with the opposite objective to extend an effect from retrospective to prospective metamemory judgments, see Monds et al., 2019). Past research has shown that experience-based factors such as retrieval fluency or latency play a major role in confidence (e.g., Kelly & Lindsay, 1993; Weber & Brewer, 2006), but there is little research on the effect of theory-based factors and beliefs in confidence. Thus, our main objective is to explore the effects of metamemory beliefs on confidence. Both prospective and retrospective judgments are usually understood as stemming from the same monitoring process that examines and evaluates our memories (Nelson & Narens, 1990; Schwartz & Metcalfe, 2017). However, McDonough et al. (2021) found
that monitoring was better conceptualized as two related sets of processes: one at
encoding and another at retrieval. This raises the question of whether beliefs affect
retrospective confidence at retrieval similarly to JOLs at encoding. The answer to this
question will help us identify similarities and differences between both judgments and
allow us to better understand the monitoring process(es) at encoding and retrieval.

To study the role of beliefs on confidence, it is important to ensure that there is a
belief. Some manipulations of perceptual information might prove useful, namely the
font size and the sound volume. As per font size, participants rate materials presented in
larger font size with higher JOLs (Luna, Albuquerque, & Martín-Luengo, 2019; Rhodes
& Castel, 2008; Undorf & Zimdahl, 2019; for a meta-analysis, see Luna et al., 2018). As
per sound volume, words presented in louder volume are also rated with higher JOLs
(Frank & Kuhlmann, 2017; Peynircioğlu & Tatz, 2019; Rhodes & Castel, 2009).

Current knowledge strongly suggests that these results are mostly due to people’s
beliefs (Blake & Castel, 2018; Luna, Nogueira, & Albuquerque, 2019; Undorf &
Zimdahl, 2019): 1 People believe that words in large font or louder volume are better
remembered than words in small font or lower volume (Frank & Kuhlmann, 2017;
Kornell et al., 2011; Mueller et al., 2014). Thus, perceptual information provides the
perfect opportunity to test the effects of beliefs on confidence. In this research, we focus
on font size for practical reasons, but we also review relevant studies about volume.

In the remaining of this introduction, we first describe studies suggesting that font
size and volume, in general, do not affect confidence or that the effect is, at best, small.

Then, we argue that there are theoretical and empirical reasons to expect an effect of
1 There was a debate on whether experience-based factors (e.g., fluency) could also account for
that effect, but Luna, Nogueira, and Albuquerque (2019) and Undorf and Zimdahl (2019) found that with
very large font sizes the effect of font size on JOLs was higher, but processing fluency was lower. These
results are incompatible with the idea that experience-based factors are mostly responsible for the font-
size effect on JOLs.
beliefs on confidence and offer an explanation for the results from previous studies.

Finally, we present the current research and the hypotheses described in the registered report (available at

https://osf.io/z62gk/?view_only=6497c613a6604d97a8aae05e2470cdf8).

**Does Perceptual Information Affect Confidence Ratings?**

To our knowledge, only seven experiments in three papers manipulated font size (six experiments) or volume (one experiment) and collected confidence ratings. All the experiments collected JOLs and showed the expected effect, i.e., higher JOLs for large than small font and for loud than quiet volume. All the experiments but two showed no effect of perceptual information on confidence.

Three of the experiments that collected confidence ratings were presented in Luna et al. (2018; see the Supplemental Materials of that article for a full description). In Study 1, participants read words presented in either small or large font size, completed a recognition test with confidence ratings, and font size did not affect confidence.

However, there was no information about font size at test, so participants may have just ignored that cue. Thus, the information that could be obtained from that experiment is limited. In Studies 2 and 5, participants read word pairs in small and large font and completed a cued recall test in which the cues were presented in the same font as during the study phase. Again, font size did not affect confidence. However, in a cued recall test participants have to generate the answer, which may produce a set of cues that may be either more diagnostic or weigh more in the confidence rating (e.g., retrieval fluency or speed of retrieval, Kelley & Lindsay, 1993), thus reducing the need of other cues such as font size.
Maybe in a recognition memory test participants will not have the cues stemming from the process of generating the answers and, thus, may rely on other cues for confidence ratings, including beliefs. For example, McDonough and Gallo (2012) presented words twice, in either small, large, or both font sizes. Then, participants completed several yes/no recognition tests with studied and new words presented in intermediate font size. In the “large” test, participants had to answer “yes” if the word was studied in large font and in the “small” test they had to answer “yes” if the word was studied in small font. In both tests, participants also rated their confidence. Across three experiments, confidence was higher for hits in the “large” test (i.e., answers “large” to words studied in large size) than in the “small” test (i.e., answers “small” to words studied in small size), although these effects were small ($d$ between 0.10$^2$ and 0.43) and were not statistically significant in Experiment 2. The authors concluded that “These differences are consistent with the idea that participants carried a false expectation that large words were better remembered than small words” (p. 395). In other words, the results from McDonough and Gallo (2012) suggest that, under certain conditions, beliefs about perceptual information may affect confidence.

In particular, three procedural details in McDonough and Gallo (2012) may have helped beliefs affect confidence. First, items were presented twice, which may have increased participants’ ability to recollect font size at test. Second, the difference between small (25 points) and large font size (125 points) was larger than in other studies. When the difference between small and large font size increases, the effect on JOLs is larger (Halamish, 2018; Undorf & Zimdahl, 2019), and, thus, a larger font size

---

2 McDonough and Gallo (2012) did not report Cohen’s $d$ for non-significant results. We computed this effect size from the reported descriptive statistics, and thus this number should be considered an approximation to the actual figure.
difference may have increased the chances of detecting an effect on confidence. And third, at test participants were requested to recognize only words in large or in small font, which may have made font size salient at test. If font size affects metamemory judgments via beliefs, those beliefs must be active to affect metamemory (Kornell et al., 2011). Thus, it may be that by making font size salient, the task activated the belief that font size affects memory. These procedural details may have increased the sensitivity of the test to detect the effect of beliefs on confidence. However, some other procedural details may have worked in the opposite direction. For example, McDonough and Gallo presented 144 words twice at study. The high number of words at study may have made font size recollection for a specific item more difficult.

