



# Is the perceptual disfluency effect moderated by working memory capacity? Direct replication of Lehmann et al. (2016)

Sophia C. Weissgerber<sup>1</sup> · Denia Indah Permatasari Terhorst<sup>1</sup> · Ralf Rummer<sup>1</sup>

Received: 25 January 2022 / Accepted: 21 November 2023 / Published online: 30 December 2023  
© The Author(s) 2023

## Abstract

According to an aptitude-treatment interaction experiment (Lehmann et al., *Metacognition and Learning*, 11, 89–105, 2016,  $N=47$ , published in *Metacognition and Learning*), perceptually disfluent texts facilitated retention and comprehension performance (but not transfer performance) only for learners with higher working memory capacity (WMC). No effects of WMC for a fluent text were found (albeit theoretically, fluency may be more advantageous for learners with lower WMC). The findings of our (pre-registered) direct replication (supervised online sample of  $N=96$ ) show a substantial deviation from the original results: In contrast to the interaction effect (disfluency and WMC) of the primary study, we obtained null results for disfluency, WMC, and their interaction for all learning outcomes. Our replication data are not indicative of WMC as a boundary condition moderating the disfluency effect on learning. We discuss discrepancies in the results of the primary study and our direct replication regarding particular methodological and analytical decisions, questioning the robustness and generalizability of Lehman et al.'s results beyond their primary study.

**Keywords** Disfluency effect · Learning with expository texts · Working memory capacity (WMC) · Direct replication · Aptitude-treatment interaction (ATI)

The perceptual disfluency effect—improving learning performance by harder-to-read fonts in text materials (Diemand-Yauman et al., 2011)—has the potential for easy and cost-efficient applicability in educational settings. It may seem counterintuitive to (perceptually) increase reading difficulties during learning. However, the theoretical rationale of the perceptual disfluency effect stems from research on reasoning (Alter & Oppenheimer, 2009; Alter et al., 2007) and is rooted in a switch to deeper cognitive processing from shallower processing (from Type 1 to Type 2 processing; cf. Evans & Stanovich, 2013). Such deeper

---

✉ Sophia C. Weissgerber  
scweissgerber@uni-kassel.de

✉ Ralf Rummer  
rummer@uni-kassel.de

Denia Indah Permatasari Terhorst  
denia.terhorst@gmail.com

<sup>1</sup> Institute of Psychology, University of Kassel, Kassel, Germany

processing is mentally more effortful and resonates with the concept of desirable difficulties (Bjork, 1994). Desirable difficulties are empirically well-supported learning conditions (e.g., Dunlosky et al., 2013). They systematically increase difficulties in the learning process to change cognitive processing and thereby facilitate (long-term) learning gains. As such, introducing perceptual difficulties with harder-to-read fonts is argued to benefit learning through deeper processing (Diemand-Yauman et al., 2011).

The (perceptual) disfluency effect—whether an increased perceptual difficulty of harder-to-read fonts is desirable for learning, reasoning, and memory—has been widely disputed (e.g., Dunlosky & Mueller, 2016; Eitel & Köhl, 2016). Difficulties in replicating the disfluency effect led researchers to question its mere existence or at least its generalizability and effect size (e.g., Ebersbach et al., 2023; Eitel et al., 2014; Geller et al., 2018; Rummer et al., 2016; Weissgerber & Reinhard, 2017). A recently published meta-analysis (Xie et al., 2018) reported a null effect ( $d=0.01$ ), thus questioning the study's theoretical underpinning and practical applicability because there were no systematic differences in learning outcomes between fluent and disfluent fonts (see also Weissgerber et al., 2021).

As an alternative to questioning the universality of the disfluency effect, the search for moderators of the disfluency effect (cf. Köhl et al., 2014) is rooted in the idea that boundary conditions are of critical importance. Given that some studies found a positive disfluency effect (e.g., Weltman & Eakin, 2014), but other studies found no disfluency effect (e.g., Rummer et al., 2016) or even negative disfluency effects (e.g., Geller et al., 2018), various moderators have been proposed and tested. These moderators encompass learning material characteristics (e.g., Magreehan et al., 2016; Sidi et al., 2016), disfluency manipulation (e.g., Ebersbach et al., 2023; Rummer et al., 2016; Seufert et al., 2016; Weissgerber & Reinhard, 2017) and learner characteristics (e.g., Eitel & Köhl, 2016; French et al., 2013; Geller & Peterson, 2021; Köhl et al. 2014; Lehmann et al., 2016). Learners' characteristics are particularly interesting to examine as a boundary condition because the occurrence of the disfluency effect may be tied to the interaction between learners' aptitudes and the learning treatment administered.

Based on this concept of *aptitude-treatment interaction* (ATI; Snow, 1989), Lehmann et al. (2016) were the first to theoretically identify working memory capacity (WMC) as a key aptitude that could determine whether perceptual disfluency enhances learning performance. Their ATI experiment tested whether WMC (the amount of information that can be successfully held in mind, attended to, or maintained in a rapidly accessible state at a time, e.g., Cowan, 2016) is a crucial moderator for the disfluency effect. Working memory is a limited cognitive resource (e.g., Baddeley, 2000; Cowan, 2001; for an overview see Schweppe & Rummer, 2014) that can be assessed by tests that relate to central-executive functioning, like an operation-span task (e.g., Oberauer et al., 2000). The theoretical value of the paper of Lehmann et al. (2016) lies in the consideration of cognitive demands by perceptual disfluency (Sweller et al., 2011) and learners' available cognitive resources related to the deeper processing hypothesis of the disfluency literature.

Lehmann et al.'s reasoning was based on the idea that disfluent texts create additional load on working memory and, therefore, only participants with sufficient WMC may profit from these additional strains. Additional demands, like perceptually processing the text, are characterized as a particular type of cognitive load in Cognitive Load Theory (Sweller et al., 2011). This type is called extraneous cognitive load (ECL) because the learner needs to expend cognitive resources that are not directed to mastering the learning task. Cognitive resources that are directed to learning itself, for example, to deeper understanding and processing of the learning material, are distinguished and characterized as germane cognitive load (GCL). Accordingly, Lehmann et al. argued (p. 92 ff.):

Additional demands on working memory caused by disfluent texts are possibly just usable by learners with a high WMC. Only learners with a high WMC might have enough capacity for the higher ECL caused by less legible texts and could engage in deeper processing and a more analytic and elaborative thinking rather than a heuristic and intuitive reasoning. WMC may work as an enhancer: The instructional strategy of using disfluent text is only effective with sufficient WMC. By contrast, learners with a low WMC should not be able to handle a higher ECL caused by disfluent material. Instead, the increased ECL will exceed the resources available, and learners cannot allocate germane resources to the learning process. Thus, the construction of a situational model is hindered. Overall, disfluency should not be beneficial for those learners. In the case of learning with fluent material, ECL should not be increased which would result in learners with high as well as low WMC having similar learning outcomes.

To test this interaction hypothesis, Lehmann et al. set up the following experiment. Before learning the materials, participants' previous knowledge was assessed with a self-developed test. Participants then learned (self-paced) a printed text on the topic of "Time and date differences on earth" adapted from Schnotz and Bannert (1999). Half of the participants had received the disfluent text and the other half the fluent text (between-subjects). A disfluent display of the text was achieved by a grey-scaled Haettenschweiler font (size 12 pt), whereas the fluent display used a black Arial font (size 12 pt). After learning, participants answered items to assess potential confounding variables (task enjoyment / interest / motivation) and items to assess cognitive load (Klepsch & Seufert, 2012). This was followed by the performance test (without time limit) of participants' retention, comprehension, and transfer, after which participants again filled out the items regarding confounding variables. Working memory had been measured previous to the experiment and was later linked to the data. After filtering their raw sample of  $N=65$  for medium previous knowledge, their final sample consisted of 47 participants. The sample was comprised mainly of young ( $M_{age}=22.9$ ) female (85%) students.

The results (mostly) confirmed Lehmann et al.'s interaction hypothesis (see pages 98 ff.) regarding retention ( $\beta=11.66$ ,  $t(44)=2.46$ ,  $p=0.02$ ) and comprehension ( $\beta=12.62$ ,  $t(46)=2.33$ ,  $p=0.03$ ) but not transfer ( $\beta=-2.27$ ,  $t(45)=-0.23$ , n.s.). In the disfluent condition, an effect of WMC on retention and comprehension was found: Higher WMC resulted in higher retention ( $\beta=10.66$ ,  $t(44)=2.86$ ,  $p=0.01$ ) and comprehension scores ( $\beta=10.66$ ,  $t(46)=2.86$ ,  $p<0.05$ ). By contrast, WMC did not influence learning outcomes in the fluent condition for retention ( $\beta=-1.01$ ,  $t(44)=-0.35$ , n.s.) or for comprehension ( $\beta=-3.90$ ,  $t(46)=-1.20$ , n.s.). In fact, working memory had no effect overall on learning outcomes.

Lehmann et al.'s (2016) study introduced a plausible explanation for the inconsistent appearance of the perceptual disfluency effect. Their key hypothesis is that only participants with high working memory spans should benefit from hard-to-read texts, whereas participants with low memory spans should not. The data support this hypothesis, at least partially.

In sum, for high-span subjects, learning outcomes were higher with disfluent than with fluent texts. In addition, in the disfluent condition, learning outcomes were higher with disfluent texts than with fluent texts. These findings fit Lehmann et al.'s crucial hypothesis. Unexpectedly, however, low-span subjects performed better than high-span subjects in the fluent condition, a result that is rather implausible and questions a straightforward

interpretation of the critical two-way interaction. In addition, for low-span subjects, there was no difference between the two experimental conditions, which is at least somewhat surprising.

There are also numerous methodological limitations. The first is that the experiment was underpowered. The final sample on which their interaction effects were based consisted of  $N=47$  participants in a between-subjects design. Based on a simulation according to Finsaas et al. (2021) and Baranger et al., (2022), Lehmann et al. (2016) had around 45% power to detect an interaction effect with the interaction effect size obtained in the primary study. The post hoc power for their regression models, including the interaction and respective control variables, ranged from 0.62 to 0.69. See the Method section of this paper for details on power analysis.

