Table 1

Descriptive statistics for the DVs broken down by experimental condition.

Measure	Condition	n	M	SE_M	SD	min	max
Total correct	Content, feedback	41	13.22	0.508	3.25	2	19
	Content, no feedback	42	12.79	0.465	3.02	7	19
	General, feedback	40	10.97	0.533	3.37	1	17
	General, no feedback	40	10.47	0.449	2.84	5	16
	Rereading	40	10.88	0.443	2.80	4	17
Total intrusors	Content, feedback	41	3.15	0.258	1.65	0	7
	Content, no feedback	42	3.38	0.257	1.67	0	7
	General, feedback	40	4.17	0.318	2.01	0	8
	General, no feedback	40	4.58	0.288	1.82	1	9
	Rereading	40	4.62	0.350	2.21	1	10

Exclusion criteria

Prior to analysing the data, we have excluded participants based on a priori set criteria. Participants who have spent less than or equal to 90 seconds on the practice text were excluded (1 exclusion). Further, we wanted to exclude participants who have had no correct answers on the final test (0 exclusions). Finally, we have excluded participants who have stated that they have reading deficits (3 exclusions). This left us with a total sample of 203 participants. The descriptives for the sample are shown in Table 1.

Interpolated activity effect

Our first two hypotheses are concerned with the effects of different interpolated activities on the total number of correct answers and total number of intrusive distractors chosen. To test these hypotheses, we have focused only on the groups which have not received feedback ($n = 122$). This was done because there was no feedback option for the rereading group, and we did not want to treat the feedback and no-feedback general-knowledge and content-related testing groups as equivalent without strong evidence supporting that assumption. We conducted a one-way MANOVA with interpolated activity as the independent variable and the total number of correct and intrusive options chosen as dependent variables. The correlation between our DVs calculated on the whole sample is $r(201) = -.707$ (95% CI: [-.77, -.63], $p < .0001$).

Pillai's V for the analysis is .126, $p = .004$ (Wilks' $\Lambda = .875$, $p = .003$). The effect size, calculated as $\omega^2_{mult} = .109$ (bootstrap median[1] = .132, BC_α 95% CI = [.012, .201]). To further inspect the relationship of the interpolated activities with our dependent variables, we have conducted a Roy-Bargmann stepdown analysis,

[1] All bootstrap estimates taken from 10000 replications.

as suggested by **?** (**?**; a linear discriminant analysis with the same aim is available in the supplementary materials). The total number of correct answers was a priori chosen to be the higher priority variable. According to **?**, the higher priority variable can be chosen based on theoretical or practical grounds. Since, the total number of correct answers is the criterion that determines a student's success in a testing context, we have chosen this dependent variable as the higher priority one. Therefore, we first conducted an ANOVA with interpolated activity type as the independent variable and the total number of correct answers as the dependent variable.

As could be expected, the ANOVA points to an interpolated activity effect, with $F(2, 119) = 7.541$, $p = .001$. Following the ANOVA, we conducted an ANCOVA, with the total number of correct answers as the covariate, and the total number of intrusors as the dependent variable. The results imply a main effect of the total number of correct answers ($F(1, 118) = 79.674$, $p < .0001$), but after taking into account the number of correct answers, there is no evidence for an effect of interpolated activity on the total number of chosen intrusors ($F(2, 118) = 0.844$, $p = .433$). For now, we may claim that we do not have any evidence to support our second hypothesis that the type of interpolated activity will have an effect on the number of intrusors.

In order to test our first hypothesis, we have contrasted (i) the rereading group with the two test groups, and (ii) the two test groups with each other, taking only the total number of correct answers as the DV. The first contrast finds no evidence of a difference between the rereading group and the two test groups ($t(119) = 1.355$, $p = .178$, $g_s = 0.19$, 95% CI = [-0.19, 0.57], Cohen's $U_{3,g_s} = 57.6\%$, probability of superiority = 55.39%). However, there is a difference between the two test groups ($t(119) = 3.62$, $p = .0004$, $g_s = 0.66$, 95% CI = [0.21, 1.1], Cohen's $U_{3,g_s} = 74.43\%$, probability of superiority = 67.88%). Participants in the content related test group scored higher on the final test than participants in the general knowledge test condition. These two findings are not in line with our predictions.