Finally, Frank and Kuhlmann (2017) asked participants whether they thought they were going to remember better words in quiet or in loud volume. Seventy-five percent of participants reported believing that loud words were going to be remembered better, thus ensuring that they had a belief in the expected direction. Then, participants listened to words in quiet or loud volume and completed a written old/new recognition test with confidence. Importantly, for the confidence rating the authors provided information about the volume at study. If participants paid attention to that information, that should have activated the pre-existing belief as in McDonough and Gallo (2012). Despite that, the results did not show an effect of volume on confidence.

In sum, the above results suggest that, in general, perceptual information, either font size or sound volume, has a very limited effect on confidence ratings. Most of the experiments did not find such effect, but under certain conditions it can be detected. One explanation for this pattern of results is that theory-based factors do not affect
confidence and the positive results were Type-I errors. We now provide some evidence that suggests otherwise.

**Do Beliefs Affect Confidence Ratings?**

Metamemory judgments are inferences made from cues available at the time of the judgment (e.g., Koriat, 1997), and many researchers have included metacognitive beliefs in their general understanding of the cues for confidence ratings (e.g., Buratti & Allwood, 2015; Koriat et al., 2008). In addition, there are also empirical examples of beliefs affecting confidence.

For example, Leippe et al. (2009, p. 195) proposed the cue-belief model of retrospective confidence, based on Koriat’s (1997) cue-utilization approach to JOLs. The cue-belief model proposes that confidence is an inference made from cues emanating from three sources: one’s beliefs, one’s mental processes, and the testing situation. In line with their model and the proposed relevant role of beliefs, Leippe et al. found that the belief that if memory is good for one aspect of an event, then memory will also be good for other aspects of the same event explained why feedback increased confidence. Perfect (2004) offered another example of beliefs affecting retrospective confidence. Before the experiment, participants stated that they would perform better at recognizing faces than answering questions about sports, and although there were no differences in actual performance, they still rated faces with higher confidence, consistently with their initial belief.

Similarly, Brewer et al. (2005) also proposed that retrospective confidence was affected by metamemory beliefs. In particular, they tested the effect of the belief that a complete recall is an accurate recall on confidence. The authors presented sentences that promoted a synonym substitution, e.g., from “Russian is difficult to pronounce” to
“Russian is hard to pronounce”, and later measured memory, confidence, and whether participants thought that they recalled the sentence completely, partially, or not at all. In the cued recall test, participants committed many synonym substitutions and, despite the warning of the strict correction criteria, they believed that their recall was complete. Importantly, participants rated the synonym substitutions with similar confidence than actually complete responses and sentences that did not promote a synonym substitution. These results were interpreted as indicating that participants hold the metamemory belief that a complete recall is a good indicator of accuracy and thus rated confidence accordingly (for a related study with similar results and conclusions, see Brewer & Sampaio, 2006).

Other examples of retrospective confidence models include beliefs or theory-based factors in the form of previous knowledge used as a cue to confidence ratings. The theory of probabilistic mental models (Gigerenzer et al., 1991) prominently features prior knowledge's effect on memory and confidence ratings. Similarly, the consensuality principle (Koriat, 2008), the empirical result that confidence in two-alternative forced-choice questions is correlated with the proportion of people endorsing one of the alternatives and not with the accuracy of the answer, is also consistent with theory-based factors.

In sum, the distinction between theory-based and experience-based factors seems transversal to metamemory judgments, and thus it appears that beliefs should influence any type of metamemory judgment, including confidence. Then, why does the belief that memory is better for words presented in large font have a robust and consistent effect on prospective metamemory judgments but not on confidence?

The Present Research
We propose that the conditions in previous studies were not optimal to find the expected effect of beliefs on confidence. In particular, all the experiments included a memory task in which participants had to retrieve or recognize the items. Experience-based cues stemming from item retrieval, such as familiarity (Yonelinas, 1994), subjective vividness (Robinson et al., 2000), or retrieval fluency (Kelley & Lindsay, 1993), might be more relevant than theory-based factors for retrospective judgments (Frank & Kuhlmann, 2017). This may explain why previous studies mostly failed to show an effect of beliefs on retrospective confidence. To overcome this limitation, we propose a memory test focused on font size to highlight theory-based cues (i.e., a font-size test), which might provide more room to detect an effect of beliefs on confidence. In the font-size test, participants indicate the font size in which words previously studied were presented.

The proposed font-size test is similar to a source memory test, in which participants are requested to indicate the source of an item or to indicate some characteristic present at study. However, note that in a typical source memory test, item retrieval is also mixed with the source task, that is, participants had to indicate whether an item was presented by source 1, source 2, or it was new. In the font-size test, we only presented studied words and asked participants to identify font size, which simplifies the task. Despite this difference, past studies on source memory can inform us whether participants remember details such as the voice that read a statement or the colour in which it was presented. In general, the results confirm that memory for characteristics of the items is good, even after a delay or in a surprise test. For example, in an immediate test young adults correctly identified 77% of the time that the voice that read a sentence was male or female, and after 24 hours they still identified it correctly 64%
of the time (Dodson, Bawa, & Slotnick, 2007; see also Dodson, Bawa, & Krueger, 2007, for similar results). Likewise, in a surprise test participants correctly recognized the colour in which a sentence was studied 69% of the time and the voice that read it 68% of the time (Fazio & Marsh, 2009). These results show that memory for item characteristics is good in general, and thus we expect that during the font-size test participants will retrieve the font size at study.

