Additional analyses

Because it is theoretically interesting to see whether there is evidence for absence of a difference between certain conditions, or no effect of certain manipulations, we conducted a Bayesian reanalysis of the two Roy-Bargmann stepdown procedures. Since these analyses had not been planned, we decided to use the default priors provided in the *BayesFactor* (Morey & Rouder, 2018) package.[1]

Bayesian reanalysis of the first Roy-Bargmann procedure

As was earlier done in a frequentist setting, we first fit an ANOVA model with the total number of correct answers as the dependent variable, and the type of interpolated activity as the predictor. All effects are expressed as deviations from the estimated posterior subsample mean of 11.381. The estimated mean of the effect of content related testing is 1.254 (95% HDI = [0.553, 2.005]). The 95% highest density interval of the posterior indicates that there is a fair amount of uncertainty around the exact magnitude of the effect of content-related testing. However, most of the probability density is quite far above zero, implying that there really is a positive effect. The means of the posterior distributions for the general-knowledge-test and rereading conditions bs are -0.805 (95% HDI = [-1.549, -0.116]) and -0.449, (95% HDI = [-1.125, 0.257]) respectively. Most of the posterior distribution for the effect of general knowledge testing lies below zero, pointing to a negative effect on the total number of correct answers, although the distance is not as marked as in the content-related condition. On the other hand, there is a lot of uncertainty about the effect of rereading, compared to the other two estimates. Still, 89.8% of the posterior lies below zero, which lead us to believe that the effect is most likely negative.

Furthermore, we wanted to explore the difference between the rereading and general-knowledge-test conditions, given their somewhat similar coefficient and HDI estimates, as well as sample means. To do this, we conducted a Bayesian t-test, again with the *BayesFactor* package's default priors. The estimated posterior mean of the difference in the total number of correct answers between the two groups is -0.362 (95% HDI = [-1.49, 0.856]). As can be seen from the HDI, there is a lot of uncertainty around the estimate of the difference, which points to a lack of evidence for any claim regarding the effect.

In the second step of the Roy-Bargmann procedure, we fit an ANCOVA model with the total number of correct answers as the covariate and the total number of intrusive options chosen as the dependent variable. Effects are again expressed relative to the estimated posterior subsample mean of 4.193. There is uncertainty around the estimates of the effects of the different experimental conditions — content related testing b

[1] All posteriors obtained from 6000 simulations.

= -0.214 (95% HDI = [-0.583, 0.146]), general-knowledge testing $b = 0.072$ (95% HDI = [-0.288, 0.424]), rereading $b = 0.142$ (95% HDI = [-0.216, 0.494]). The HDIs show that there could be either a slight increase or a slight decrease in the number of intrusors, which prevented us from making a conclusion about the nature of the effects. However, given the current data and priors, we find the following — 87.43% of the posterior for the effect of content related testing falls below zero; 65.57% of the posterior for the effect of general knowledge testing falls above zero; 77.68% of the posterior for the effect of rereading falls above zero. Given the stated, there is some evidence implying that content related testing decreases the number of intrusors chosen, after controlling for the effect of the total number of correct answers. Further, there is some, albeit weaker evidence that rereading leads to an increase in the number of chosen intrusive distractors. Lastly, the posterior of the general knowledge testing effect points to no particular direction. A stronger test of these claims is desired.

Bayesian reanalysis of the second Roy-Bargmann procedure

In the second Roy-Bargmann analysis, we wanted to test whether there is an effect of the type of interpolated activity, receiving feedback, and their interaction on the total number of correct answers and chosen intrusors. Again, we first fit an ANOVA model with the two predictors and the total number of correct answers as the dependent variable.

Effects are expressed relative to the estimated posterior subsample mean of 11.868. We found that content related testing leads to an increase in the total number of correct answers, $b = 1.086$ (95% HDI = [0.589, 1.559]), compared to the general knowledge testing. This is aligned with the finding obtained in the frequentist setting. The mean of the posterior for the effect of receiving feedback is 0.218 (95% HDI = [-0.251, 0.679]). The HDI around the estimate precludes any firm conclusions regarding the effect of receiving feedback. However, we will mention that 82.25% of the posterior lies above zero, implying a possible positive effect on learning. Finally, the estimate for the interaction effect (being in the content condition and receiving feedback) is -0.013 (95% HDI = [-0.46, 0.432]). This could point to there not being a relevant interaction effect. According to the collected data and the priors, we could claim that the effect is practically equivalent to zero if we were not interested in a half-point increase or decrease in the average scores (i.e. defining a region of practical equivalence (ROPE) between [-0.5, 0.5]). Still, greater precision, which would require further data collection, is desired.

We continue with the ANCOVA model, taking the total number of correct answers as the covariate. The estimate of the intercept is 3.821 (95% HDI = [3.6, 4.03]). The estimate for the effect of content

related testing on the total number of intrusive distractors chosen is b = -0.118 (95% HDI = [-0.325, 0.092]), compared to general knowledge testing. There is some evidence for a slight decrease in the number of intrusive distractors chosen in the content related testing condition. However, an increase is also possible, but less likely and negligibly small. The estimate for the effect of receiving feedback is -0.091 (95% HDI = [-0.302, 0.121]). Although the mean of the posterior is close to zero, the lower bound of the HDI shows that values which may be considered non-negligible are still somewhat probable. Therefore, we shall refrain from making a judgement regarding the effect of feedback on choosing intrusive distractors. Finally, the estimate of the interaction effect is b = 0.047 (95% HDI = [-0.153, 0.244]). The mean of the posterior is close to zero, and we could declare the effect to be practically equivalent to zero with a ROPE of approximately [-0.25, 0.25].

As previously stated, all these analyses were not planned a priori. This warrants certain caveats. The *BayesFactor* package's default priors were used. The appropriateness of these priors should certainly be questioned. However, we decided to use them because we did not want to choose priors after already seeing the data, which would have been more problematic. Further, the statements about effects made in this section are noncommittal. Whether a 0.5 increase or decrease in the total number of correct answers is practically equivalent to zero or not is left to the reader.

References

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes Factors for Common Designs [Computer software]. Retrieved from `https://CRAN.R-project.org/package=BayesFactor` (Version 0.9.12-4.2.)