Journal of Memory and Language 116 (2021) 104185

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Discrepant findings on the relation between episodic memory and retrieval practice: The impact of analysis decisions

ARTICLE INFO

Keywords
Retrieval practice
Episodic memory
Individual differences
Open science

In this letter, we (1) highlight discrepancies in the correlation between episodic memory and retrieval practice and, (2) demonstrate that contrasting conclusions can reached after making different analytic choices when combining data. Brewer & Unsworth (2012), Pan, Pashler, Potter, & Rickard (2015), and Robey (2019) all examined individual differences in episodic memory and retrieval practice in their original studies. Brewer and Unsworth (2012) assessed relations between a retrieval practice score derived from a cued recall task and individual differences in multiple measures of episodic memory, attention control, working memory capacity, and general fluid intelligence. Participants who exhibited poorer performance profiles in episodic memory (based on a composite from delayed free recall, cued recall, picture source recognition, and gender source recognition tasks) showed larger retrieval practice effects. Since this publication, a great deal of effort has been put into better examining individual differences in retrieval practice using a variety of samples and research designs.

Pan, Pashler, Potter, and Rickard (2015) attempted to replicate the Brewer and Unsworth findings, but found no significant relations between retrieval practice effects and the same episodic memory (EM) composite in two separate studies. Pan et al. (2015) suggested that Brewer and Unsworth's (2012) sample was likely composed of predominantly high ability participants, which influenced the correlations. Given Pan et al.'s reasoning of possible differences in the ability ranges for the samples across studies, Unsworth (2019) combined data from both Brewer and Unsworth (2012) and Pan et al. (2015) in order to widen the overall ability distribution and increase the overall sample size to better examine relations between EM and retrieval practice. With this combined sample (N = 349), Unsworth reported a correlation of $r = -.19$, $p < .001$ between EM and retrieval practice suggesting a weak negative relation.

In a more recent study, Robey (2019) attempted to replicate and extend these findings. In Study 1, where EM was assessed, Robey suggested that the results were inconclusive (based on Bayes Factors) in terms of the relation between EM and retrieval practice. To better examine these relations, Robey combined her Study 1 data with Brewer and Unsworth's (2012) data and Pan et al.'s (2015) data resulting in a much larger sample size. Robey reported no significant relationship between EM abilities and retrieval practice. However, Unsworth (2019) independently analyzed the same combined data from all three papers and came to a very different conclusion. Specifically, after combining the same data (N = 548) Unsworth reported a correlation of $r = -.22$, $p < .001$ between EM abilities and retrieval practice effects.

Why is it that the two analyses from the same data resulted in very different results and conclusions? The discrepancy in results comes from a specific decision to either mean-center the variables in each study or not mean-center the variables. Specifically, Unsworth (2019) did not mean-center the variables with the logic being that differences across studies reflect real differences in abilities that need to be taken into account. Robey (2019), however, did mean-center the variables within each study, with the logic being that there were methodological differences between the tasks used, and several of these differences are known to influence participant performance (see below). This decision to mean-center variables within each sample can drastically change the resulting correlations by removing potentially important between-sample variation necessary for recovering correlations between constructs of interest. In particular, changes in the rank-ordering of individuals can occur when mean-centering within each sample and then combining those samples as opposed to mean-centering the combined data across samples. For example, consider the delayed free recall task used as a measure of episodic memory in each study described earlier. In this task, participants were presented with six lists of 10 words each and had to recall the list of words after a short distractor task. Note, Brewer and Unsworth and Pan et al. used the exact same delayed free recall task, but Robey used a slightly different task. Shown in Fig. 1a are proportion correct scores on delayed free recall for each study. As can be seen, Brewer and Unsworth's distribution of scores was fairly wide with scores ranging from very high to very low. Pan et al.'s distribution of scores tended to be lower and tighter, whereas Robey's distribution of scores was more in the middle of the other studies. Given that Brewer and Unsworth and Pan et al. used the exact same task, it is reasonable to conclude that Pan et al.'s participants had generally lower delayed free recall scores than Brewer and Unsworth. Now, what happens if we mean-center these scores within each sample and then combine the mean-centered scores? Shown in Fig. 1b are the mean-centered

https://doi.org/10.1016/j.jml.2020.104185
Received 20 December 2019; Received in revised form 12 October 2020; Accepted 18 October 2020
Available online 9 November 2020

