Supplemental Material for

The Development of Metacognitive Accuracy in Working Memory Across Childhood

Contents

1. Sample Size Calculations

Experiment 1

The sample size was selected based on standard group sizes in previous, related, published work in this laboratory. We retrospectively used Bayes Factors design analysis (BFDA; Stefan et al., 2019) to estimate our ability to detect an effect (estimated as Cohen's d = 0.65) of this size with N = 30 participants per group. Specifically, a Bayes Factors design analysis (BFDA) with 10000 simulations for a between-subjects design (using a Bayes Factor > 3 as a decision criteria), revealed that with 30 participants per group we could correctly detect an effect in 53.9% of samples (BF > 3), whilst 42.4% were inconclusive (0.3333 < BF < 3), and 3.7% showed evidence for the null hypothesis (BF < 0.333). With a slightly larger estimated true effect size (d = .85), 80.3% of samples showed evidence for an effect, 19.2% were inconclusive, and 0.2 % incorrectly showed evidence for the null hypothesis. In a simulated sample with no true effect (d = 0.00), and 30 participants per group, using a decision boundary of BF > 3, 1.8%

of samples showed a false positive, 43.5% provided inconclusive evidence (0.3333 < BF < 3), and 54.7% of samples correctly rejected the null hypothesis (BF < 0.3333).

We ran similar simulations to estimate our ability to find evidence for a correlation, assuming a true effect in the sample ($r = 0.30$), or no effect ($r = 0.00$). With 151 participants, 84.3% of samples showed evidence for a correlation, 13.2% were inconclusive, and 2.5% incorrectly indicated no correlation. Assuming no correlation in the population ($r = 0$), 0.6% of samples showed a false positive, 6.6% were inconclusive, and 92.9% correctly indicated evidence against a correlation. These simulated estimates indicate that 151 participants should be sufficient for results to be of interest to others in the field, regardless of the outcome.

Experiment 2

Data from fewer child participants was collected in Experiment 2 due to practical limitations combined with the observation of large effect sizes in Experiment 1. We retrospectively simulated our ability to detect an effect of meta-memory accuracy using the effect size from Experiment 1 (Cohen's *d* for the absolute meta-WM inaccuracy in the youngest children and college adults, $d = 1.57$). Using a Bayes Factors design analysis (BFDA) with 10000 simulations for a between-subjects design (using a Bayes Factor > 3 as a decision criteria), revealed that with 20 participants per group we could correctly detect an effect in 99.3% of samples (BF > 3), whilst 0.7% of samples were inconclusive. Assuming no true effect for this central contrast ($d = 0.00$, $N = 20$ per group, decision boundary of BF > 3), 2.0% of samples showed a false positive, 64.4% provided inconclusive evidence (0.3333 < BF < 3), and 33.6% of samples correctly rejected the null hypothesis (BF < 0.3333).

We ran similar simulations to estimate our ability to find evidence for a correlation, assuming a true effect in the sample ($r = 0.51$; the correlation between *k* and absolute meta-WM

inaccuracy), or no effect ($r = 0.00$). With 85 participants, 99.3% of samples showed evidence for a correlation, 0.7% were inconclusive, and 0.0% incorrectly indicated no correlation**.** Assuming no correlation in the population ($r = 0$), 0.8% of samples showed a false positive, 9.2% were inconclusive, and 90.0% correctly indicated evidence against a correlation.

2. Working Memory Capacity (*k*) Estimation

Working Memory Capacity (*k*) was estimated in this task as follows. Correct identification of sameness ($1 - f$) occurs if the probe item is in WM, or if a correct 'no change' guess is made (occurring with probability $1 - g$):

$$1 - f = \left(\frac{k}{N}\right) + \left(1 - \left(\frac{k}{N}\right)\right) \times (1 - g)$$

An incorrect indication of sameness ($1 - h$) can only occur based on guessing: $1 - h = 1 - g$. By combining these equations, *k* can be obtained:

$$k = N \times \left(\frac{h - f}{h}\right)$$

This model has been called the 'Reverse Pashler' model because Pashler (1988) developed a model for a whole-array probe situation, to which the present model seems closely related but complementary (see Cowan et al., 2013, for the first proposal of the Reverse-Pashler formula). We used this method to measure actual WM capacity across developmental stages. We sought to test how well the actual development of *k* with age would correspond to the number of items participants believed they held in mind – which would indicate awareness of WM limitations.