The design of the experiment and analysis of the data included uncommon methodological and analytical decisions. All relevant individual variables of the participants, including WMC, were already measured in a separate prior study. The WMC data were therefore already obtained and then linked to participants who took part in the new experiment. Moreover, the items of prior knowledge were first filtered based on solution probabilities, with the consequence of eliminating three of the six items to measure previous knowledge. The remaining three items comprised the previous knowledge score, which served as both a selection criterion for including participants and a control variable in the analysis. This is noteworthy when considering that the test for prior knowledge was self-developed and had not been validated previously.

Although there are no clear standards about how to analytically use and change variables (including previous knowledge), the “if” of inclusion and the “how-to” naturally entail many degrees of freedom and subjective decisions. For example, the filtering of solution probabilities between 10% and 90% was a subjective decision, as was the selection of participants with medium previous knowledge between 25% and 75%. Most importantly, results will vary as a consequence of these degrees of freedom and the specific subjective choices. This can make many patterns of results possible (cf. the garden of forking paths: Gelman & Loken, 2019), especially when not preregistered, even when no misuse is intended (Wicherts et al., 2016). For example, the interaction effect for comprehension is rendered nonsignificant when removing previous knowledge as a control variable in the model. The  $p$  value of the interaction changes from 0.025 to 0.075. Accordingly, the working memory effect in the disfluent condition is no longer significant when previous knowledge is removed. The  $p$  value changes from  $p=0.05$  to  $p=0.13$ . The interaction effect of retention and the lack of an interaction effect for transfer are robust despite deciding to include or exclude previous knowledge.

We would like to clarify that we do not imply that Lehmann et al. (2016) misused previous knowledge. Lehmann et al. had valid methodological and theoretical reasons to select participants with medium previous knowledge. However, the specific decisions regarding item probability and the use of control variables might have resulted from subjective choices due to degrees of freedom during the data analysis. Such subjective choices, when made during data analysis, can compound the issue of being overly tailored to the current dataset, making the model highly relevant to that particular dataset but misfit for other datasets. In conclusion, apart from being underpowered, these aspects raise questions about the generalizability of the results beyond the primary study.

This lack of generalizability was also demonstrated by two experiments (Miyakawa & Hattori, 2017; Strukelj et al., 2016) that also examined the role of WMC on the disfluency effect and showed discrepant results to those of Lehmann et al. (2016). Strukelj et al. (2016) could not find any moderating effect of WMC on disfluency, and Miyakawa and

Hattori (2017) showed that participants with *lower* WMC performed better in the disfluency condition (for participants with higher WMC, performance was similar in the fluent and disfluent conditions). The operationalizations in WMC, disfluency, and outcome measures in these experiments vary, and the sample sizes were small. This complicates the comparability of the results and the evaluation of the interaction hypothesis (an in-depth consideration can be found in the Discussion section of this paper). Consequently, a direct replication is necessary to evaluate the robustness and generalizability of the findings by Lehmann et al. and thus of the interaction hypothesis.

## Motivation and Aim of the Present Replication

Lehmann et al. deserve recognition for their consideration of the role of working memory regarding the disfluency effect. In their paper, they argued from an ATI perspective that optimal learning with disfluent texts can only occur when additional cognitive demands by harder-to-read fonts are met by sufficient WMC. Their consideration of the disfluency effect has practical relevance. The original disfluency hypothesis postulates a universal effect. If disfluency effects are indeed universal, no boundary conditions and thus no interaction effects are expected. If, however, disfluency effects are not universal, then they should be tied to boundary conditions. Applying disfluency to study materials is then either to be recommended in general or only for a particular subgroup of learners (The theoretical insights and findings of Lehman et al. suggest an application only to learners with higher WMC).

If the perceptual disfluency effect has a null effect in a meta-analysis (Xie et al., 2018), then why does it merit any additional research? Assuming that the disfluency effect indeed does not exist, one practical consequence would be that neither moderated effects of disfluency nor universal effects are to be expected. Reliably finding either interactions or main effects (i.e., replicating interactions or main effects) questions the nonexistence of the effect. Failure to find convincing evidence for any disfluency effect (moderated or unmoderated) further reinforces the position that the perceptual disfluency effect is fickle. The practical consequence is to not recommend the use of disfluency.

The moderation hypothesis of disfluency by WMC is crucial. Yet, to our knowledge, few experiments exist that test this important interaction (Lehmann et al., 2016; Miyakawa & Hattori, 2017; Strukelj et al., 2016), and the results of them are inconsistent. Considering the divergent results in the disfluency literature at large and considering that the study by Lehmann et al. is the only existing study demonstrating the proposed relation between WMC and the disfluency effect, a direct replication of Lehmann et al. is called for to bolster their primary findings.

The aim of the current direct replication was, therefore, to replicate the results of the primary study based on a larger sample to verify the robustness of the findings. To this end, we doubled the final sample size. We used the same learning materials, working memory measurement, outcome measures, and control variables as in Lehmann et al. (2016). We followed the same analytical approach and applied it to our data. This means to analyze the replication data, we used the criteria and analysis set out by Lehmann et al., and we filtered the solution probabilities between 10% and 90% to select the eligible previous knowledge items. We also applied the criteria of selecting participants with medium previous knowledge levels between 25% and 75% and selected significant control variables in our models by correlating the control variables with the outcome measures and testing for differences

across experimental conditions. Lastly, the replication attempt was evaluated as successful if the following criterion was met: a significant interaction effect of WMC and disfluency, which is in the same direction as in the primary study.

## Method

### Participants

Participants were recruited through the university's research participation system, various Facebook groups, and online forums; they were compensated either by course credits or participation in a lottery to win coupons for an online bookstore. Participants were randomly assigned to the experimental conditions (double blind through automation within the software) in the online experiment (only computers allowed, no tablets or smartphones) and to either the disfluent learning condition or the fluent learning condition.

A total of 137 participants completed the experiment. The preregistered (AsPredicted #45,573; <https://osf.io/54vqc/><sup>1</sup>) stop rule for collecting participants was set to a data collection period until 27 August 2020 or until 200 participants were reached, whichever came first. We planned to increase power relative to the primary study ( $N=47$ ), given that primary findings are often less sizable (e.g., Nosek et al., 2015; Camerer et al., 2018).

Exclusion of participants followed the preregistration (<https://osf.io/54vqc/>), except that we decided to include participants with dyslexia, as the primary study of Lehmann et al. (2016) did not report excluding these participants in the analysis. We had to exclude seven participants whose answers to some of the test questions seemed similar to the text displayed on Wikipedia or Google. This additional criterion was applied to account for unauthentic answers given the online setup of the study. (Note, participants were supervised by an experimenter throughout the online data collection.)

In line with Lehmann et al. (2016) and our preregistration (<https://osf.io/54vqc/>), we excluded participants with less than 25% of the maximum test score and participants with more than 75% of the maximum test score in the test for prior knowledge, including only participants with medium knowledge. Accordingly, we excluded 34 participants who scored too low (less than 25% of maximum score, i.e., lower than 2 of maximum 8 points) or too high (more than 75% of maximum score, i.e., higher than 6 of maximum 8 points) in the test about prior knowledge (cf. Lehmann et al., 2016:  $n=18$ ). Thus, our final sample included 96 participants (76% female) with a medium level of prior knowledge (cf. Lehmann et al., 2016:  $n=47$ ; 85% female), which exceeded the minimum required number of participants ( $N=86$ ) for the regression model of retention with four predictors according to a power analysis conducted a priori (see below and see <https://osf.io/6mzt7/>). The distribution of included participants across experimental conditions was  $n_{fluent}=47$  and  $n_{disfluent}=49$  (cf. Lehmann et al., 2016:  $n_{fluent}=24$ ,  $n_{disfluent}=23$ ). For further characteristics of our replication sample, please see Table 1.

<sup>1</sup> We would like to add that we included the target sample size and stop rule for data collection in the preregistration but not in the power analysis. Also, we followed the analytical approach outlined in the original study and specified our covariates in the preregistration under point 8. Per default, our inclusion criteria for participants and covariates matched the inclusion criteria of the original study, applied to our data.

**Table 1** Comparison of Demographic Characteristics: Primary and Replication Study

Characteristic	Lehmann et al. (2016)		Direct Replication	
	n	%	n	%
Total participants	47	100	96	100
Gender				
Female	40	85	76	79
Male	7	15	20	21
Distribution Disfluency Manipulation				
Fluent	24	51	49	51
Disfluent	23	49	47	49
Occupation				
Student	47	100	62	65
Non-Student	-	-	34	35
German As Native Language				
German Native Speaker	Unknown	77	80	
Non-German Native Speaker		19	20	
Age	Value			
Mean	22.9	28.8		
SD	3.77	12.3		

**Power Analysis** We conducted an a-priori power analysis using G-Power 3.1.9.7 based on the effect of the overall model. In the primary study, the regression model of retention included four predictors (disfluency, working memory, their interaction, motivation during assessment) and yielded an effect size of  $R^2=0.19$ . We calculated the effect size ( $f^2$ ) using the following formula:  $f^2 = \frac{R^2}{1-R^2} = \frac{.19}{.81} = 0.23$ . With  $\alpha=0.05$  and  $1-\beta=0.95$ , the power analysis yielded 86 participants (57 participants for conventional power of 0.80).

**Simulation of Power and Sensitivity for the Interaction Effect of Lehmann et al. (2016)** Our simulation was based on free-to-use tools (R package InteractionPower and shiny app) by Baranger et al. (2022) and Finsaas et al. (2021). The tools enable power analyses for regression models in cross-sectional data sets, specifically focusing on interactions between two variables (moderation analyses). The package and app take variable correlations, reliability, and variable distributions into account to derive simulation-based power estimates.

The necessary correlation values of X1-Y, X2-Y, X1-X2 for the simulation were computed based on the primary study's dataset (regression model with retention as dependent variable and as predictors disfluency, standardized WMC, their interaction effect, and standardized motivation during assessing learning outcomes). The reliability (Cronbach's Alpha) of the retention test was 0.492 based on the five items for retention of the primary data. We assumed reliability for X1 and X2 to be 0.90. In the primary study, two participants had no score for motivation during assessing learning outcomes, decreasing the inputted sample size from 47 to 45. Given that the X1X2-Y from the primary data was  $r=0.386$ , the simulation shows that the primary study had around 45% power to detect an interaction effect of this size. At conventional levels of 0.80 power, the sample of 47 participants of Lehmann et al. (2016) would have been sufficient to detect an interaction effect with X1X2-Y of  $r>0.6$ , see <https://osf.io/6mzt7/>.



