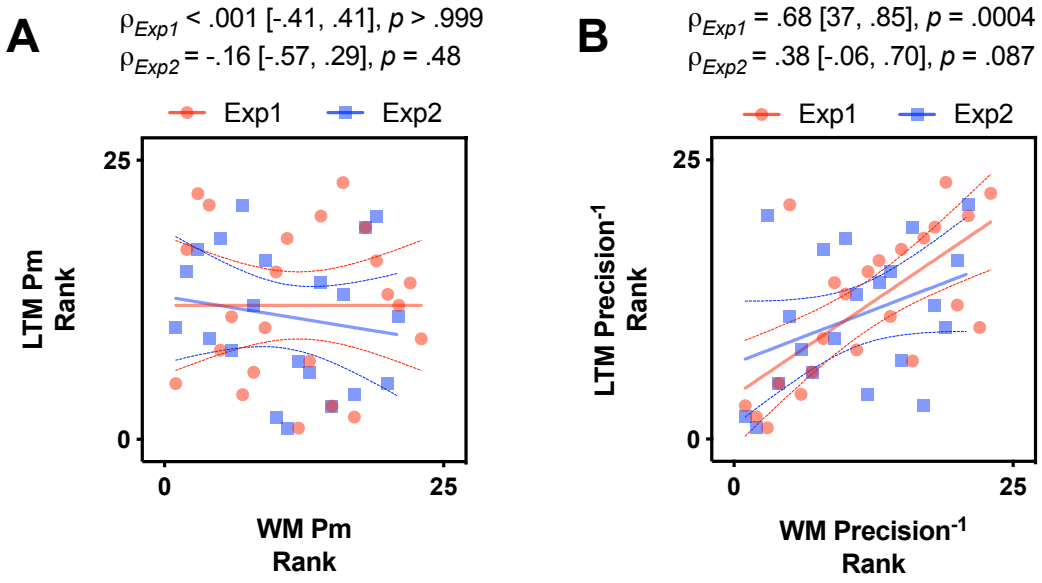
Supplementary Information

Spearman Rank-order Correlation of Estimated Parameters in Biderman et al.

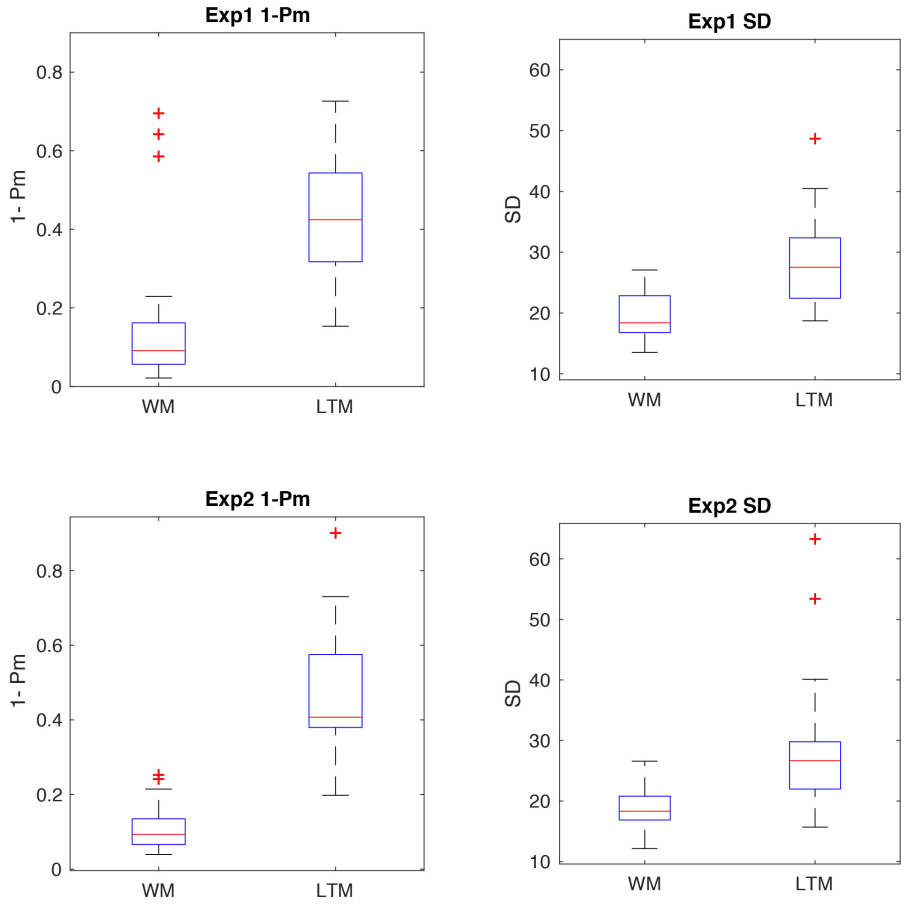
We further evaluated the correlated patterns of parameters estimated by Biderman and colleagues using Spearman rank-order correlation to mitigate issues concerning potential outliers. The meta-analysis procedure was the same as that used for Pearson correlation. These meta-analyses steps were outlined in Rosenthal and DiMatteo (2001). Similar to the results using Pearson correlation reported in the main text, WM and LTM precision estimates were highly correlated with one another (Supplementary Figure S1). When they were meta-analytically combined to increase statistical power, there was a strong correlation between these estimates (ρ = .56 [95% confidence intervals: .32, .72], Z = 3.73, p = .00019). In comparison, the qualitative aspect of WM and LTM, estimated by the probability of successful recall (Pm), did not show a significant association across WM and LTM in each individual experiment or in the meta-analysis across experiments (all ps > .40). Critically, the significant correlation in precision estimates across WM and LTM was significantly different from the non-significant correlation in Pm across WM and LTM (Z = 3.48, p = .0005) by comparing these correlated but non-overlapping meta-analytically combined correlations across experiments (Raghunathan, Rosenthal, & Rubin, 1996).