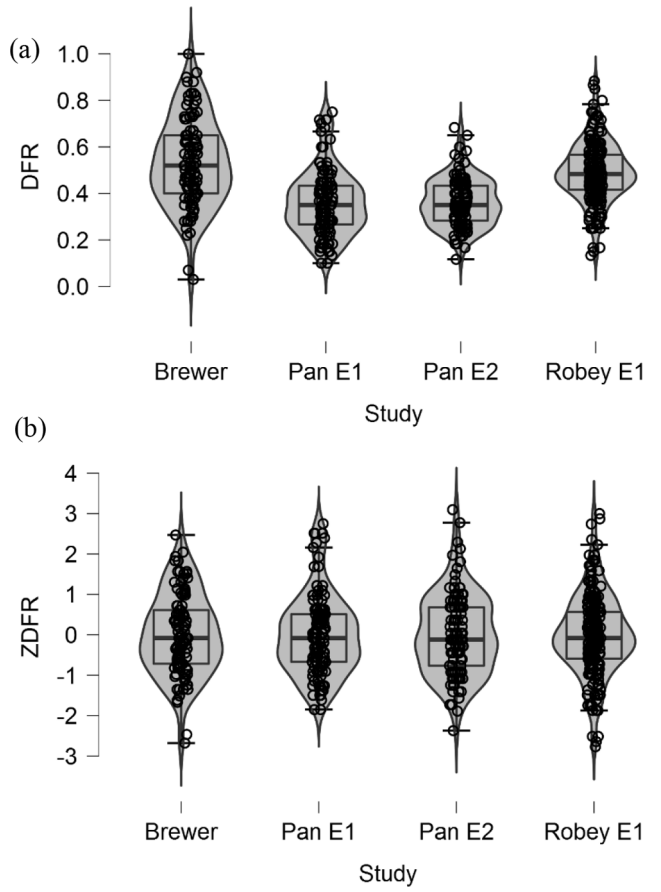

Fig. 1. (a) Violin plots of raw delayed free recall scores in each study. (b) Violin plots of mean-centered delayed free recall scores in each study.

proportion correct scores on delayed free recall for each study. As can be seen, the distribution of scores in Pan et al. have now been altered (i.e., moved up) compared with the distributions for Brewer and Unsworth and Robey. Thus, although the rank order of participants within each sample has not changed, if we now combine the data, the rank ordering of individuals in the combined data will change compared to the raw data. That is, some participants will have similar z-scores (around .60, for example) but who have very different raw delayed free recall scores (.42 and .30). Thus, mean-centering in this manner can potentially result in meaningful individual differences being obscured because some participants rankings are increasing (e.g., mostly from Pan et al.'s study in this example) to the same level as others (e.g., mainly for Brewer and Unsworth's study in this example).

To see if differences in mean-centering account for discrepancies across published studies, we reanalyzed the data both with and without mean-centering in each sample before aggregating studies together. Correlations from all individual and combined studies both mean-centered and not mean-centered can be found in Table 1. These correlations will not match the correlations reported in the original papers as a different method was used to create the composite scores. In all three original papers composites scores were created using z-score composites (i.e., taking the mean of the standardized scores for each variable) in which each task makes an equal contribution to the composite. In the present analysis composite scores were created by extracting the scores from a factor analysis using principal axis factoring in which each task makes a weighted contribution to the composite. When mean-centering, we first mean-centered each EM task and the testing effect scores within each sample. With mean-centering within each sample for all studies combined, the correlation was $r = -0.01$, $p = .89$, suggesting no relation between EM abilities and retrieval practice (see Fig. 2a) consistent with

Table 1
Original, not mean centered, and mean centered correlations between episodic memory composites and the magnitude of the testing effect. Note: Robey (2019, Study 2) did not include episodic memory measures and is therefore not included in the present analyses.

Study	Correlation between episodic memory composite and testing effect
Individual Studies	
Brewer & Unsworth (2012)	r = −0.27, t(105) = −2.85, p = 0.01
Pan et al., (2014; Study 1)	r = 0.14, t(118) = 1.50, p = 0.14
Pan et al., (2014, Study 2)	r = 0.13, t(120) = 1.43, p = 0.15
Robey (2019, Study 1)	r = −0.03, t(197) = −0.38, p = 0.71
Combined (not mean centered)	
Pan et al. (Study 1 & Study 2)	r = 0.12, t(240) = 1.92, p = 0.06
Brewer & Unsworth & Pan et al (Study 1 & Study 2)	r = −0.19, t(347) = −3.51, p = 0.001
All studies	r = −0.17, t(546) = −3.92, p < .001
Combined (mean centered)	
Pan et al. (Study 1 & Study 2)	r = 0.13, t(240) = 1.98, p = 0.05
Brewer & Unsworth & Pan et al (Study 1 & Study 2)	r = 0.01, t(347) = 0.14, p = 0.89
All studies	r = −0.01, t(546) = −0.13, p = 0.89