**Simulation of Necessary Sample Size for the Interaction Effect of Retention** All input values were the same as in the previous section (see above), except we varied inputs of the sample size from 40 to 200 participants to simulate achieved power (simulation based on Baranger et al., 2022; Finsaas et al., 2021). Given the effect sizes of Lehman et al.'s data, 150 participants would be necessary to achieve conventional power of 0.80.

## Materials

Based on the original material of Lehmann et al.,<sup>2</sup> we created predefined solutions using the information available in the learning material as a base for our coding scheme of participant performance (see <https://osf.io/u9jpf/>), mirroring the scoring of Lehmann et al.. All other materials utilized in the replication are identical to the primary study. The experiment was created using SoSciSurvey (Leiner, 2019) and was made available to users via <https://www.soscisurvey.de> (the programming file can be found here: <https://osf.io/ye547/>).

## Prior Knowledge Test

Prior knowledge was measured by the same six open questions on “Time and date differences on earth” used by Lehmann et al.. The six questions (<https://osf.io/97ehv/>) encompassed three categories parallel to the types of learning outcome (retention, comprehension, and transfer) with two questions measuring each type. The prior knowledge test included a table listing eight different cities all over the world and their time difference to Greenwich. Each answer to the questions was compared to our predefined solution based on the text information (<https://osf.io/u9jpf/>).

Concerning the prior knowledge test, Lehmann et al. (2016) applied a selection criterion of permissible items to have a solution probability of more than 10% or less than 90%. The application of this criterion to our replication sample (contingent on our data) led to an exclusion of two items (Item 2 and Item 3) for prior knowledge with solution probabilities under 10% and 90%. By contrast, Lehmann et al. excluded three of six items from further analysis based on this criterion and their data (Item 1, Item 2, and Item 3). Thus, the maximum score for previous knowledge is not 12 points, but actually 8 points for overall prior knowledge for the four included items (2 points for each question) according to the criteria of Lehmann et al. applied to our sample. For means and standard deviations of previous knowledge, see Table 1. The code for items selection based on the 10% to 90% criterion can be found here: <https://osf.io/ed9b3/>.

## Learning Material

Utilizing the same learning materials as in Lehmann et al. (2016), the text-based learning material revolved around the topic of “Time and date differences on earth” (adapted from Schnotz & Bannert, 1999). The text consisted of 1070 words and included the same table previously shown in the prior knowledge test. For the fluent condition, the text was written in Arial, 12pt, black (see <https://osf.io/5rp2j/>); for the disfluent condition, Haettenschweiler, 12pt, grayscale 35% was used (<https://osf.io/fyh3s/>). Exemplary text is displayed in Fig. 1.

<sup>2</sup> We would like to thank Tina Seufert for providing all the materials used by Lehmann et al..



Fluent (Arial, 12pt, black)	Disfluent (Haettenschweiler, 12pt, grayscale 35%)
For the purpose of ship- and airplane-navigation, the earth can usually be viewed as a sphere without significant errors. As a convention, the surface of said sphere was divided into time zones using so-called meridians.	For the purpose of ship- and airplane-navigation, the earth can usually be viewed as a sphere without significant errors. As a convention, the surface of said sphere was divided into time zones using so-called meridians.

**Fig. 1** Examples of the Text Font Manipulation of the Learning Material. Note. The exemplary text passage, which was originally in German, was translated into English for this figure

## Learning Outcomes

The same criterion test for learning outcomes as in Lehmann et al. (2016) was used in the present experiment. This final test addressed three different categories: retention, comprehension, and transfer with each category consisting of five open questions. For the comprehension and transfer category, a table identical to the table of the prior knowledge test was presented in addition to the open questions. Similar to the prior knowledge test, the answers in each category were compared to a predefined solution with a maximum score of 2 for each item (for the items see <https://osf.io/vw8g2/> and for scoring, see: <https://osf.io/u9jpf/>). The five items with maximum of 2 points per correct answer led to a maximum score of 10 points for each category. All three categories together thus totaled a maximum of 30 points. The code for the computation of the performance results is available here: <https://osf.io/ed9b3/>

## Working Memory Capacity

We utilized the same test to measure working memory as Lehmann et al. (2016) based on the Numerical Updating Memory subtest of the WMC test by Oberauer et al. (2000). In this test, participants were presented with a chessboard-like grid of fields. Participants were instructed to memorize numbers (ranging from one to nine) that were shown for 1000 ms consecutively in each of these fields. Subsequently, four arrows pointing upward or downward were presented in these fields, randomly one by one, for 1000 ms each.

Participants' task was to do simple mathematical calculations to the number in the field previously memorized when arrows were shown. Arrows pointing downward prompted to subtract 1 point from the number, and arrows pointing upward indicated to add 1 point to the number. After the last arrow was presented, a question mark followed by an on-screen keyboard in which participants had to fill in the updated number of the promoted fields, appearing randomly on each of the fields. Short feedback was then given to let participants know whether their answer was correct.

Overall, the test consisted of seven levels with an increasing number of fields being shown the higher the level was. The first level contained three fields, and nine fields were presented on the highest level. To advance to the next level, 75% of the fields that had been previously shown had to be answered correctly. When this was not achieved, the test ended, and the final score was calculated according to the level participants successfully solved.

As the test started with three fields being presented, participants could achieve a minimum score of 2 (first level not solved) and a maximum score of 9 (final level solved). Note, we implemented a minor procedural deviation before starting the working memory test: In contrast to Lehman et al., who immediately started the span test after participants had read the instruction, our participants received two practice trials with one field, and when this field was answered correctly, then two fields. The code for computing the working memory scores is available here: <https://osf.io/ed9b3/>.

### Potential Confounding Variables

*Enjoyment* during task performance, *interest* in the task, and *motivation* to solve the task were measured using the items developed by Lehmann et al. (2016). The question “How much did you enjoy the task performed?” with a 7-point scale ranging from 1 (*very little*) to 7 (*very much*) was used to assess *enjoyment* during task performance. *Interest* in the task and *motivation* to solve the task were measured through the items “I was interested in the task” and “I was motivated to solve the task,” respectively, on a 7-point scale ranging from 1 (*absolutely not true*) to 7 (*completely true*). These control variables were correlated with the outcome variables and tested for differences across experimental conditions, and in case of a significant result, they were controlled for in the regression model (see <https://osf.io/kgxr9/>). This approach followed the criteria of Lehmann et al.

**Further variables** Following the same procedures as the primary study, the replication also assessed cognitive load (Sweller, 1994) by means of the Cognitive Load Questionnaire (Klepsch & Seufert, 2012). These variables, however, were not relevant for testing the target effect of WMC and disfluency and are therefore not reported in the results section.

Based on the study of Eitel and Kühn (2016), which showed reading comprehension being a covariate of retention performance, the replication study included reading comprehension and reading speed in the first session (otherwise only a 10 min WMC test would have been administered, making it odd for participants to sign up for a two-session study; yet the two-session replication was necessary because Lehman et al. had obtained the WMC data separately at a different date). The inclusion of the reading comprehension test (LGVT 6–12; Schneider et al., 2007) thus had the advantage of mimicking the measurement of different individual aspects of participants in a separate session, similar to Lehmann et al.

**Demographic Questions** The demographic questions consisted of gender, age, country of residence, professional status, whether participants are German native speakers, and whether participants have dyslexia. Participants were also given the chance to provide feedback about the study at the end of the experiment.

### Procedure

Concerning the procedure, the replication study was matched as closely as possible to the primary study. However, due to the COVID-19 situation at the time the study was conducted, a modification had to be made; specifically, the experiment was conducted online via video conference. The experimenter supervised multiple participants during online data collection. In the sessions, upon entering the video conference room, participants were given the link to the online experiment and asked to work on the study while staying in the video conference room. Hence, similar to a laboratory setting, the experimenter watched

participants. Moreover, participation was only possible on a computer—that is, anyone trying to access the experiment via tablet or smartphone was technically denied access by our programming.

As mentioned, the primary study measured WMC separately in a prior study. The replication study was therefore divided into two sessions to simulate these conditions; an example of Session 1 and Session 2 in German can be found here: <https://osf.io/at7ey/>. The primary study did not provide information about the time interval between the measurement of WMC and data collection in the disfluency study. To match this procedure as closely as possible, the replication study conducted the first and second sessions on different days, allowing participants to choose the dates of the experiments themselves, with the first session always being before the second session.

The first session lasted about 25 min. The session was conducted as follows: Participants were first asked about their consent, followed by creating a personal code used to match their data from the first and second sessions. Subsequently, the WMC test was conducted. In addition, and in contrast to the primary study, the reading comprehension and speed tests were carried out (but in the primary study, other concepts were measured<sup>3</sup>).

The second session took around 50 min and was similar to the primary study, except the demographic questions were presented at the end of the session instead of at the beginning. Participants were first asked to type in the code they already created in the first session followed by the prior knowledge test. Afterward, participants were asked to read a text about “Time and date differences on earth” (either fluent or disfluent). They were instructed to read it carefully and to take as long as they wanted, without time limits. After participants finished reading, the Cognitive Load Questionnaire and questions about enjoyment, interest, and motivation during learning were presented. Subsequently, five retention questions followed by five comprehension and five transfer questions were presented. Next, participants were asked to answer the Cognitive Load Questionnaire and the questions about enjoyment, interest, and motivation during answering the learning outcome tests. Finally, participants answered demographic questions, and at the end of the experiment, a short explanation of the study was shown.

## Results

### Data Preparation

The following analyses are based on the same analytical approach set out by the primary study. We adopted the same data preprocessing criteria and model inclusion criteria as Lehmann et al. (2016) and applied them to our replication sample. The raw data can be found here: <https://osf.io/59xt3/> and <https://osf.io/gvps6/>. The data preparation file—including filtering and participant selection according to the criteria of Lehmann et al.—can be found here: <https://osf.io/ed9b3/>. The preprocessed data set can be accessed here: <https://osf.io/qdbcv/>.