Supplementary Figure S1. Correlated patterns of probability of successful recall (Pm, namely one minus probability of guessing, A) and mnemonic precision^{-1} (estimated from standard deviation of recall error distribution from the mixture model, B) across WM and LTM, using the Maximum Likelihood Estimation (MLE) parameters in Experiment 1 and 2 provided by Biderman et al. (osf.io/93cvs/). Rank-ordered data in each experiment are evaluated by Spearman rank-order correlation, with each data point representing the ranks of estimated parameters for each participant. Correlated patterns are plotted in different colors for different experiments, and are analyzed separately with the correlation coefficients, 95% confidence intervals, and p values shown on top of the figures. The solid lines represent linear fits of the data, and the dashed lines represent 95% confidence intervals of the linear fits.

Reliability of Parameter Estimates in Biderman et al.

Measurement of reliability is pivotal for correlation analysis. We thus evaluated the reliability of parameter estimates in Biderman et al.'s data. First, because a restricted range (e.g., ceiling level performance) in a measure can limit the reliability of the measure (Sackett, Laczo, & Arvey, 2002), we first assessed whether this issue can account for the presence of a significant correlation between WM and LTM measures in SD but not in Pm. As shown in Supplementary Figure S2 (also in Fig. 2 of Biderman et al), WM Pm and SD both have restricted data ranges as compared with those in the LTM condition. It is thus less likely that restricted data range alone could account for the presence of a significant correlation in SD between WM and LTM, but the lack of a significant correlation in Pm between WM and LTM.



Supplementary Figure S2. The boxplots from Experiment 1 and 2 of Biderman et al. Note, to keep consistent with the original data reported in Biderman et al., we reported 1- Pm (i.e., the probability of guessing, PG in Biderman et al.). As shown in the figure, the issue of restricted data range in WM exists for both Pm and SD, such that the LTM data, in general, are more variable. Although there are several outliers in this boxplot based on default algorithm in Matlab, these data have already been outlier-corrected by Biderman et al. (1-Pm < 0.05 and SD > 80, as well as data that exceed 3 standard deviations of a variable), and hence we do not think that one should further remove the outliers. The same data can be seen in Fig. 2 of Biderman et al.

Next, we tried to directly estimate the reliability of model parameters in the current dataset using a bootstrap method (Davison & Hinkley, 1997). In brief, for each subject and experiment, we generated pairs of Pm estimates and pairs of SD estimates, based on resampling the trial-by-trial data with replacement. We repeated this procedure for the whole data set 100 times and took the mean correlation between pairs as our estimate of reliability for each measurement. This method has been used in the modeling literature to estimate the reliability of a certain measure (see Bays, 2018 as an example). The table below summarizes reliability estimates in the current dataset.

Supplementary Table S1. Reliability of Parameter Estimates in Biderman et al.

Reliability estimates from 100 resampling iterations	Exp 1		Exp 2		Naively Averaging across Experiments	
	Pm	SD	Pm	SD	Pm	SD
WM	0.8196	0.7368	0.7038	0.6929	0.7617	0.7148
LTM	0.6288	0.4207	0.7059	0.5792	0.6673	0.5000

Overall, the reliability estimates for the probability of successful recall are quite reasonable (0.6 ~ 0.7). This is not particularly different from the psychometric properties of major personality scales (assuming that personality traits are stable, e.g., Hahn, Gottschling, & Spinath, 2012). The lowest reliability estimates are actually that for the SD in LTM. That said, it is still not very low either. Using these reliability measures to form an upper bound for the possible correlation estimates between WM and LTM would be 0.7129 for Pm and 0.5978 for SD, respectively. The meta-analytically combined correlation estimates did not seem to substantially exceed these upper bounds. This calculation is based on Spearman's correction for correlation attenuation formula (see Supplementary Figure S1).