Robey (2019). Next, we examined the relationship between the EM factor composite and retrieval practice without mean-centering in each sample. The correlation was $r = -.17$, $p < .001$, suggesting a weak negative correlation between EM abilities and retrieval practice (see Fig. 2b) consistent Unsworth (2019). Note that the correlation reported in Unsworth (2019) and the current correlation are different because Unsworth (2019) was based on the original combined data reported in Robey (2019). However, there was an error in the delayed free recall scores for the combined data. The current correlation is based on the data with the corrected delayed free recall scores. These re-analyses suggest that discrepancies across the two papers when all data were combined are due to the decision to mean-center within each sample or not. Mean-centering within each sample results in no correlation between EM abilities and retrieval practice, whereas not mean-centering within each sample results in a small negative correlation between EM abilities and retrieval practice. Thus, different results can arise depending on the specific data analytic decisions that are made, as pattern that has been previously shown by Silberzahn et al. (2018).

Within the context of the current dataset and question, mean-centering the data may be problematic given that it can change the rank ordering of individuals in ways that drastically reduce individual differences resulting in greatly attenuated correlations. However, at the same time, we acknowledge that there are important methodological differences across the studies that need to be taken into account. Not accounting for methodological differences may result in significant results that are due to inflated variability due to methodological differences and not real effects. For example, even though Pan et al. (2015) used the same tasks as Brewer and Unsworth (2012), they made a number of changes that could influence the relations (e.g., changing the timing of some of the tasks). Robey (2019) also used similar tasks as the prior studies, but with a few important differences. First, the stimuli for both the cued-recall and testing effect tasks were modified in order to capture variability in strategy use. These changes resulted in easier-to-remember stimuli for both the EM cued-recall task and testing effect task compared to the previous studies. Second, in Brewer and Unsworth and Pan et al., the delay between the initial retrieval practice task and the final test was 24 h, whereas in Robey it was only 30 min. Both the difficulty of the to-be-remembered items and delay between practice and final test are known to influence the magnitude of the testing effect (e.g., Pyc & Rawson, 2009; Roediger & Karpicke, 2006). More specifically, these differences are known to attenuate the testing effect. There were a number of additional differences among the studies. These differences across studies make it difficult to draw firm conclusions when raw data is pooled across studies.