First, we applied the selection criterion of permissible items of the prior knowledge test to have a solution probability of more than 10% or less than 90%. We included four out of

---

<sup>3</sup> We are aware from personal communication that students’ personality characteristics are assessed in a battery at the beginning of their undergraduate studies and (if of interest) linked to various experiments over the course of their 3-year coursework term at the university where data collection took place.

six items (Items 1, 4, 5, and 6). Furthermore, we excluded participants with low previous knowledge scores (less than or equal to 2 out of 8 points, representing 25% of the maximum score). We further excluded participants with high prior knowledge scores (more than or equal to 6 out of 8 points, representing 75% of the maximum score in the prior knowledge test; see also section “[Participants](#)”).

## Learning Outcomes

The analysis code can be found here: <https://osf.io/kgxr9/>. All regression analyses were carried out using dummy-variable coding according to Aiken and West (1991) for the disfluency manipulation. The fluent condition was first coded as 0 and the disfluent condition as 1, followed by the same regressions with a reversed coding of the fluent condition as 1 and the disfluent condition as 0. Furthermore, dependent variables were converted into percentages, and all the predictors (except the disfluency manipulation) were standardized. The dependent variables and predictors were identical to the primary study. We followed the analytical approach of the primary study including selection of covariates (i.e., correlating the potential covariates outlined in Lehmann et al. with the learning outcomes; in case of a significance, they were included in the respective model). An overview of the potential covariates and learning outcomes of both the primary study and the replication study is provided in Table 2.

**Retention** Using retention as a dependent variable, the predictors in the regression model were disfluency manipulation (fluent = 0; disfluent = 1), standardized WMC, the interaction term of disfluency \* standardized WMC, and the standardized control variables enjoyment, interest, motivation during learning, and enjoyment, interest, motivation during the assessment of the learning outcomes.

The regression model was significant,  $F(9, 84) = 2.76$ ,  $p = 0.007$ ,  $R^2_{adj} = 0.15$ . Nevertheless, the only significant predictor of retention was motivation during learning ( $\beta = 7.32$ ,  $SE = 3.24$ ,  $t(84) = 2.26$ ,  $p = 0.026$ ). Disfluency manipulation ( $\beta = 0.82$ ,  $SE = 4.77$ ,  $t(84) = 0.17$ ,  $p = 0.864$ ), WMC (both in fluent ( $\beta = -0.07$ ,  $SE = 3.94$ ,  $t(84) = -0.17$ ,  $p = 0.865$ ; disfluent  $\beta = 3.52$ ,  $SE = 3.04$ ,  $t(84) = 1.16$ ,  $p = 0.25$  condition), and the interaction between disfluency manipulation \* WMC, ( $\beta = 4.19$ ,  $SE = 4.92$ ,  $t(84) = 0.85$ ,  $p = 0.397$ ), were not predictors of retention (see Fig. 2 and Table 3). Based on the criterion of significance and an effect in the same direction, these results do not replicate the primary results of Lehmann et al. (2016).

**Comprehension** Except the control variables, all predictors were the same as in the retention model. Thus, the comprehension model included the disfluency manipulation (fluent = 0; disfluent = 1), standardized WMC, the interaction term of disfluency \* standardized WMC, and the standardized control variables: enjoyment and interest during learning; enjoyment, interest, and motivation during assessment of the learning outcomes, as well as prior knowledge.

The regression model was significant ( $F(9, 85) = 3.17$ ,  $p = 0.002$ ,  $R^2_{adj} = 0.17$ ). The only significant predictor for comprehension was prior knowledge ( $\beta = 9.19$ ,  $SE = 2.81$ ,  $t(85) = 3.27$ ,  $p = 0.002$ ). Disfluency manipulation ( $\beta = 8.90$ ,  $SE = 5.54$ ,  $t(85) = 1.61$ ,  $p = 0.11$ ), WMC (fluent:  $\beta = 6.11$ ,  $SE = 4.60$ ,  $t(85) = 1.33$ ,  $p = 0.19$ ; disfluent:  $\beta = -0.67$ ,  $SE = 3.59$ ,  $t(85) = -0.19$ ,  $p = 0.85$ ), and their interaction ( $\beta = -6.78$ ,  $SE = 5.77$ ,  $t(85) = -1.18$ ,  $p = 0.243$ ) were not significant predictors of comprehension (see Fig. 3 and Table 4). Note, the slope of the interaction effect is negative in our replication data and positive in the

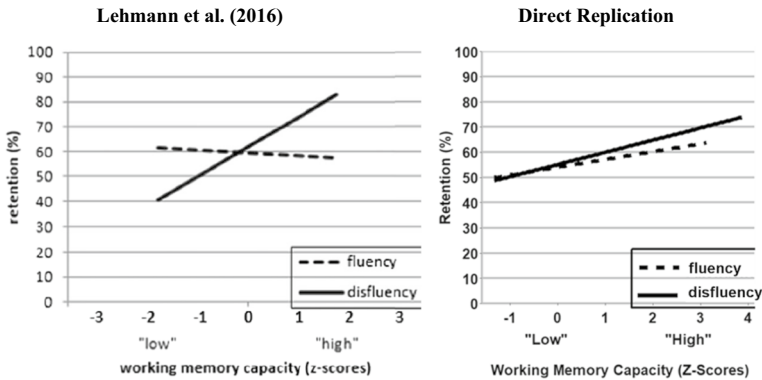
**Table 2** Comparison of Mean and SD Values of the Variables in the Primary Study and the Direct Replication

Variables	Fluent		Disfluent		Overall							
	Lehmann et al. (2016)		Lehmann et al. (2016)		Lehmann et al. (2016)		Lehmann et al. (2016)		Direct Replication			
	M	SD	M	SD	M	SD	M	SD				
Working Memory Capacity	3.96	1.23	<b>3.91</b>	<b>1.18</b>	4.09	0.95	<b>3.67</b>	<b>1.49</b>	4.02	1.09	<b>3.79</b>	<b>1.34</b>
<b>Learning Outcomes</b>												
Retention (%)	59.79	17.35	<b>54.3</b>	<b>23.3</b>	61.52	18.12	<b>54.7</b>	<b>26.2</b>	60.6	17.6	<b>54.5</b>	<b>24.7</b>
Comprehension (%)	73.54	22.63	<b>66.8</b>	<b>27.1</b>	71.96	15.28	<b>71.4</b>	<b>31.1</b>	72.8	19.2	<b>69.2</b>	<b>29.2</b>
Transfer (%)	53.59	29.76	<b>43.8</b>	<b>34.3</b>	49.78	32.88	<b>41.6</b>	<b>32.9</b>	51.7	31.1	<b>42.7</b>	<b>33.4</b>
<b>Potential Confounding Variables</b>												
Prior Knowledge	1.98 <sup>a</sup>	1.57 <sup>a</sup>	<b>3.89<sup>b</sup></b>	<b>1.51<sup>b</sup></b>	2.09 <sup>a</sup>	2.04 <sup>a</sup>	<b>3.73<sup>b</sup></b>	<b>1.51<sup>b</sup></b>	2.03 <sup>a</sup>	1.8 <sup>a</sup>	<b>3.81<sup>b</sup></b>	<b>1.50<sup>b</sup></b>
Enjoyment during learning	3.21	1.56	<b>3.53</b>	<b>1.89</b>	3.30	1.40	<b>3.24</b>	<b>1.53</b>	Unknown	Unknown	<b>3.39</b>	<b>1.71</b>
Interest during learning	3.83	1.76	<b>5.02</b>	<b>1.57</b>	4.09	1.70	<b>4.48</b>	<b>1.57</b>	Unknown	Unknown	<b>4.75</b>	<b>1.58</b>
Motivation during learning	4.75	1.29	<b>5.00</b>	<b>1.65</b>	4.70	1.58	<b>5.20</b>	<b>1.29</b>	Unknown	Unknown	<b>5.11</b>	<b>1.47</b>
Enjoyment during assessment	3.36	1.79	<b>3.96</b>	<b>1.98</b>	3.30	1.64	<b>3.55</b>	<b>1.81</b>	Unknown	Unknown	<b>3.75</b>	<b>1.90</b>
Interest during assessment	3.95	1.84	<b>4.85</b>	<b>1.84</b>	3.87	1.89	<b>4.33</b>	<b>1.76</b>	Unknown	Unknown	<b>4.58</b>	<b>1.81</b>
Motivation during assessment	5.27	1.52	<b>5.06</b>	<b>1.58</b>	4.70	1.79	<b>5.18</b>	<b>1.55</b>	Unknown	Unknown	<b>5.12</b>	<b>1.56</b>

Note. Total number of participants in Lehmann et al. (2016)  $N=47$ , fluent condition  $n=24$ , disfluent condition  $n=23$ . Total number of participants in the Direct Replication  $N=96$ , fluent condition  $n=47$ , disfluent condition  $n=49$

<sup>a</sup> Maximum Score = 6 (based on three prior knowledge items with solution probabilities between 10 and 90% out of six possible items)

<sup>b</sup> Maximum Score = 8. (based on four prior knowledge items with solution probabilities between 10 and 90% out of six possible items)



**Fig. 2** Retention: Interaction Working Memory Capacity and Disfluency Manipulation in the Primary Study and Direct Replication. Note.  $N=47$  in Lehmann et al. (2016).  $N=94$  in the replication study (R’s regression model excluded 2 persons due to missings)