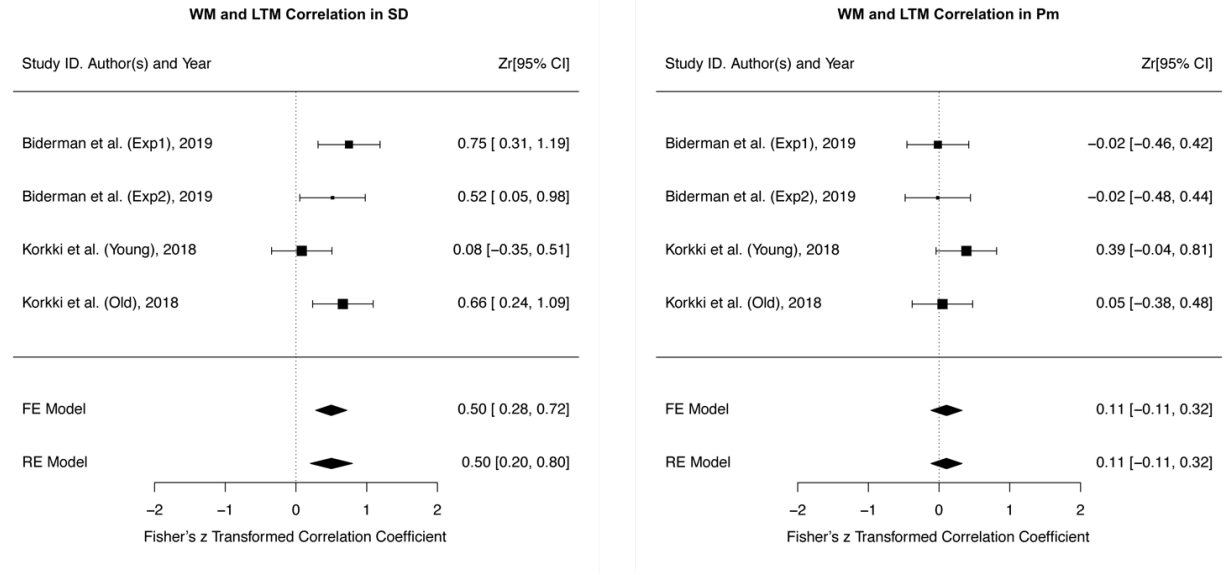
Observed $r_{x,y} = r_{x',y'} * \mathrm{Sqrt}(r_{x,x} * r_{y,y})$.

where $r_{x',y'}$ is the underlying true correlation between two variables X and Y (upper bound =1); $r_{x,x}$ represents the reliability estimate of variable X; and $r_{y,y}$ represents the reliability estimate of variable Y.

Mini Meta-Analysis of Findings in the Literature

After combining findings from additional data available in the literature (Korkki, Richter, Jeyarathnarajah, & Simons, in press) using steps outlined in Rosenthal and DiMatteo (2001), there was still a significant correlation between WM and LTM precision estimates in both fixed- and random-effect models (*p*s < .005; see Supplementary Figure S2 for meta-analyzed results). In contrast, the correlation of Pm between WM and LTM was not statistically significant in either fixed- or random-effect models (*ps* > .30). On average, the effect size of the correlation between WM and LTM precision seems to be 4 to 5 times larger than the effect size of the correlation between WM and LTM Pm. Overall, these results from a mini meta-analysis (Goh, Hall, & Rosenthal, 2016) suggest that observations from Biderman et al. should not be limited to a single dataset.

Several caveats of this mini-meta-analysis should be noted. First, the effect in Experiment 2 of Biderman et al., and the young group of Korkki et al. were attenuated, which could be due to many factors, such as different groups of subjects and different time points or experimental contexts. These factors are assumed to be taken into account when aggregating the results across studies (especially in a random-effect model). Furthermore, the results in Korkki et al. are different from Biderman et al. in many ways, including task design, subject groups (young vs. old due to their research aims in the paper), and parameter estimates (Korkki et al. used kappa as a concentration parameter instead of circular standard deviation). These could also add additional variance in the estimation of the targeted correlation.