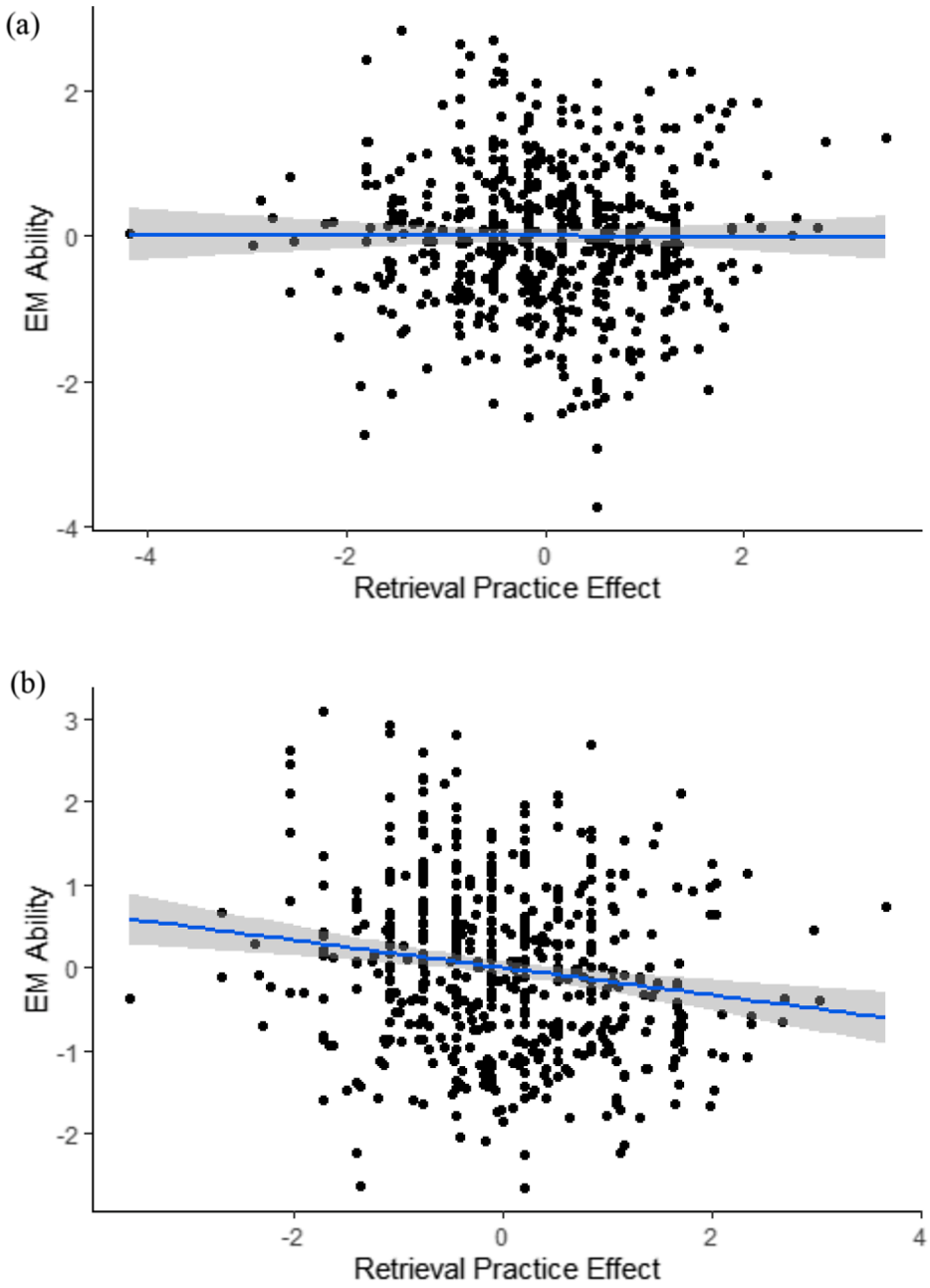

Fig. 2. (a) Scatter showing the relation between episodic memory (EM) abilities and the magnitude of the retrieval practice effect after mean-centering each task within each sample. (b) Scatter showing the relation between episodic memory (EM) abilities and the magnitude of the retrieval practice effect without mean-centering each task within each sample. Error bands represent 95% confidence intervals.

This example highlights the need for researchers to be completely transparent in their analysis choices. Although Robey (2019) explicitly stated that the combined data were mean-centered within study, it took the authors several days of email correspondence to determine where the differences were coming from. Recently, there has been a push for open science practices within the psychological sciences (e.g., Nosek et al., 2015). These discussions have tended to focus on pre-registering study plans (Chambers, 2013; Nosek & Lakens, 2014) and sharing data (e.g., Kidwell et al., 2016) which we agree are important. The present example, however, also highlights the importance of also sharing analysis code. As such, the analysis code and data for the present paper can be found on the open science framework (https://osf.io/dgcaz/). Providing analysis code is an important aspect of open science which will become increasing important as we attempt to increase the reproducibility of our field.

Ultimately, to fully address the question of whether episodic memory serves as an individual difference in retrieval practice, additional research is needed, as there are limitations to both prior attempts to combine previously collected data that was not designed for that purpose. Specifically, a large-scale pre-registered study that includes participants from across a wide range of abilities is needed in order to better assess possible relations between EM abilities and individual differences in retrieval practice. This same practice would also be beneficial for the assessment of other potential individual differences measures that may account for a larger percentage of variance in the effect of retrieval practice.

References

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory & Language, 66*, 407–415.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*, 609–610.
Kidwell, M. C., Lazarevic, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., … Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective, method for increasing transparency. *PLOS Biology, 14*, Article e1002456.
Nosek, B., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 1422–1425.
Nosek, B., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137–141.
Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language, 83*, 53–61.
Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*, 437–447.
Robey, A. (2019). The benefits of testing: Individual differences based on student factors. *Journal of Memory and Language, 108*, Article 104029.
Roediger, H. L. I. I. I., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improved long-term retention. *Psychological Science, 17*(3), 249–255.
Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E. C., Nosek, B. A., et al. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances and Methods in Psychological Science, 1*, 337–356.
Unsworth, N. (2019). Individual differences in long-term memory. *Psychological Bulletin, 145*, 79–139.

Gene A. Brewer
Arizona State University, United States

Alison Robey
University of Maryland, United States

Nash Unsworth
University of Oregon, United States