**Table 3** Regression Analysis of Retention: Comparison of Primary Study and Replication Study

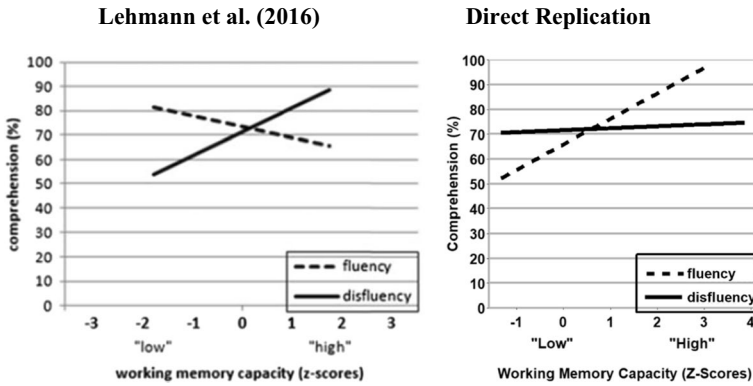
Effect	Lehmann et al. (2016)			Direct Replication		
	Estimate	SE <sup>a</sup>	t(44)	Estimate	SE	t(84)
Disfluency Manipulation	2.42	-	0.53	0.82	4.77	0.17
WMC: Fluent	-1.01	-	-0.35	-0.67	3.94	-0.17
WMC: Disfluent	10.66*	-	2.86	3.52	3.04	1.16
Disfluency * WMC	11.66*	-	2.46	4.19	4.92	0.85
Enjoyment during learning	-	-	-	-2.29	3.12	-0.73
Interest during learning	-	-	-	-2.84	3.95	-0.72
Motivation during learning	-	-	-	7.32	3.23	2.26*
Enjoyment during assessment	-	-	-	0.46	3.55	0.13
Interest during assessment	-	-	-	4.76	4.28	1.11
Motivation during assessment	5.59*	-	2.38	5.24	3.05	1.72
Regression Model	$F(4, 44)$	$R^2_{adj}$		$F(9, 84)$	$R^2_{adj}$	
	3.56*	0.19		2.76***	0.15	

Note.  $N=47$  for Lehmann et al. (2016).  $N=94$  for the replication study. SE=Standard Error

<sup>a</sup> Lehmann et al. (2016) did not report the standard errors of the coefficients (SE)

\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$

primary study. Thus, the direction of the interaction is reversed in the replication data (positive effect: Higher WMC in the fluent condition appears associated with higher comprehension performance) as opposed to the pattern in the primary study (negative effect: Higher WMC in the fluent condition related to lower comprehension performance). Moreover, our replication data showed no WMC effect in the disfluent condition, whereas the pattern in the primary study showed a positive WMC effect in the disfluent condition. In sum, the replication findings do not replicate the primary results, neither in terms of significance nor in terms of the direction of the effect.



**Fig. 3** Comprehension: Interaction between Working Memory Capacity and Disfluency Manipulation in the Primary and Replication Study. *Note.*  $N=47$  in Lehmann et al. (2016).  $N=95$  in the replication study (R’s regression model excluded 1 person due to missings)

**Table 4** Regression Analysis for Comprehension: Comparison between Primary Study and Replication Study

Effect	Lehmann et al. (2016)			Direct Replication		
	Estimate	SE <sup>a</sup>	<i>t</i> (46)	Estimate	SE	<i>t</i> (85)
Disfluency Manipulation	-2.36	-	-0.46	8.90	5.54	1.61
WMC: Fluent	-3.90	-	-1.20	6.10	4.60	1.33
WMC: Disfluent	10.66*	-	2.86	-0.67	3.57	-0.19
Disfluency * WMC	12.62*	-	2.33	-6.78	5.77	-1.18
Enjoyment during learning	-	-	-	2.44	3.56	0.68
Interest during learning	-	-	-	1.16	4.17	0.28
Enjoyment during assessment	-	-	-	0.14	4.13	0.03
Interest during assessment	-	-	-	3.69	4.70	0.78
Motivation during assessment	-	-	-	1.88	3.34	0.56
Prior Knowledge	7.90**	-	3.04	9.19	2.81	3.27**
Regression Model	$F(4, 46)$	$R^2_{adj}$		$F(9, 85)$	$R^2_{adj}$	
	3.33*	0.17		3.17**	0.17	

*Note.*  $N=47$  for Lehmann et al. (2016).  $N=95$  for the Replication Study (R’s regression model excluded 1 person due to missings). *SE* = Standard Error

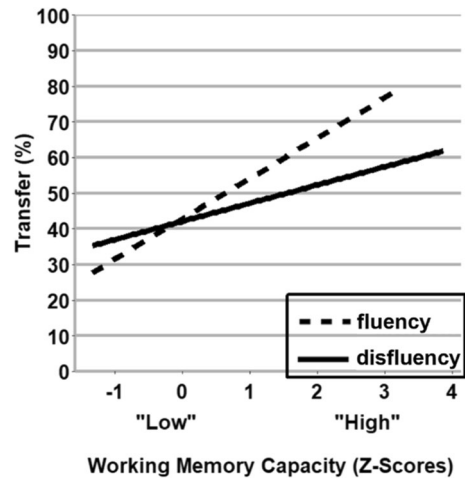
<sup>a</sup> Information was not provided in Lehmann et al. (2016)

\*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$

**Transfer** Except the control variables, all predictors were the same as with retention and comprehension as dependent variables. The regression was conducted using transfer performance as the dependent variable and disfluency manipulation (fluent=0; disfluent=1), standardized WMC, and the interaction term of disfluency \* (standardized) WMC as predictors. Furthermore, the regression model included the standardized control variables enjoyment, interest, motivation during learning; and enjoyment, interest, motivation during assessment of the learning outcomes; as well as prior knowledge.



**Fig. 4** Transfer: Interaction of Working Memory Capacity and Disfluency (Replication Study). *Note.* Lehmann et al. (2016) did not include any graphical representation for transfer performance.  $N=94$  in the replication study (R's regression model excluded 2 persons due to missings)



**Table 5** Regression Analysis for Transfer: Comparison between Original and Replication Study

Effect	Lehmann et al. (2016)			Replication Study		
	Estimate	SE <sup>a</sup>	<i>t</i> (45)	Estimate	SE	<i>t</i> (83)
Disfluency	-4.26	-	-0.46	1.48	6.09	0.24
WMC: Fluent <sup>a</sup>	-	-	-	5.87	5.01	1.17
WMC: Disfluent <sup>a</sup>	-	-	-	4.30	3.88	1.11
Disfluency * WMC	-2.27	-	-0.23	-1.57	6.28	-0.25
Enjoyment during learning	-	-	-	-1.09	3.97	-0.27
Interest during learning	-	-	-	-0.77	5.03	-0.15
Motivation during learning	-	-	-	2.55	4.12	0.62
Enjoyment during assessment	-	-	-	6.52	4.56	1.43
Interest during assessment	-	-	-	2.18	5.45	0.40
Motivation during assessment	-	-	-	1.68	3.92	0.43
Prior Knowledge	-	-	-	14.06***	3.11	4.52
Regression Model	<i>F</i>	<i>R</i> <sup>2</sup> <sub>adj</sub>		<i>F</i> (10, 83)	<i>R</i> <sup>2</sup> <sub>adj</sub>	
	< 1	< 0.01		4.70***	0.28	

*Note.*  $N=47$  for Lehmann et al. (2016).  $N=96$  for the replication study. *SE* = Standard Error

<sup>a</sup> Information was not provided by Lehmann et al. (2016)

The regression model ( $F(10, 83)=4.70$ ,  $p<0.001$ ,  $R^2_{adj}=0.28$ ) was significant, but the only significant predictor was prior knowledge ( $\beta=14.06$ ,  $SE=3.11$ ,  $t(83)=4.52$ ,  $p<0.001$ ). Disfluency manipulation ( $\beta=1.48$ ,  $SE=6.09$ ,  $t(83)=0.24$ ,  $p=0.808$ ), WMC (fluent:  $\beta=5.87$ ,  $SE=5.01$ ,  $t(83)=1.17$ ,  $p=0.245$ ; disfluent:  $\beta=4.30$ ,  $SE=3.88$ ,  $t(83)=1.11$ ,  $p=0.270$ ), and their interaction ( $\beta=-1.57$ ,  $SE=6.28$ ,  $t(83)=-0.25$ ,  $p=0.804$ ) did not significantly predict transfer performance. Because Lehmann et al. (2016) also failed to find an ATI effect on transfer, and given that both their effect size and our effect size were negative, the two studies' findings align. The primary results (null finding) thus could be replicated (see Fig. 4 and Table 5).

## Discussion

Our direct replication of the ATI study of Lehmann et al. (2016) tested the hypothesis that the perceptual disfluency effect regarding text-based learning is moderated by WMC (for other investigations of a moderation effect of WMC, see Miyakawa & Hattori, 2017; Strukelj et al., 2016). Lehmann et al. theorized that perceptually disfluent texts entail greater processing demands, which only would allow learners with higher WMC to engage in deeper and more analytical/elaborative processing rather than heuristic reasoning (see also Diemand-Yauman et al., 2011; Ebersbach et al., 2023; Meyer et al., 2015; Rummer et al., 2016). The primary study by Lehmann et al. found such an ATI effect of WMC on disfluency for retention and comprehension (but not transfer). Particularly, in the disfluent condition, the higher the WMC, the better retention and comprehension. Yet in the fluent condition, no WMC effect was obtained. Note that WMC in general had no positive main effect on learning outcomes in the primary study.

Given the discrepant findings in the disfluency literature, the practical implications of a disfluency\*working memory interaction effect, the small power, and the (at least partly) idiosyncratic statistical decisions of Lehmann et al. (2016), we aimed to bolster the initial results by an independent direct (and preregistered) replication. Thus, we utilized the same learning materials and measures and followed as closely as possible the experimental design and the analytical approach of the primary study (for a more detailed discussion of similarities and differences, see below). To substantiate our empirical findings, we tripled the number of collected participants, which after exclusions resulted in a final sample of  $N=96$ . This is approximately twice the size of the sample size of Lehmann et al. ( $N=47$ ), Strukelj et al. (2016;  $N=53$ ), and Miyakawa and Hattori (2017;  $N=40$ ). We applied the criteria of finding a statistically significant effect ( $p < 0.05$ , two-sided) in the same direction to evaluate the success of the replicability of the primary results (cf. Camerer et al., 2018; Nosek et al., 2015).