As mentioned, with the font-size test we expect to elicit cues related to font size. Some of these cues will be experience-based, but we expect that beliefs about the effect of font size will also be activated and used to rate confidence. If beliefs affect retrospective confidence in a similar way than JOLs, then we expect that participants responding that a word was studied in large font size will access the belief that these words are better remembered, which should increase confidence. The underlying process could be exemplified by a participant thinking “I remember that this word was presented in large font size; of course I remember, words in large font are better remembered”. Such thinking will not happen with words in small font.

**Hypotheses**

To test the effect of font size and beliefs in confidence, we conducted two experiments approved by the local Ethics Committee. In both experiments, we presented participants with words in either small or large font and requested JOLs. We expected higher JOLs for words in large than in small font. After that, participants completed the font-size test, in which they had to identify the font size in which words were studied and rated their confidence in the response. In this test, we expected no differences in font-size identification. The effect of beliefs on confidence may be shown in two non-exclusive ways. First, confidence may be higher for responses “large” than
for responses “small”. Second, the effect may be limited to correct responses, that is, there may be higher confidence for “large” correct responses than for “small” correct responses. This may be the case if participants access the actual font size at study. In that case, beliefs may boost confidence mainly for correct responses to words in large font, with no effect on incorrect “large” responses.

Finally, participants also completed a standard yes/no recognition test with confidence to provide a comparison with previous studies. Most studies on font size have found no effect of font size on memory (Kornell et al., 2011; Mueller et al., 2014; Rhodes & Castel, 2008), although others did (Halamish, 2018; Luna et al., 2018; Price & Harrison, 2017; Undorf & Zimdahl, 2019). When present, the effect of font size on memory is usually small and limited to studies with free recall test (Chang & Brainerd, in press; Luna et al., 2018). Thus, we do not expect any effect of font size on memory. As per confidence, the results may depend on the strength of the effect of beliefs in confidence. If the effect is strong, then all the analyses could show higher confidence for words in large than in small font size, but if the effect is weak it may be limited to some types of response, particularly hits.

Finally, participants answered a question to test their beliefs directly. Following previous research (e.g., Mueller et al., 2014), we expect that most participants will believe that words in large font are better remembered.

**Experiment 1**

**Method**

**Participants**

McDonough and Gallo (2012) found effect sizes ranging from 0.10 to 0.43. We expect this experiment to be more sensitive to detect the effects of beliefs on
confidence, and aimed at a sample size able to detect an effect size within that range, 

\textit{Cohen's }d = 0.35. A power analysis with the package \textit{pwr} (Champely, 2018) in R (R Core Team, 2018) showed that 66 participants were necessary to obtain a power = .80 with alpha = .05. A total of 72 university students took part in the experiment (55 females, age M = 20.76, SD = 3.38) in exchange for course credits. The only requisite was that participants were native speakers of the language in which the experiment was conducted.

\textit{Design}

The only variable was font size (small, 18 points; large, 100 points), manipulated within subject. We used a 100 points large font size instead of the usual 48 points because the font-size effect is stronger with larger font sizes (Luna, Nogueira, & Albuquerque, 2019; Undorf & Zimdahl, 2019). With a larger-than-usual font size, we expected to increase the chances of finding an effect on confidence.

\textit{Materials}

One hundred and four singular nouns were selected from a linguistic database. For counterbalancing purposes, these words were separated into six lists of 16 words, matched in number of letters and word frequency. The remaining eight words were used as primacy and recency buffers during study. Four lists of 16 words were used during the study phase, two of them were later used in the font-size test and the other two in the recognition test. The two final lists were used as non-studied words for the recognition test (see below).

\textit{Procedure}

The experiment was programmed in LiveCode (version 9.5.1), and participants completed the experiment in individual soundproof booths. Executables and code from
the original experiments are available at the OSF. All responses were provided by selecting the appropriate radio buttons. First, participants provided informed consent and basic demographic information. Then, they read the instructions that encouraged them to pay attention to words because later there would be a memory test. To increase the chances that participants encoded font size and later recollected it during the font-size test, instructions mentioned that they would have to remember the item and the font size in which it was presented.

The experiment included a study phase and a test phase with two memory tests. Figure 1 shows a summary of the procedure. In the study phase, each trial started with a fixation cross presented for 500 milliseconds, followed by the word for four seconds. Words were presented in the centre of the screen, half in small (18 points) and half in large font size (100 points), counterbalanced. After each word, participants rated the chances that they would recall the word later on a scale from 0 to 100 in deciles. The first and last four words were buffers and were not analysed. Words were presented in a pseudo-random order with no more than two words in the same font size in a row. After 72 trials, participants completed a filler task of writing European cities for two minutes.

Figure 1

*Summary of the procedure used in Experiments 1 and 2*
The test phase included a font-size and a standard recognition test, both without time limit. In the font-size test, instructions mentioned that 32 studied words would be presented again and that participants had to indicate the size in which they were presented. Words were presented in a different order than at study. After a fixation point of 500 ms, each word was presented in both the small (18 pt) and large (100 pt) font sizes on the same screen; radio buttons below allowed participants to indicate the font size at study. The position of the large and small words was counterbalanced across participants. Immediately below, participants indicated the confidence that they had selected the correct font size on a scale from 1 (extremely low) to 7 (extremely high), following McDonough and Gallo (2012). This scale allowed a clear separation of confidence and JOLs (on a percentage scale). After the font-size test, participants
completed another filler task in which they listed fruits, mammals, trees, and fishes for one minute each.