The interaction between feedback and interpolated activity type

The remaining hypotheses deal with the effect of feedback on the total number of correct answers and the total number of intrusors. Therefore, these analyses are carried out only on the data from participants in the general and content related test conditions ($n = 163$). To test these hypotheses, we first conducted a two-way MANOVA with interpolated activity and feedback as independent variables, and total number of correct answers and total number of intrusors as the dependent variables.

Pillai's V for the interpolated activity effect (calculated with type III sums of squares) is .071, $p = .003$

(Wilks' $\Lambda = .929$, $p = .003$) confirming the main effect of interpolated activity type. The effect size $\omega^2_{mult} = .065$ (bootstrap median = .072, BC_α 95% CI = [.007, .139]).

On the other hand, we find no evidence for an effect of giving feedback on the linear combination of our two dependent variables — Pillai's V = .003, $p = .800$ (Wilks' $\Lambda = .997$, $p = .800$). The effect size is ω^2_{mult} = -.003 (bootstrap median = .003[2]).

Furthermore, we find no evidence for an interaction effect between activity type and feedback — Pillai's V = .001, $p = .941$ (Wilks' $\Lambda = .999$, $p = .941$). The effect size ω^2_{mult} = -.005 (bootstrap median = .003[3]). Both the feedback and the interaction estimates of ω^2_{mult} are to be considered to be zero, given their negative values.

Again, we have conducted a follow-up Roy-Bargmann stepdown analysis. In the ANOVA model with the total number of correct answers as the dependent variable and the type of interpolated activity, feedback and their interaction as predictors, only the type of activity seems to be relevant ($F(1, 159) = 11.2, p = .001$). This result also shows that participants in the content related test condition scored higher on the final test than the participants in the general knowledge test condition, which should be no surprise given the results of the first stepdown analysis. In the second step, we fit an ANCOVA model with the total number of correct answers as the covariate. In this model, the type of interpolated activity ceases to be a relevant predictor ($F(1, 155) = 0.175, p = .676$). The full models are shown in Table 2.

Contrary to our expectations, we find no evidence of an effect of feedback on the total number of correctly answered questions (**??**). Also, we found no evidence for an interaction effect of feedback and type of interpolated activity on the total number of correct answers (**??**). The same findings apply to the predictions regarding the total number of intrusors chosen (**??** and **??**).

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 have conducted a Bayesian reanalysis of the two Roy-Bargmann stepdown procedures. Since these analyses were not planned, we have decided to use the default priors provided in the *BayesFactor* (**?**) package.[4]

[2] The BC_α 95% CI for this estimate is $[-.006, .004]$.

[3] The BC_α 95% CI = $[-.006, -.005]$. Our guess is that this odd result is due to the fact that most of the density is concentrated around 0, causing an unreliable estimate. The same could be said for the CI in footnote 2.

[4] All posteriors obtained from 6000 simulations.

Table 2

ANOVA and ANCOVA models for the second Roy-Bargmann procedure.

Term	*SS*	*df*	*F*	*p*
ANOVA				
Activity	109.393	1	11.200	.001
Feedback	3.904	1	0.400	.528
Activity x Feedback	0.045	1	0.005	.946
Residuals	1553.046	159		
ANCOVA				
Activity	0.301	1	0.175	.676
Feedback	0.173	1	0.100	.752
Total correct	63.216	1	36.760	< .0001
Activity x Feedback	0.813	1	0.473	.493
Activity x Total correct	0.862	1	0.501	.480
Feedback x Total correct	0.130	1	0.075	.784
Activity x Feedback x Total correct	1.229	1	0.715	.399
Residuals	266.551	155		

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 *b*s 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, leading 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 = -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, preventing 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 find 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 prevents us from making 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 have 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. We conclude by reminding the reader that the data is available online, at `https://osf.io/gk9a3/?view_only=33f2a1207acf4e2093d489e46063b6e0`.

Deviations from the preregistered analysis plan

Initially, we have planned to do a robustness check of our findings using data with an additional exclusion criterion, based on the number of times each participant has read each of the three parts of the main text. This analysis was never conducted because (i) applying this criterion would have lead to unacceptably low power and (ii) the participants' estimates of the number of times they have read each part were similarly distributed across all conditions. Further, we have planned to conduct a TOST procedure to test whether there is no difference between the content-related and general-knowledge testing groups. This analysis was

not conducted because we have found a difference. A Bayesian t-test was also considered for the same comparison, but was dropped early on due to some conceptual concerns.