The findings of our direct replication are not in line with the results of Lehmann et al. (2016). We did not find an interaction effect of WMC and disfluency on retention performance, albeit our post hoc power of 0.90 exceeded the post hoc power of 0.70 of Lehmann et al. for the overall model. Our only significant predictor in the regression model for retention was motivation during learning. Likewise, we obtained no interaction effect for comprehension performance (our post hoc power = 0.84; Lehmann et al. = 0.62). Moreover, the direction of the interaction effect for comprehension descriptively reversed in our replication sample (negative effect) as opposed to the primary study (positive effect). Because our interaction effect was not significant, we do not interpret this aspect of our data pattern but do note its opposing direction (cf. Miyakawa & Hattori, 2017). For transfer performance, we replicated the primary findings of Lehmann et al.: a null effect (with the same negative coefficient). In our study, the only significant predictor of both comprehension and transfer performance was prior knowledge. In summary, our replication data do not support the hypothesis of a cognitively moderated disfluency effect with WMC acting as a boundary condition.

It is worth noting that our findings are based on the same analytical approach used in the primary study. We thus applied the same data preprocessing criteria and model inclusion criteria as outlined by Lehmann et al. (2016) contingent on our replication sample. This means that we also retained only prior knowledge items with a solution probability of less than 10% or higher than 90%, which in our case were four out of six (instead of three) items. Based on this selection, we filtered for participants with medium prior knowledge (between 25% and 75% points in the prior knowledge test). Moreover, we included prior

knowledge in our regression model (of comprehension and transfer) according to the criterion of Lehmann et al. (=significant correlation of previous knowledge with the learning outcomes or when a  $t$  test showed differences across the disfluency manipulation). We disagree with such extensive use of prior knowledge as a filtered variable itself (three out of six items of prior knowledge retained based on  $> 10\%$  and  $< 90\%$  solution rates), and selection criteria for participants (with medium knowledge,  $> 25\%$  and  $< 75\%$ ), as well as a control variable (e.g., see Table 4), especially given the unvalidated, self-developed nature of the prior knowledge test.

In this respect, we note that the number of motivational control variables we had to include in our models as a result of the criteria by Lehmann et al. (=a significant correlation of the control variables with learning outcomes or significant differences in the motivational control variables across the disfluency manipulation without correction for multiple testing) was larger than the number of control variables Lehmann et al. (2016) included; this was likely due to our increased sample size. With less power, fewer control variables can show a significant correlation or group differences, avoiding the criteria. We had twice as many participants in our final sample as Lehmann et al.. In our replication sample, almost all six motivational control variables fulfilled the criteria for all regression models. Lehmann et al. included different control variables in each model (for retention: motivation during assessment of learning outcomes; for comprehension: prior knowledge). In our view, the adoption of such criteria is not only subjective but also largely power-dependent and data-contingent to the obtained sample. Results thus can be more prone to (random) error, less reliable, and less replicable. Case in point, removing previous knowledge as a control variable in the regression model for comprehension of Lehmann et al. rendered their interaction effect nonsignificant.

Although our power for the overall regression models is higher than the power of the primary study, we need to briefly consider power aspects regarding the interaction effects and alternative analysis. Using the interaction effect size observed in the primary study for retention and given their sample size, Lehmann et al. (2016) had an estimated power of only 45% to detect the interaction effect, based on a simulation tool developed by Finnsaas et al. (2021) and Baranger et al. (2022). A simulation of the necessary sample size to find the interaction effect for retention with a power of 80% required a sample size of 150 participants.

We thus reanalyzed our data with the full sample of 130 participants in various sensible ways. We still found no support for an interaction effect in any of the models, especially no interaction effect for retention. In those rare cases, where the  $p$  value was close to 0.10, albeit only for comprehension, the sign of the coefficients was different from the primary study. This indicated that the interaction had a different pattern, for example, the comprehension effect in the primary study: Disfluency  $B = -2.36$ , WMC: Fluent  $B = -3.90$ , Disfluency \* WMC  $B = 12.62$ ; comprehension effect in our replication: Disfluency  $B = 6.25$ , WMC: Fluent  $B = 11.87$ , Disfluency \* WMC  $B = -9.18$  (no filtering for medium previous knowledge participants and no control variables,  $N = 130$ ).

Note, whether we filtered for previous knowledge in various ways or included control variables in various ways, there was no change in the overall implications of our data. We did not find any support for the proposed interaction effect and the descriptive pattern of the interaction effect (regardless of significance) is dissimilar to the primary study, which suggests that the lack of the interaction effect is not a result of insufficient statistical power.

Such aspects (power and analytical choices) add another layer of reasons for discrepancies and contribute to inconsistencies, which are characteristic of the deviating results of the existing studies examining the interaction between WMC and disfluency. For instance,

Miyakawa and Hattori (2017), in contrast to Lehmann et al. (2016), discovered that participants with *lower* WMC performed better in the disfluency condition. Strukelj et al. (2016), however, found no moderating effect of WMC on disfluency, similar to our present findings. Thus, there is one experiment demonstrating that the disfluency effect is restricted to high-span subjects (Lehmann et al., 2016), one experiment demonstrating that it is restricted to low-span subjects (Miyakawa & Hattori, 2017), and two experiments demonstrating a null effect (Strukelj et al., 2016; the present study). Interpreting these discrepancies proves difficult; the issue is not merely differences in power and analytical choices but also the methodological deviations in the manipulation of disfluency, the measurement of working memory, and the choice of dependent variables. To follow, we consider these differences.

Miyakawa and Hattori (2017), Lehmann et al. (2016), and our replication of Lehmann et al. used a grey-scaled font manipulation, whereas Strukelj et al. (2016) used low-pass filtering to achieve blurriness. Miyakawa and Hattori used a within-subjects manipulation of disfluency (and applied a median split for working-memory capacity), whereas all other studies used a between-subjects design. Only Miyakawa and Hattori found a significant advantage of disfluency, whereby participants learned vocabulary and the dependent variable was free recall; this is in line with a study by Yue et al. (2013) in which participants were presented with word lists suggesting that the disfluency effect emerges only when disfluency was manipulated within-subjects and not between-subjects. The other studies used text materials and either assessed only recall of information regarding how airplanes achieve lift (Strukelj et al.) or recall, comprehension, and transfer on time and date differences across the world (Lehmann et al., 2016, as well as the present replication study).

One explanation for why Miyakawa and Hattori (2017) found an effect and all other studies did not could rest on sample differences (Japanese vs. European participants) and/or material differences (vocabularies vs. expository texts). Another explanation could rest on the proposition that a robust disfluency effect does not exist at all (e.g., Xie et al., 2018; also see Weissgerber et al., 2021). Based on this explanation, the results of Miyakawa and Hattori would be a false positive. However, despite their small sample, their within-subjects manipulation achieved greater power to detect a disfluency effect as opposed to Strukelj et al., who also had a small sample but a between-subjects design. In addition, there is some evidence that the disfluency effect emerges when disfluency was manipulated within-subjects and not between-subjects (e.g., Yue et al., 2013).

If the disfluency effect exists, and it is not practically equivalent to zero, the consideration of power is an important aspect apart from the methodological differences. In line with this, it could be that the effect, albeit nontrivial and existent, is very small. Thus, detection would require a large sample size, especially in between-subjects designs. Assuming a small disfluency effect of  $d=0.20$ , alpha of 0.05, and beta of 0.80 (one-sided) would require 620 participants to detect a significant mean difference between a fluent and a disfluent group. Thus, neither the original nor our experiment were sufficiently powered to detect a potential main effect of disfluency manipulated between-subjects. Importantly, however, neither the goal of the primary study nor of our replication was to detect a main effect. Moreover, a small disfluency effect severely limits its practical relevance, despite its scalability due to easy applicability.

The reasons for the occurrence (Miyakawa and Hattori, 2017; Strukelj et al., 2016) or the lack of a working memory effect (Lehmann et al., 2016; the present replication) across the experiments are ambiguous. Given that the lack of a working memory effect occurs in the primary study and in our study, it could be specific to the learning material/working memory test and procedural choices. In theory, the used WMC test requiring numerical

updating (Oberauer et al., 2000) and the measurement of the dependent variables also requiring numerical computations (e.g., transfer question: “You will depart from Anchorage on June 25 at 10 pm. After a 19-h flight, you land in Moscow. What is the date and time in Moscow?”) allows the possibility of a WMC effect. It could be, nevertheless, that such an effect would only occur when the performance measurement is restricted (system-paced) rather than self-paced (as in the primary study of Lehmann et al. and our replication).

Strukelj et al. (2016) used the OSPAN working memory test (Unsworth et al., 2005) and participants studied (self-paced) how airplanes achieve lift (Mautone & Mayer 2001). The measurement of the recall performance was, however, time restricted. Strukelj et al. found a main effect of working memory on recall performance but no interaction effect with disfluency (manipulated by low-pass filtering the text to appear blurry). However, Miyakawa and Hattori’s (2017) assessment of the recall performance of vocabulary was time restricted (as was the study time in the learning phase), and they did not obtain a working memory effect. Their measurement of WMC was based on two instruments: an operation span test (Izeki, 2010) and a revised reading span test (Mosaka, 2002). Thus, all three studies investigating a possible moderation of the disfluency effect by WMC used different measures of WMC, albeit the operationalization of WMC in terms of a span task overlaps.

It may seem plausible to attribute differences in the results across experiments to different measures of WMC. Although there might be some truth to it, this explanation probably falls short. Strukelj et al. (2016) found no moderating effect of WMC on disfluency, but our replication results were similar in terms of the lack of a moderation effect despite a different WMC measure and learning material. Lehmann et al. (2016) and Miyakawa and Hattori (2017) used different WMC measures and materials, but both experiments found a moderation effect. Thus, an overinterpretation of discrepancies as mere consequences of WMC measurements should be cautioned. It is worthwhile to further consider the hypothesis and obtained effects.

The original hypothesis of Lehmann et al., (2016, p. 94) states:

In the fluency condition, the WMC should not influence retention (Hypothesis 1a), comprehension (Hypothesis 2a), or transfer (Hypothesis 3a). In the disfluency condition, the WMC should affect learning outcomes: The higher the WMC, the better the expected retention (Hypothesis 1b), comprehension (Hypothesis 2b), and transfer performance (Hypothesis 3b) in the disfluency condition.