Then, participants completed a standard recognition test. The 32 studied words not used in the font-size test and 32 new words (counterbalanced) were presented in pseudo-random order, one at a time, with no more than two words of the same type in a row. Studied words were presented in the same font size as during the study phase, and half of the new words were presented in small (18 pt) and half in large font size (100 pt). Instructions mentioned that font size of studied words corresponded to that during the study phase. This served to avoid participants selecting “no” for a word they recognized but they thought it was studied in a different font. Participants answered yes/no the word was studied. On the same screen, participants also rated their confidence that they selected the correct answer on the same scale from 1 to 7 as in the font-size test.

The order of the tests was not counterbalanced for two reasons. First, our main objective was better tested with the font-size test and the recognition test was included only for comparison purposes. Thus, it was important to prevent the secondary recognition test from interfering with the main font-size test, which could happen in the recognition-test-first order (see McDonough & Gallo, 2012, for a similar justification). And, second, administering the recognition-test-first order would increase the delay between the study phase and the font-size test, which may make recollection of the font size too difficult due to forgetting or interference from the new items in the recognition test.

Finally, to check participants’ beliefs about the effect of font size on memory, they answered the following question, partially based on Mueller and Dunlosky (2017,
Experiment 5): “In this experiment, you read words in small and large font size. Which font size do you think leads to a better memory?”. The three alternative responses were:

- I think that, in general, words in large font size are better remembered.
- I think that, in general, words in small font size are better remembered.
- I think that, in general, they will be similarly remembered.

Results

We expected that font size would affect confidence, but it may not. To get evidence in support or against our hypotheses, the analyses of confidence directly related to our key hypotheses were conducted with both the NHST and Bayesian perspectives. For the NHST analyses, alpha was set to .05. For the Bayesian analyses, we compared the hypothesis of differences between conditions (H1) against the hypothesis of no differences (H2). When the reported $BFs$ are higher than 1 they support H1, and when lower they support H2. We used the default Cauchy prior and the standard criteria of 0.33, 1, 3, and so on to determine the strength of the evidence in favour of H1 or H2. All the analyses were conducted with R (R Core Team, 2018) and a copy of the data is available at the OSF.

Question about Beliefs

Seventy-four percent of participants indicated that they thought they would remember better words in large font, 4% that they would remember better words in small font, and 22% that they would remember both equally. These results confirmed that most participants held the belief that words in large font are better remembered.

Judgments of Learning
Participants attributed higher JOLs to words in large than in small font, \( t(71) = 9.43, p < .001, d_{av} = 0.87 \) (see Table 1). This result replicates the typical font size effect in JOLs extensively reported before.

**Table 1**

*Mean (SD) for JOLs in Experiments 1 and 2*

<table>
<thead>
<tr>
<th></th>
<th>Small</th>
<th>Large</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>37.94 (16.81)</td>
<td>52.72 (17.07)</td>
<td></td>
</tr>
<tr>
<td>E2 – Better Small</td>
<td>38.52 (15.47)</td>
<td>43.79 (15.73)</td>
<td>41.16 (14.67)</td>
</tr>
<tr>
<td>E2 – Better Large</td>
<td>32.23 (16.11)</td>
<td>51.07 (16.01)</td>
<td>41.65 (14.10)</td>
</tr>
<tr>
<td>E2 – Total</td>
<td>35.35 (16.06)</td>
<td>47.46 (16.23)</td>
<td></td>
</tr>
</tbody>
</table>

*Note:* E1: Experiment 1; E2: Experiment 2; JOLs on a scale from 0 to 100%.

**Font-Size Test**

We conducted two analyses, one with the proportion of correct identifications of the font size at study and another with confidence in the response. First, participants identified better the font size when the words were studied in small font \( (M = .78, SD = .13) \) than in large font \( (M = .66, SD = .18) \), \( t(71) = 4.36, p < .001, d_{av} = 0.70 \).

Performance in both conditions was better than chance, for small \( t(71) = 18.13, p < .001, d_{av} = 2.14 \), and for large, \( t(71) = 7.59, p < .001, d_{av} = 0.89 \).

**Table 2**

*Mean (SD) for Confidence in the Font Size Test in Experiments 1 and 2*

<table>
<thead>
<tr>
<th></th>
<th>Answer “Small”</th>
<th>Answer “Large”</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>4.44 (0.99)</td>
<td>5.14 (0.99)</td>
</tr>
<tr>
<td>Correct Responses</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Incorrect Responses  3.68 (1.04)  4.34 (1.30)

**Experiment 2 – Better Large**
Correct Responses  4.43 (1.10)  5.07 (1.17)
Incorrect Responses  3.53 (1.06)  3.95 (1.29)

**Experiment 2 – Better Small**
Correct Responses  4.27 (1.19)  4.62 (1.17)
Incorrect Responses  3.40 (1.21)  3.60 (1.34)

*Note:* Confidence on a scale from 1 to 7.

Second, for confidence we conducted an analysis of variance 2 (type of answer: “small”, “large”) x 2 (type of response: correct, incorrect). See Figure 2 and Table 2. Six participants were lost for this analysis because they did not have data in all the cells.

Confidence was higher for answers “large” than “small”, $F(1, 65) = 30.29, p < .001, \eta_p^2 = .32, BF_{inc} = 7.71 \times 10^7$, and for correct ($M = 4.79, SD = 0.83$) than incorrect responses ($M = 4.01, SD = 0.99$), $F(1, 65) = 73.11, p < .001, \eta_p^2 = .53, BF_{inc} = 1.73 \times 10^{10}$. The interaction was not significant, $F(1, 65) = 0.04, p = .839, \eta_p^2 < .01$, with an inconclusive $BF_{inc} = 0.763$. In sum, both NHST and Bayesian analyses were consistent in that font size affected confidence in the responses in the font-size test.