Supplementary Figure S3. Meta-analytically combined effect sizes for the correlation between WM and LTM precision estimates (on the left) and for the correlation between WM and LTM probability successful recall (Pm, one the right). FE = Fixed-effect, RE = Random-effect.

Treatment of Within-subject Hierarchical Structure in a Between-subject Manner Biderman et al.

As indicated by the codes provided by Biderman and colleagues (copied below), they have improperly treated the within-subject structure of the data in a between-subject manner. Specifically, each subject's parameters were sampled differently across different conditions, ignoring any within-subject variances across experimental conditions (see the following code from *// subject-specific and condition-specific precision and pm*). This treatment is more appropriate for between-subject cases, for example in Souza (2016) for the comparison of age-effect across different age groups. However, this treatment is not optimal for within-subject design, because this treatment altered the hierarchical structure of the original dataset, making statistical inference inaccurate.

Codes extracted from osf.io/93cvs/ made available by Biderman et al. (2018)

```
model {
    // hyperparameters: subject-common, but vary between conditions.
    for (c in 1:n_cond) {
    mean_precision[c] ~ exponential(0.1);
    sigma_precision[c] ~ exponential(0.1);
    pm_a[c] ~ exponential(0.1);
    pm_b[c] ~ exponential(0.1);
    // subject-specific and condition-specific precision and pm
     for (s in 1:n_subjects) {
     pm[s,c] ~ beta(pm_a[c], pm_b[c]);
     precision[s,c] ~ gamma(precision_shape[c],precision_rate[c]);
     } // n_subjects
    } // n_cond

    for (n in 1:N_trials) {
     /* see Stan manual, Finite Mixtures, 13.4. "Log Sum of Exponentials:
     Linear Sums on the Log Scale" */
    target += log_sum_exp(log(pm[subject[n],cond[n]]) +
    von_mises_lpdf(response_color[n] | study_color[n], precision[subject[n],cond[n]]),
    log(1-pm[subject[n], cond[n]]) + uniform_lpdf (response_color[n] | 0, 2*pi())); }
} // model
```

Although one would argue that analyzing the within-subject dataset in a between-subject manner tends to be more conservative and should, in fact, reduce the chance of making Type-I errors, this potential benefit does not always justify its costs (Donner, Taljaard, & Klar, 2007; Lynn & McCulloch, 1992). For example, performing between-subject analyses on data arising from a repeated-measures design can lead to biases in parameter estimates in a statistical model. Furthermore, the validity of statistical inference based on altered data structure may also depend on the type of statistical models under testing (Lynn & McCulloch, 1992).

In the Bayesian framework, by breaking the within-subject constraint, the hierarchical structure of a dataset is altered, influencing both the parameter estimation and statistical inference, which critically relies on how the priors are set in the estimation procedure. It should be noted that the prior of each parameter would not be the same as the prior of the difference between parameters. It remains an empirical question to evaluate the extent to which these issues would impact results in different datasets. As this is still an emerging analytical method with standardized procedures yet to be developed, we would like to urge researchers using this method to at least retain, as authentically as possible, the original hierarchical data structure (Sorensen, Hohenstein, & Vasishth, 2016), without altering the data structure to make it more liberal or conservative than it should be for statistical inference, because neither is good for the progress of science.

References

Bays, P. M. (2018). Failure of self-consistency in the discrete resource model of visual working memory. *Cognitive Psychology*, *105*, 1–8.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. London, UK: Cambridge University Press.

Donner, A., Taljaard, M., & Klar, N. (2007). The merits of breaking the matches: a cautionary tale. *Statistics in Medicine*, *26*(9), 2036–2051.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini Meta-Analysis of Your Own Studies: Some Arguments on Why and a Primer on How. *Social and Personality Psychology Compass*, *10*(10), 535–549.

Hahn, E., Gottschling, J., & Spinath, F. M. (2012). Short measurements of personality - Validity and reliability of the GSOEP Big Five Inventory (BFI-S). *Journal of Research in Personality*, *46*(3), 355–359.

Korkki, S. M., Richter, F. R., Jeyarathnarajah, P., & Simons, J. S. (in press). Healthy ageing reduces the precision of episodic memory retrieval. *Psychology and Aging*.

Lynn, H. S., & McCulloch, C. E. (1992). When does it pay to break the matches for analysis of a matched-pairs design? *Biometrics*, *48*(2), 397–409.

Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, *1*(2), 178–183.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59–82.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion interrater reliability: Implications for validation research. *Personnel Psychology*, *55*(4), 807–825.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, *12*(3), 175–200.

Souza, A. S. (2016). No age deficits in the ability to use attention to improve visual working memory. *Psychology and Aging*, *31*(5), 456–470.