One could argue that this particular hypothesis is not a convincing instantiation of “WMC as a moderator of the disfluency effect” but rather of “disfluency as a moderator of an effect of WMC” on learning (because the positive effect of working memory is predicted to be “switched off” in the fluent condition and “switched on” in the disfluent condition). However, when WMC is the moderator for the disfluency effect, higher WMC should predict higher learning outcomes in the fluent condition (e.g., Swanson & Alloway, 2012; positive slope of WMC). Moreover, higher WMC should act especially as a performance booster in the disfluent condition (an even steeper positive slope of WMC), but those with lower WMC should not benefit from disfluency. Thus, a positive disfluency effect should be “switched off” for low-capacity participants and should be “switched on” for high-capacity participants.

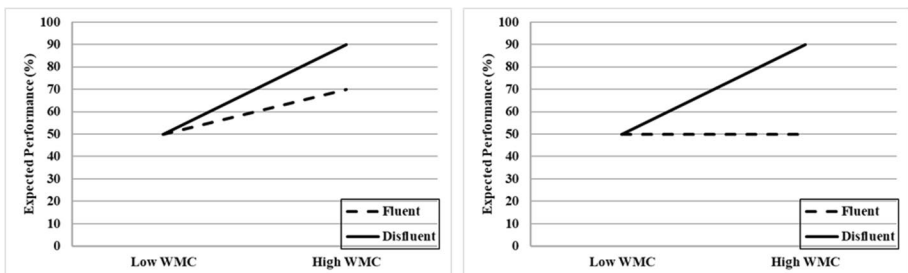
This is relevant with respect to the (surprising) lack of a positive WMC effect in the results of Lehmann et al. (2016). The slope for the WMC effect in the fluency condition is negative (albeit nonsignificant), which descriptively shows a worse performance for participants with higher WMC for the standard (fluent) learning material. By contrast,

the slope for the WMC effect in the disfluency condition is positive. It significantly shows better performance of participants with higher WMC in the disfluent condition. Both opposing slopes appear together as a crossover effect; this can be gauged in Fig. 3. Therefore, we wonder about how this unexpected coefficient for WMC in the fluent condition may have contributed to the emergence of the (crossover) interaction effect. We caution against interpreting the primary data to robustly mean that the disfluency effect is restricted to high-span subjects (“disfluency only paid off when learners had sufficient WMC,” p. 100), when it appears that this effect could partly rest on the (surprising) negative slope of fluency seemingly costing learners with sufficient WMC.

Note, the hypothesis of WMC as moderator of the disfluency effect predicts an interaction effect in the shape of an upward-spreading triangle (see Fig. 5, left). This particular interaction effect entails a positive effect of WMC, which is even further enhanced by disfluency, albeit disfluency itself has no main effect. The findings by Miyakawa and Hattori (2017) are interesting in this context because they show a disfluency effect, but no main effect of WMC, and an interaction effect, whereby disfluency is only advantageous for low-span subjects. This pattern of results is the opposite of the idea that the disfluency effect is “switched off” for low-span individuals and “switched on” for high-span individuals, as predicted by the “working-memory as moderator for disfluency” interaction. Thus, this pattern is the opposite of what should have been the original hypothesis, rather than assuming no working-memory effect in the fluency condition, and a disfluency advantage for high-span individuals (see Fig. 5).

Overall, it is preliminary to gauge the evidence that there is no reliable relation between WMC and the perceptual fluency effect in text-based learning. Given that only a few studies on the interaction effect of WMC\*disfluency exist and given the discrepant experimental approaches, we are cautious to say that the findings by Miyakawa and Hattori (2017) falsify the WMC-as-moderator-for-disfluency hypothesis. We are also cautious in claiming that the null findings of our replication falsify Lehmann et al. (2016). In our view, the discrepancies across the experiments investigating the WMC\*disfluency interaction emphasize the need for a systematic investigation of this important hypothesis.

Despite disagreeing on the formulation of this hypothesis with Lehmann et al., we emphasize the importance of the predicted interaction of WMC\*disfluency by Lehmann et al. for the investigation and theoretical advancement of the disfluency effect. We agree with Lehmann et al. that WMC is a highly interesting boundary condition to scrutinize as a possible explanation for the fickle nature of the disfluency effect. Moreover, a



**Fig. 5** Theoretically Expected Pattern of Results for the Interaction Effect of Disfluency\*WMC. *Note.* Left-side: our theoretically expected pattern for the disfluency\*WMC interaction; right-side: theoretically expected pattern of Lehmann et al. (2016)



WMC\*disfluency interaction effect would only be obtained if the theoretical explanation holds that disfluency leads to deeper processing (Diemand-Yauman et al., 2011) but not if the disfluency effect is nonexistent or is based on another mechanism (e.g., novelty/distinctiveness effect: Rummer et al., 2016). We thus reiterate the value of examining the predicted interaction effect of Lehmann et al. (2016).

An obvious candidate to contribute to discrepant replication findings in Lehmann et al.'s and our study are the data collection methods and settings. The primary study was conducted in small groups in the laboratory with the materials in paper–pencil format. Due to COVID-19 restrictions, the replication study was conducted online, reliant on participants' personal computers at home. We set up both replication sessions to be conducted in small groups (maximum five persons) via video conference to monitor participants' behavior and closely match the laboratory conditions (e.g., group setting). Again, we can only speculate if the diverging method influenced our findings.

For replications to be considered direct replications, the replication does not need to (and cannot) recreate the exact context and conditions of the primary study. Instead, the direct replication must recreate the theoretical conditions known to be necessary to create the to-be-replicated effect: Accordingly, only the critical features of a primary study need to be the same, which allows some deviation (Zwaan et al., 2018). We believe the change from a laboratory setting to an online setting to be a less probable (albeit possible) context-based post hoc explanation of the discrepant results. To foreshadow, we have limited reasons to assume that the relationship between working memory and disfluency was moderated by lab vs. online context and thus to be the main explanation for the discrepant results of the primary study and the replication. Nevertheless, we cannot rule out this possibility.

An interesting post hoc contextual factor to consider would be the effect of paper–pencil versus computer screen for the disfluency manipulation. Reading a disfluent text on a computer screen instead of on a piece of paper might generate subtle perceptual differences for which we cannot account. To our knowledge, there is at least one study that tested a potential interaction of the perceptual disfluency effect and the presentation medium: Sidi et al. (2016) found that disfluent font onscreen resulted in higher success rates than fluent font, but on-paper fluent fonts outperformed disfluent fonts. The differential effects were only found in one out of two experiments; yet, assuming there are indeed media-dependent differences to be reckoned with, we would expect the screen presentation in our replication to be a conducive condition to the creation of a disfluency effect that could be moderated by WMC. In other words, our screen presentation should not have hindered obtaining a direct replication effect but aided detection.

Another interesting post hoc explanation for the discrepant results concerns variations in participants' screens, which may have affected the effectiveness of the disfluency manipulation. For this reason, participation was exclusively possible by computer and not by tablet or smartphone. We programmed our study in a way that any attempt to access the experiment via tablet or smartphone led to a technical screen-out right away. This should have substantially reduced the variation in screen size, albeit some variation in computer screen size remained. This may have created some noise, as display size alone may have an effect in terms of the (dis)fluency manipulation. These variations may have been less consequential because the grey-scaled disfluent text in Haettenschweiler remained disfluent in smaller screen sizes and larger screen sizes (comparison of disfluency in larger screens and smaller screens) and also compared to the fluent condition in Arial; this argument holds even if participants may have compensated for disfluency by adjusting their displays (e.g., by zooming in). Only at extreme levels of zooming in (>400%) do disfluent texts become fluent but at the expense of the number of words displayed at a time.



We cannot rule out the possibility that participants zoomed in and adjusted their screen online, whereas in the lab, the printed paper does not allow the option of such an adjustment. Nevertheless, we point out that the data collection was supervised via Zoom, which should have discouraged participants from adjusting the screen (and we found no indication of zooming in during the review of participants' open comments). We cannot rule out that a lack of control (rather than the online setting per se compared to the lab setting) introduced some noise and variations in the strength of the disfluency manipulation that could have affected our findings. To repeat the argument, in most cases, Haettenschweiler would still have been disfluent despite possible differences in contrast or size (e.g., zooming), especially relative to the fluent font Arial. We thus have little reason to believe that the lack of a moderation effect is simply due to the online setting (and possible variations in the strength of the disfluency manipulation).

Finally, we cannot exclude sampling error and variations in sample characteristics. Our sample was overall "older" than Lehmann et al.'s (29 instead of 22 years), and the primary study employed a convenience student sample, but our convenience sample included students and non-students with the majority being students (65%). The age gap is not wide enough to convincingly argue for differential abilities to learn the material. We have no good theoretical or practical reasons to believe that the variations in the sample led to the discrepant results.

In summary, considering the inconsistent empirical findings of our direct replication and those studies investigating the interaction between the disfluency effect and WMC (Lehmann et al., 2016; Miyakawa & Hattori, 2017; Strukelj et al., 2016), it is inconclusive whether higher WMC is a prerequisite to profit from the disfluency effect. The hypothesis as such, however, is theoretically and practically important. To conduct a replication that provides clear and unambiguous results, it is necessary to carry out more than one replication experiment (Hedges & Schauer, 2019). Hence, further examination of the potential interplay of WMC and disfluency is necessary, especially in the laboratory.

In conclusion, this experiment revealed no support for WMC to enable a learning benefit of perceptual disfluency. The direct replication does not provide evidence of the existence of an interaction of WMC and disfluency with the utilized materials and measures. Our results add to the evidence of the disfluency effect and its potential boundary conditions to be "fickle" (e.g., Rummer et al., 2016). Until substantiated, reliance on the disfluency effect and its educational application should be treated with great caution.

**Acknowledgements** We would like to thank the DFG research group—lasting learning: cognitive mechanisms and effective instructional implementation (FOR 5254).

**Author contribution** All authors contributed to the study conception and design. Material preparation was performed by D. I. P. Terhorst. Data collection and analysis was performed by D. I. P. Terhorst and S.C. Weissgerber. The first draft of the manuscript was written by D. I. P. Terhorst, and all subsequent versions were written and rewritten by S.C. Weissgerber. All authors commented on previous versions of the manuscript. R. R. provided critical feedback. All authors read and approved the final manuscript.

**Funding** Open Access funding enabled and organized by Projekt DEAL. No funding was received to assist with the preparation of this manuscript.

**Data availability** The data and materials are publicly accessible on the Open Science Framework page (osf.io/4ja26). This study had been pre-registered (AsPredicted #45573: [aspredicted.org/blind.php?x=67um2p](https://aspredicted.org/blind.php?x=67um2p)).