**Figure 2**

*Confidence in the Font Size Test in Experiments 1 and 2. Error bars indicate SD*
Recognition Memory Test

We also conducted analysis of memory performance and confidence. The main
descriptive statistics for measures of memory performance are presented in Table 3.
There were more hits with items in large than in small font, \( t(71) = 3.29, p = .002, dav = 0.34 \), no differences in false alarms, \( t(71) = 0.85, p = .396, dav = 0.10 \), and better
accuracy, measured with \( d' \), for words in large than small font, \( t(71) = 2.77, p = .007, dav = 0.28 \).

Table 3

Mean (SD) for Hits, False Alarms, and \( d' \) in the Recognition Test in Experiments 1 and 2

<table>
<thead>
<tr>
<th></th>
<th>Hits</th>
<th>False Alarms</th>
<th>( d' )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small</td>
<td>Large</td>
<td>Small</td>
</tr>
<tr>
<td>E1</td>
<td>.78 (.16)</td>
<td>.83 (.13)</td>
<td>.28 (.15)</td>
</tr>
</tbody>
</table>
For confidence, we conducted an analysis of variance 2 (font size: small, large) x 2 (type of response: correct, incorrect). Two participants were lost because they did not have data in all the cells. Results suggested that confidence was higher for words studied in large than in small font size, but the result was not particularly clear, $F(1, 69) = 5.68$, $p = .020$, $\eta^2 = .08$, with an anecdotal $BF_{inc} = 2.04$ (see Table 4). Confidence was higher for correct than incorrect responses, $F(1, 69) = 178.61$, $p < .001$, $\eta^2 = .72$, $BF_{inc} = 2.91 \times 10^{31}$, and the interaction was not significant, $F(1, 69) = 2.50$, $p = .119$, $\eta^2 = .03$, with an inconclusive $BF_{inc} = 1.08$. In sum, the effect of font size on confidence in the recognition test was small and barely noticeable, but the results were in the same direction of those from the font-size test.

### Table 4

*Mean (SD) for Confidence in the Recognition Test in Experiments 1 and 2*

<table>
<thead>
<tr>
<th></th>
<th>Small Font Size</th>
<th>Large Font Size</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct Responses</td>
<td>5.02 (0.81)</td>
<td>5.24 (0.82)</td>
<td>5.13 (0.79)</td>
</tr>
<tr>
<td>Incorrect Responses</td>
<td>4.20 (0.99)</td>
<td>4.27 (1.00)</td>
<td>4.23 (0.91)</td>
</tr>
<tr>
<td>Total</td>
<td>4.61 (0.85)</td>
<td>4.76 (0.83)</td>
<td></td>
</tr>
<tr>
<td><strong>Experiment 2 – Better Large</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct Responses</td>
<td>4.99 (1.00)</td>
<td>5.20 (0.94)</td>
<td>5.10 (0.94)</td>
</tr>
</tbody>
</table>

3 We registered a 2 (type of response: “small”, “large”) x 2 (type of response: correct, incorrect) similar to that of the font-size test, but later realized that there are no responses “small” or “large” in this test. To follow the analyses in previous research, we replaced type of response by font size.
To further test the effect of font size in retrospective confidence, we also analysed confidence for hits and false alarms. For hits, results suggested a higher confidence for words in large \((M = 6.06, SD = 0.88)\) than small font size \((M = 5.90, SD = 0.85)\), \(t(71) = 2.07, p = .042, d_{av} = 0.18\), but with an inconclusive \(BF = 0.96\). For false alarms, both analyses were consistent in showing no effect of font size on confidence, with moderate evidence in support of no differences, \(t(71) = 1.05, p = .297, d_{av} = 0.08, BF = 0.22\) (large \(M = 4.21, SD = 1.69\); small \(M = 4.09, SD = 1.69\)).

**Exploratory Analyses**

The analyses in this section were not registered. The results above suggest that font size affects retrospective confidence in the font-size test and, to a lesser degree, also in the recognition test. However, these effects could be driven by beliefs or by other factors. One way to get information on the role of beliefs is to compare the effect of font size in participants believing and not believing that large words would be better remembered. If beliefs drive the observed effects of font size on confidence, then the effects should be reduced or eliminated for participants without the belief. We conducted a 2 (belief: yes, no) x 2 (type of answer: “small”, “large”) analysis of variance with confidence in the font-size test and found a significant interaction, \(F(1, \ldots)\)
70) = 9.52, \( p = .003 \), \( \eta^2_p = .12 \), \( BF_{inc} = 18.23 \). In the belief group, the effect of font size on confidence was significant (large \( M = 5.11 \), \( SD = 1.05 \); small \( M = 4.22 \), \( SD = 1.05 \), \( t(52) = 6.69, p < .001 \), \( d_{av} = 0.85 \), \( BF = 8.28 \times 10^5 \)), but it was not for participants in the no-belief group (large \( M = 4.77 \), \( SD = 0.84 \); small \( M = 4.67 \), \( SD = 0.71 \), \( t(18) = 0.47, p = .645 \), \( d_{av} = 0.13 \), \( BF = 0.26 \).

Also, in the recognition test a 2 (belief: yes, no) x 2 (font size: small, large) analysis with confidence showed the same results. The interaction was significant, \( F(1, 70) = 5.42, p = .023 \), \( \eta^2_p = .07 \), \( BF_{inc} = 6.10 \), and the effect of font size on confidence was significant in the belief group (large \( M = 5.01 \), \( SD = 0.86 \); small \( M = 4.73 \), \( SD = 0.90 \), \( t(52) = 4.84, p < .001 \), \( d_{av} = 0.33 \), \( BF = 1.58 \times 10^3 \)), and not significant in the no-belief group (large \( M = 5.27 \), \( SD = 0.77 \); small \( M = 5.24 \), \( SD = 0.64 \), \( t(18) = 0.28, p = .785 \), \( d_{av} = 0.03 \), \( BF = 0.25 \)). In sum, these analyses suggest that font size affects confidence only for participants believing that large words are better remembered.