3. Bayesian model comparison

By comparing models that include different predictors, this procedure allows gathering evidence for or against the presence of main effects and interactions between predictors. Typically, BF values between 1 and 3 are considered 'anecdotal' evidence of the alternative hypothesis (Wetzels & Wagenmakers, 2012), *'*not worth more than a bare mention' (Jeffreys, 1961). BF greater than 3 is considered 'substantial', between 10 and 30 'strong', 30 – 100 'very strong', and over 100 'decisive' evidence (Jeffreys, 1961; Wetzels & Wagenmakers, 2012). These labels are subjective, so we advise readers to attend to the actual BF values along with the methods. Using this kind of method, collection of more evidence not only can provide support for the non-null hypothesis; unlike traditional approaches, it also can provide clear support for the null instead (BF_{01}). We used the 'anovaBF' or 'generalTestBF' functions of the *R* package *BayesFactor* to compute our models, using 10,000 iterations (the former when all factors were categorical, the latter for continuous factors).

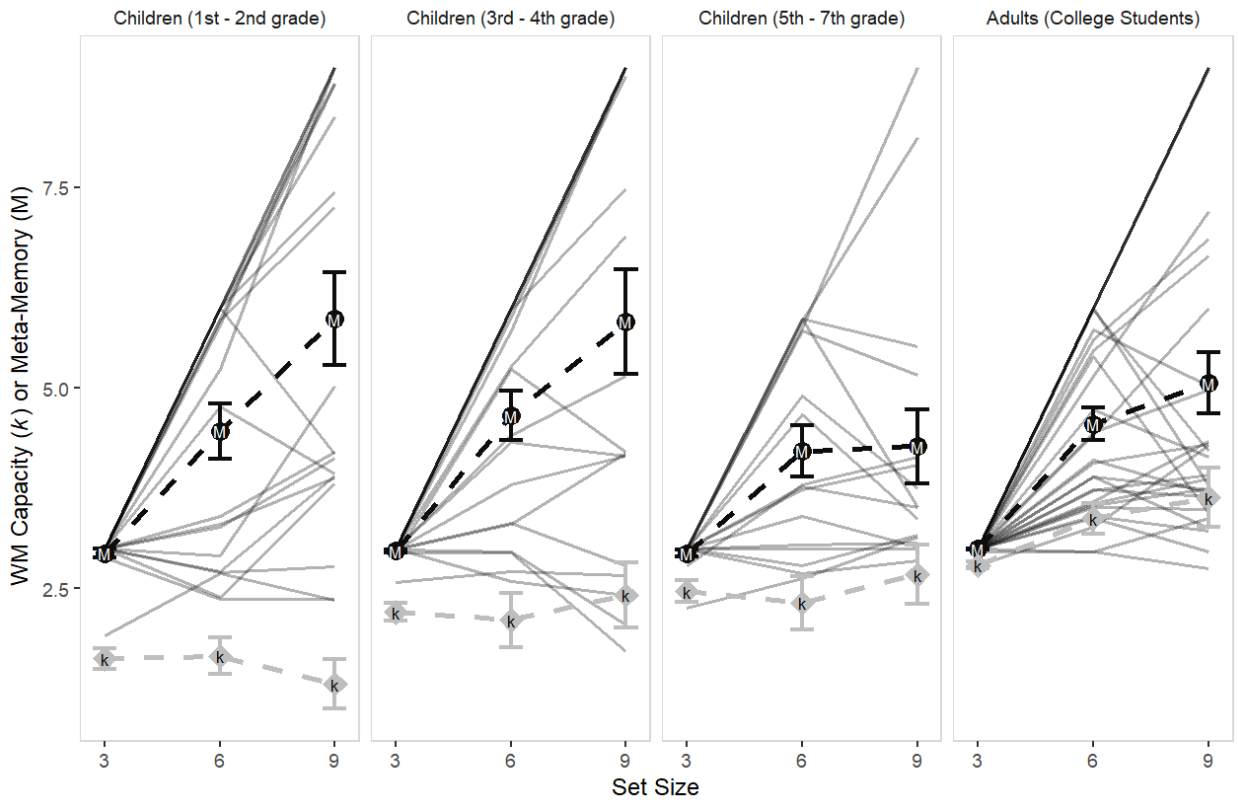

Figure S1. Experiment 2: Meta-WM judgments (M; black circles) and *k* (k; grey diamonds) across the set-sizes for each age group. Solid light grey lines represent the average meta-WM judgments in individual participants. Overlapping lines appear darker. A number of participants always used the maximum rating (1^{st} – 2^{nd} grade: 2 participants, 3^{rd} –4^{th} grade: 4 participants, 5^{th} – 7^{th} grade: 0 participants, Adults: 3 participants). Error bars represent the standard error of the mean.

Supplemental References

Cowan, N., Blume, C. L., & Saults, J. S. (2013). Attention to attributes and objects in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 731. 10.1037/a0029687

Jeffreys, H. (1961). Theory of Probability. Oxford: UK Oxford University Press.

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, *44*(4), 369-378. 10.3758/BF03210419

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E. J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042-1058. 10.3758/s13428-018-01189-8

Wetzels, R., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057-1064. 10.3758/s13423-012-0295-x