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest. All research involving human participants followed ethical guidelines and the Declaration of Helsinki. All participants voluntarily took part

in the study based on informed consent and received debriefing.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. SAGE Publications Inc.
- Alter, A. L., & Oppenheimer, D. M. (2008). Effects of fluency on psychological distance and mental construal (or why New York is a large city, but New York is a civilized jungle). *Psychological Science*, *19*(2), 7. <https://doi.org/10.1111/j.1467-9280.2008.02062.x>
- Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219–235. <https://doi.org/10.1177/1088868309341564>
- Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *136*(4), 569–576. <https://doi.org/10.1037/0096-3445.136.4.569>
- Baranger, D. A. A., Finsaas, M. C., Goldstein, B. L., Vize, C., Lynam, D., & Olino, T. M. (2022). Tutorial: Power analyses for interaction effects in cross-sectional regressions. PsyArXiv. <https://doi.org/10.31234/osf.io/5ptd7>
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about Knowing* (pp. 185–205). The MIT Press. <https://doi.org/10.7551/mitpress/4561.003.0011>
- Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Jahannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T. . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, *2*(9), 637–644. <https://doi.org/10.1038/s41562-018-0399-z>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates
- Cowan, N. (2016). *Working memory capacity* (Classic). Routledge. <https://doi.org/10.4324/9781315625560>
- Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the Bold (and the Italicized): Effects of disfluency on educational outcomes. *Cognition*, *118*(1), 111–115. <https://doi.org/10.1016/j.cognition.2010.09.012>
- Dunlosky, J., & Mueller, M. L. (2016). Recommendations for exploring the disfluency hypothesis for establishing whether perceptually degrading materials impacts performance. *Metacognition and Learning*, *11*(1), 123–131. <https://doi.org/10.1007/S11409-016-9155-9>
- Ebersbach, M., Guschlbauer, J. A., & Rummer, R. (2023). Flickering presentations do affect the judgment of learning but not the learning outcome. *Applied Cognitive Psychology*, *37*(4), 889–898. <https://doi.org/10.1002/acp.4086>
- Eitel, A., & Kühl, T. (2016). Effects of disfluency and test expectancy on learning with text. *Metacognition and Learning*, *11*(1), 107–121. <https://doi.org/10.1007/s11409-015-9145-3>
- Eitel, A., Kühl, T., Gerjets, P., & Scheiter, K. (2014). Disfluency meets cognitive load in multimedia learning: Does harder-to-read mean better-to-understand? *Applied Cognitive Psychology*, *28*(4), 488–501. <https://doi.org/10.1002/acp.3004>

- Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223–241. <https://doi.org/10.1177/1745691612460685>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Finsaas, M. C., Baranger, D. A. A., Goldstein, B. L., Vize, C., Lynam, D., & Olino, T. M. (2021). InteractionPowerR Shiny App: Power Analysis for Interactions in Linear Regression. Available at <https://mfinsaas.shinyapps.io/InteractionPowerR/> Retrieved [06.01.2023]
- French, M. M. J., Blood, A., Bright, N. D., Futak, D., Grohmann, M. J., Hasthorpe, A., Heritage, J., Poland, R. L., Reece, S., & Tabor, J. (2013). Changing fonts in education: How the benefits vary with ability and dyslexia. *The Journal of Educational Research*, 106(4), 301–304. <https://doi.org/10.1080/00220671.2012.736430>
- Geller, J., & Peterson, D. (2021). Is this going to be on the test? Test expectancy moderates the disfluency effect with sans forgetica. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(12), 1924–1938. <https://doi.org/10.1037/xlm0001042>
- 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>
- Klepsch, M., & Seufert, T. (2012, April 9). *Subjective differentiated measurement of cognitive load*. 5th International Cognitive Load Theory Conference, Tallahassee (USA).
- Kühl, T., Eitel, A., Scheiter, K., & Gerjets, P. (2014). A call for an unbiased search for moderators in disfluency research: Reply to Oppenheimer and Alter (2014). *Applied Cognitive Psychology*, 28, 805–806. <https://doi.org/10.1002/acp.3030>
- Lehmann, J., Goussios, C., & Seufert, T. (2016). Working memory capacity and disfluency effect: An aptitude-treatment-interaction study. *Metacognition and Learning*, 11(1), 89–105. <https://doi.org/10.1007/s11409-015-9149-z>
- Leiner, D. J. (2019). SoSci Survey (Version 3.1.06) [Computer Software]. <https://www.sosicisurvey.de>
- Miyakawa, N., & Hattori, M. (2017). Effects of disfluency of letters on memorizing words: A perspective from working memory capacity. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 24(3), 450–456. <https://doi.org/10.11225/jcss.24.450>
- Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—Facets of a cognitive ability construct. *Personality and Individual Differences*, 29(6), 1017–1045. [https://doi.org/10.1016/S0191-8869\(99\)00251-2](https://doi.org/10.1016/S0191-8869(99)00251-2)
- Owen, A. M., Lee, A. C. H., & Williams, E. J. (2000). Dissociating aspects of verbal working memory within the human frontal lobe: Further evidence for a “process-specific” model of lateral frontal organization. *Psychobiology*, 28(2), 146–155. <https://doi.org/10.3758/BF03331974>
- Rummer, R., Schweppe, J., & Schwede, A. (2016). Fortune is fickle: Null-effects of disfluency on learning outcomes. *Metacognition and Learning*, 11(1), 57–70. <https://doi.org/10.1007/s11409-015-9151-5>
- Schmidt, R., & Bjork, R. (1992). New Conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3(4), 207–217. <https://doi.org/10.1111/j.1467-9280.1992.tb00029.x>
- Schneider, W., Schlagmüller, M., & Ennemoser, M. (2007). *Lesegeschwindigkeits- und verständnistest für die Klassenstufen 6–12 (LGVT 6–12) [Reading speed and comprehension test for class levels 6–12]*. Hogrefe.
- Schnotz, W., & Bannert, M. (1999). Einflüsse der Visualisierungsform auf die Konstruktion mentaler Modelle beim Bild- und Textverstehen [Influence of the type of visualization on the construction of mental models during picture and text comprehension]. *Experimental Psychology*, 46(3), 217–236. <https://doi.org/10.1026/0949-3964.46.3.217>
- Schweppe, J., & Rummer, R. (2014). Attention, working memory, and long-term memory in multimedia learning: An integrated perspective based on process models of working memory. *Educational Psychology Review*, 26(1), 285–306. <https://doi.org/10.1007/s10648-013-9242-2>
- Sidi, Y., Ophir, Y., & Ackerman, R. (2016). Generalizing screen inferiority—Does the medium, screen versus paper, affect performance even with brief tasks? *Metacognition and Learning*, 11(1), 15–33. <https://doi.org/10.1007/s11409-015-9150-6>
- Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory*, 4(6), 592–604. <https://doi.org/10.1037/0278-7393.4.6.592>

- Snow, R. E. (1989). Aptitude-treatment interaction as a framework of research in individual differences in learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences: Advances in theory and research* (pp. 13–59). Freeman.
- Song, H., & Schwarz, N. (2008). Fluency and the detection of misleading questions: Low processing fluency attenuates the Moses illusion. *Social Cognition*, 26(6), 791–799. <https://doi.org/10.1521/soco.2008.26.6.791>
- Strukelj, A., Scheiter, K., Nyström, M., & Holmqvist, K. (2016). Exploring the lack of a disfluency effect: Evidence from eye movements. *Metacognition and Learning*, 11(1), 71–88. <https://doi.org/10.1007/s11409-015-9146-2>
- Swanson, H. L., & Alloway, T. P. (2012). Working memory, learning, and academic achievement. In K. R. Harris, S. Graham, T. Urdan, C. B. McCormick, G. M. Sinatra, & J. Sweller (Eds.), *APA educational psychology handbook, Vol. 1. Theories, constructs, and critical issues* (pp. 327–366). American Psychological Association. <https://doi.org/10.1037/13273-012>
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.
- Taylor, A., Sanson, M., Burnell, R., Wade, K. A., & Garry, M. (2020). Disfluent difficulties are not desirable difficulties: The (lack of) effect of Sans Forgetica on memory. *Memory*, 28(7), 850–857. <https://doi.org/10.1080/09658211.2020.1758726>
- Tulving, E., & Kroll, N. (1995). Novelty assessment in the brain and long-term memory encoding. *Psychonomic Bulletin & Review*, 2(3), 387–390. <https://doi.org/10.3758/BF03210977>
- Weissgerber, S. C., & Reinhard, M.-A. (2017). Is disfluency desirable for learning? *Learning and Instruction*, 49(1), 199–217. <https://doi.org/10.1016/j.learninstruc.2017.02.004>
- Weissgerber S. C., Brunmair, M., & Rummer, R. (2021). Null and void? Errors in meta-analysis on perceptual disfluency and recommendations to improve meta-analytical reproducibility. *Educational Psychology Review*, 33, 1–27. <https://doi.org/10.1007/s10648-020-09579-1>.
- Weltman, D., & Eakin, M. (2014). Incorporating unusual fonts and planned mistakes in study materials to increase business student focus and retention. *INFORMS Transactions on Education*, 15(1), 156–165. <https://doi.org/10.1287/ited.2014.0130>
- Wessel, I., Albers, C. J., Zandstra, A. R. E., & Heininga, V. E. (2020). A multiverse analysis of early attempts to replicate memory suppression with the Think/No-think Task. *Memory*, 28(7), 870–887. <https://doi.org/10.1080/09658211.2020.1797095>
- Wicherts, J. M., Veldkamp, C. L. S., Augustejn, H. E. M., Bakker, M., Van Aert, R. C. M., & Van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7(1), 1832. <https://doi.org/10.3389/fpsyg.2016.01832>
- Xie, H., Zhou, Z., & Liu, Q. (2018). Null effects of perceptual disfluency on learning outcomes in a text-based educational context: A meta-analysis. *Educational Psychology Review*, 30(3), 745–771. <https://doi.org/10.1007/s10648-018-9442-x>
- 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(1), 229–241. <https://doi.org/10.3758/s13421-012-0255-8>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.