**Discussion**

Experiment 1 replicated the effect of font size on JOLs and showed an effect of font size on confidence in the font-size test. In the recognition test, results were in the same line but were less clear. Existing metamemory beliefs could explain these results, an interpretation supported by the exploratory analyses. However, font size could also affect confidence through experience-based factors. With the font-size test, we tried to reduce experiential cues, but they may still provide a reasonable explanation. We tested the influence of beliefs in the effect of font size on confidence in Experiment 2.

**Experiment 2**

In Experiment 2, we manipulated the strength of participants’ beliefs by presenting information either in support of a better memory for words in large or in
small font size (a counter-belief). If beliefs drive the effect of font size on confidence, then we expected that the counter-belief would reduce, eliminate, or revert that effect.

Method

Participants and Design

Experiment 2 was a 2 (font size: small, 18 points; large, 100 points) 2 (instructions: better memory for small font size, better memory for large font size) mixed design, with instructions manipulated between participants. In Experiment 1, we found moderate to large effects of font size in confidence (font-size test $\eta_p^2 = .32$, and recognition $\eta_p^2 = .08$), but our hypotheses pertained to the interaction. Thus, we determined the sample size to find a medium effect in the interaction. We registered the use of GPower (Faul et al., 2007) to compute sample size, but after reading Brysbaert (2019), we were unconvinced by the result that a total of 34 participants were needed to get power = .80 in the interaction in a 2x2 mixed design. Brysbaert (2019) suggests 130 participants (see Table 7 of Brysbaert, 2019). This result is consistent with other power analysis software, MorePower (Campbell & Thompson, 2012) and WebPower (Zhang & Yuan, 2018). Thus, we aimed at a sample size of 130 with 65 participants in each group, but we ended up collecting more participants to match the number of participants by counterbalance group. A total of 145 participants completed the experiment in exchange for course credits (117 female, age $M = 20.75, SD = 4.15$), with participants randomly allocated to each instruction (72 to the better small and 73 to the better large).

Materials and Procedure

Experiment 2 used the same materials and procedure as in Experiment 1, with the exception that the instructions at the start of the experiment and before the font-size test presented a brief justification for the better memory for words in small or large font.
The instructions were based on Blake and Castel (2018, Experiment 2): “Research has shown that, for college-age participants, words in [larger/smaller] fonts are easier to recall than words in [smaller/larger] fonts”.

**Results**

**Questions about Beliefs**

In the group with better large instructions, 55% of the participants indicated that they thought they would remember better words in large font, 4% that they would remember better words in small font, and 41% that they would remember them equally well. In the group with better small instructions, 47% thought that they would remember better words in large font, 10% that they would remember better words in small font, and 42% that they would remember them equally. These results suggest that our manipulation of beliefs had an effect, albeit small, on participants’ beliefs.

**Judgments of Learning**

An analysis of variance 2 (font size: small, large) x 2 (instructions: better small, better large) showed higher JOLs for words in large than small font size, $F(1, 143) = 121.24, p < .001, \eta^2_p = .46$, no effect of instructions, $F(1, 143) = 0.04, p = .838, \eta^2_p < .01$, and a significant interaction, $F(1, 143) = 38.13, p < .001, \eta^2_p = .21$ (see Table 1).

The effect of font size on JOLs was present in both groups, for the better small $t(71) = 4.21, p < .001, d_{av} = 0.34$ and for the better large $t(72) = 10.46, p < .001, d_{av} = 1.17$, and the effect was smaller in the former than in the latter, $t(143) = 6.17, p < .001, d_{av} = 1.03$.

Consistent with the results above, JOLs showed that the manipulation of the beliefs was successful, but that the effect was small because the counter-belief did not eliminate the font size effect.

**Font-Size Test**
For memory performance, we found better memory for the font size of words presented in small ($M = .76$, $SD = .14$) than in large font size ($M = .63$, $SD = .16$), $F(1, 143) = 60.33$, $p < .001$, $\eta_p^2 = .30$. We found no effect of instructions, $F(1, 143) = 0.24$, $p = .623$, $\eta_p^2 < .01$, nor the interaction, $F(1, 143) = 0.84$, $p = .361$, $\eta_p^2 < .01$. Performance was better than chance for both small $t(144) = 22.68$, $p < .001$, $d_{av} = 1.88$, and large words, $t(144) = 9.47$, $p < .001$, $d_{av} = 0.79$.

As per confidence, we conducted a 2 (type of answer: “small”, “large”) x 2 (type of response: correct, incorrect) x 2 (instructions: better small, better large). Twelve participants were lost for this analysis. Confidence was higher for answers “large” ($M = 4.31$, $SD = 1.12$) than “small” ($M = 3.91$, $SD = 1.04$), $F(1, 131) = 27.24$, $p < .001$, $\eta_p^2 = .17$, $BF_{inc} = 3.93 \times 10^{15}$, and for correct ($M = 4.60$, $SD = 1.07$) than incorrect responses ($M = 3.62$, $SD = 1.05$), $F(1, 131) = 204.69$, $p < .001$, $\eta_p^2 = .61$, $BF_{inc} = 4.74 \times 10^{82}$ (see Table 2). There was no effect of instructions, $F(1, 131) = 2.51$, $p = .115$, $\eta_p^2 = .02$, $BF_{inc} = 1.55$. These results replicate those from Experiment 1 in which font size affected confidence.

The interaction between font size and instructions was not significant $F(1, 131) = 2.67$, $p = .105$, $\eta_p^2 = .02$, but the Bayesian analysis showed moderate evidence in support of differences, $BF_{inc} = 4.44$ (see Figure 2). This interaction tested our main hypothesis, so we conducted further analyses to explore it. With both instructions, confidence was higher for words in large than small font size. With the better large instructions the effect size was medium and the $BF$ showed extreme evidence in support of differences, $t(66) = 6.52$, $p < .001$, $d_{av} = 0.58$, $BF = 1.06 \times 10^5$. With the better small instructions the effect size and $BF$ were smaller at a descriptive level, $t(65) = 3.38$, $p = .001$, $d_{av} = 0.32$, $BF = 21.24$. However, the difference between instructions was not large
enough as to show in the statistical comparisons, $t(131) = 1.87, p = .063, d_{av} = 0.33,$ including an inconclusive $BF = 0.91$. These results hint that the counter-belief may have reduced the effect of font size on confidence, but that effect was not particularly clear. This goes in line with the small effect of the beliefs manipulation.

**Recognition Memory Test**

For memory performance, there were no differences in hits and false alarms, but accuracy measured with $d'$ was higher with words in large ($M = 1.71, SD = 0.75$) than in small font size ($M = 1.59, SD = 0.76$), $F(1, 143) = 4.38, p = .038, \eta_p^2 = .03$ (see Table 3).

For confidence, we conducted an analysis of variance 2 (font size: small, large) x 2 (type of response: correct, incorrect) x 2 (instructions: better small, better large). Four participants were lost for this analysis. Confidence was higher for words in large ($M = 4.72, SD = 0.96$) than in small font ($M = 4.58, SD = 1.00$), $F(1, 139) = 10.87, p = .001, \eta_p^2 = .07, BF_{inc} = 1.07 \times 10^3$, and also higher for correct ($M = 5.14, SD = 0.90$) than incorrect responses ($M = 4.17, SD = 1.09$), $F(1, 139) = 316.22, p < .001, \eta_p^2 = .69, BF_{inc} = 6.10 \times 10^{150}$ (see Table 4). There was no effect of instructions, $F(1, 139) < 0.01, p = .983, \eta_p^2 < .01, BF_{inc} = 0.38$, and no interactions.

We also analysed confidence for hits and false alarms. A 2 (font size: small, large) x 2 (instructions: better small, better large) analysis of variance showed higher confidence for hits with words in large font size ($M = 6.01, SD = 0.85$) than small ($M = 5.85, SD = 0.93$), $F(1, 143) = 11.82, p = .001, \eta_p^2 = .08, BF_{inc} = 20.89$, and no other effects. A similar analysis with confidence in false alarms showed no effects. In particular, font size did not affect confidence in false alarms with strong evidence in
support of no differences, $F(1, 143) = 0.09, p = .760, \eta_p^2 < .01, BF_{inc} = 0.09$ (for words in large font, $M = 4.23, SD = 1.60$, and for words in small font, $M = 4.18, SD = 1.78$).

**Exploratory analyses**

The analyses in this section were not registered. We replicated here the analysis of Experiment 1. Seventy-five participants believed that memory would be better for words in large font size and 70 believed otherwise. For the font-size test, we conducted a 2 (belief: yes, no) x 2 (type of answer: “small”, “large”) analysis of variance with confidence and found a significant interaction, $F(1, 143) = 22.14, p < .001, \eta_p^2 = .13, BF_{inc} = 9.11 \times 10^3$. The effect of font size on confidence was significant in the belief group (large $M = 4.99, SD = 0.93$; small $M = 4.19, SD = 1.01$), $t(74) = 7.89, p < .001, d_{av} = 0.83, BF = 4.31 \times 10^8$, and not significant in the no-belief group (large $M = 4.25, SD = 1.20$; small $M = 4.08, SD = 1.15$), $t(69) = 1.79, p = .079, d_{av} = 0.14, BF = 0.59$. The difference in confidence for answers large and small was higher in the belief group than in the no-belief group, $t(143) = 4.70, p < .001, d_{av} = 0.78, BF = 2.75 \times 10^3$.

For the recognition test, a 2 (belief: yes, no) x 2 (font size: small, large) also showed an interaction, $F(1, 143) = 6.18, p = .014, \eta_p^2 = .04, BF_{inc} = 3.11$, with similar results. The effect of font size on confidence was significant in the belief group (large $M = 5.08, SD = 0.86$; small $M = 4.85, SD = 0.88$), $t(74) = 5.41, p < .001, d_{av} = 0.27, BF = 2.00 \times 10^4$, and not significant in the no belief group (large $M = 4.99, SD = 1.02$; small $M = 4.91, SD = 1.07$), $t(69) = 1.77, p = .081, d_{av} = 0.08, BF = 0.58$. Also, the difference in confidence between large and small words was higher in the belief group than in the no-belief group, $t(143) = 2.49, p = .014, d_{av} = 0.41, BF = 2.93$.

**Discussion**
The results from Experiment 2 were consistent with those from Experiment 1: Font size influenced confidence in both the font-size and the recognition tests. The results also suggest that beliefs play a role in the effect, in line with past research with JOLs (Luna, Nogueira, & Albuquerque, 2019). When we introduced the counter-belief that words in small font size were recalled better, the effect of font size on confidence was smaller, although, in general, the effect of the counter-belief was small. Also, the results from the exploratory analyses showed that only participants who believe that words in large font are better remembered rate words in large font with higher confidence. The effect vanishes when there is no such belief.

In Experiment 2, the proportion of participants believing that words in large font are better remembered was numerically lower (55% and 47%) than in Experiment 1 (74%). This may be due to an unexpected reactivity to the information in Experiment 2 about the effect of font size on memory. That information may have drawn attention to font size and made participants reflect on its likely impact on memory. Upon reflection, some participants may have concluded that font size does not affect memory. To our knowledge, this is the first time that reactivity effects to instructions manipulating metamemory beliefs have been reported. Future research should try to replicate and study that effect.

**General Discussion**

This research tested whether metamemory beliefs, known to affect prospective judgments, also affect retrospective judgments. In particular, we tested whether the belief that words in larger font size are better remembered also affected confidence. The main results were that (1) font size affected confidence and that (2) the effect
depended on participants’ beliefs. We discuss each result in turn and then comment on memory results.

This research was the first to comprehensively study the effect of font size on confidence. Metamemory judgments are inferential and based on cues, and we should not take for granted that a cue affecting one type of judgment (e.g., JOL) would affect another (e.g., confidence). Similarly, there is no good theoretical reason to assume that cue utilisation and cue validity are similar between prospective and retrospective metamemory judgments. In support of differences between prospective and retrospective judgments, McDonough et al. (2021) found that monitoring at encoding and retrieval may be different. Despite these arguments, we found that font size affects both JOLs and confidence. This should not be interpreted as supporting that similar cues, in general, affect both judgments, but instead that the specific manipulation here, i.e., font size, is used as a cue to both JOLs and confidence. Other cues may have different effects, and we join McDonough et al. (2021) in suggesting that future research should include both prospective and retrospective metamemory measures to better understand the monitoring process(es).

More relevant, this research showed that beliefs also affect confidence. Participants who did not believe that large fonts were better remembered did not rate words in large font size with higher confidence than words in small font size. Similarly, the presentation of a counter-belief seemingly reduced the effect of font size on confidence, although this effect was not particularly clear because, in general, the counter-belief had a small effect on confidence. These results suggest that the debate on the role of theory-based and experience-based factors should be extended and encompass retrospective judgments. We make no claims on the contribution of
experiential factors in the font-size effect on confidence because the experiments were not designed to test them (but see below for some speculation). Instead, we would like to stress the effect of beliefs, which suggests that theory-based factors are transversal to both JOLs and confidence. Previous studies seeking an effect of font size on confidence were not sensitive enough to find it, either because the belief may not have been active at retrieval or because the font size at study may not have been retrieved and, thus, ignored as a cue for the judgment. All these factors should be taken into account in future research studying the effect of theory-based factors on confidence.

Although we did not study nor measured experiential cues, some results may give us a first insight into the weights of beliefs and experiential factors in our experiments. In the font-size test, participants identified better words in small font size (i.e., answer “small”) but rated their answers “large” with higher confidence. These results suggest that, in this particular research setting, experiential factors during retrieval may have been less relevant as cues for confidence than metamemory beliefs. However, the counter-belief in Experiment 2 had a clear effect on JOLs, a barely noticeable effect on the font-size test, and no effect on recognition, which suggests a higher role of beliefs on JOLs and a lesser role on confidence. These results are apparently contradictory, but are consistent with the idea that the belief must be active to influence metamemory judgments. Participants read the counter-belief before the study phase and again before the font-size test, but not before the recognition test. By then, most participants may have forgotten about the counter-belief (which may also show in the small effect of the counter-belief in the final question about beliefs). Our conclusion, largely speculative, is that when beliefs are active, they may have a stronger role in the font-size effect on confidence than experiential factors, but future research should test this idea.
Memory was secondary to this research, and as font size usually has a small effect on memory, we expected no effect of font size on memory. However, we consistently found better accuracy with large words in the recognition test. Past meta-analyses have shown that font size also affects memory (Chang & Brainerd, in press; Luna et al., 2018), but it seems that the effect of font size on memory is no longer part of the meta-analysis realm. Several individual experiments have reported similar effects (e.g., Halamish, 2018; Undorf & Zimdahl, 2019). Thus, maybe it is time to reconsider the repeated claim that font size does not affect memory and start acknowledging that it has a consistent effect when statistical power is large enough.

In a similar line, we did not expect font size to affect correct identifications in the font-size test. However, we found better memory for words presented in small than large font. We have two tentative explanations for this result. First, it may be due to participants remembering better the font size for small words, with particular mechanisms yet to be identified. And, second, it may be the consequence of some compensatory strategy. If participants cannot retrieve that a word was presented in large font, they could consider that they should have remembered the font size for a large word because large words are better remembered. Then, they could turn and answer “small”. Similar compensatory mechanisms have been proposed before in related research (e.g., see the distinctiveness heuristic in McDonough & Gallo, 2012). In support of this interpretation, participants reported more answers “small” than “large” (Experiment 1: 56% vs. 44%; Experiment 2: 57% vs. 43%). These explanations are highly speculative and should be tested in future research.

In sum, this research found that beliefs affect confidence in a similar way to JOLs. This may be specific to the font size manipulation, but it suggests that research
should pay more attention to theory-based factors in retrospective metamemory judgments.

References


JASP Team (2018). JASP (Version 0.9) [Computer software]. Available at: https://jasp-stats.org


METAMEMORY BELIEFS AND CONFIDENCE


Figure 1.

**Study phase**
- 500ms
- 4s
- No time limit
- 32 large font size (100pt)
- 32 small font size (18pt)
- 8 buffer

**Test phase**

**Font-Size Test**
- 500ms
- No time limit
- x32 studied words:
  - 16 studied large
  - 16 studied small

**Recognition Test**
- No time limit
- Different studied items at font-size and recognition tests

- Virtue yes/no confidence 1...7
- Banana yes/no confidence 1...7
- 32 studied
- 16 studied large
- 16 studied small
- 32 new
- 16 new large
- 16 new small