Participants and design

Undergraduate and graduate phontecis and psychology students (80.8% female, median age = 21, IQR = 3, range = [18, 31], total N = 207) from the University of Zagreb participated in the study in exchange for course credit. Participants were randomly assigned to one of five groups which differed in the type of activity they engaged in between parts of the text they have read and in whether they received feedback on their intermittent test achievement or not.

Materials and procedure

Materials

Participants read a text on the evolution, ecological and biological characteristics of weeds. The text was taken from a chapter in a Croatian university-level textbook. Some sentences and passages were slightly modified, so as to avoid odd language constructions; Latin plant names were translated to Croatian, and some plants were removed from the text to make it less difficult for the target participant population. The text was divided into three parts of 874, 754, and 835 words, respectively. Additionaly, there was a practice text taken from the same chapter, but unrelated to any of the other three parts of the text (768 words).

Forty-four content related questions with four response options were generated from the presented texts. Four questions were presented after the practice text, ten after each of the first two parts (only to the participants in the content related test condition), and twenty after the third part of the text (to all participants). Starting from the second ten-question-set, the distractor options were chosen so that (a) two distractors were plausible, but unrelated to the text, and (b) one distractor was a term or concept mentioned in the previous part of the text — this was considered to be the "intrusive" option.

An example question is:

> Compared to younger weeds, older weeds:
> (a) have a stronger allelopathic effect
> (b) contain more phytotoxins
> (c) **contain less inhibitory matter**
> (d) *show greater plasticity.*

Option (c) is the correct answer, and option (d) is the intrusive distractor.

Further, twenty-four general knowledge questions were generated. These questions were presented to participants in the general knowledge test condition, after the first two parts of the text. An example general knowledge question is:

> The name of Kurt Vonnegut's famous anti-war novel is:
> (a) **Slaughterhouse Five**
> (b) All Quiet on the Western Front
> (c) A Farewell to Arms
> (d) Journey to the End of the Night.

At the beginning of the session, participants' ID, age and sex information was collected. At the end of the session, participants were asked to estimate how much of each text they have read. The texts and questions were presented on a personal computer, in an application constructed using the open source `oTree` framework (version 2.1.35, Chen, Schonger, & Wickens, 2016) for the Python programming language (version 3.6.4, October 20, 2018).

Procedure

Participants were first given a brief introduction to the study, and were encouraged to carefully read and follow the written instructions. Then, they were led to one of six compartments containing a computer, which was running a fullscreen instance of the `oTree` application with a randomly chosen experimental condition. There, participants read the informed consent form and, in case there were no questions, started the experiment.

After entering their personal information, participants were presented with instructions for their first task, which was to read the practice text at a speed that comes naturally to them. They were to click a button at the bottom of the text when they have finished reading it. Unbeknownst to the participants, the time they took to read the practice text was recorded, and used as the basis for determining the reading time limits for the remaining texts. Results of a pilot study using different participants have shown that most participants found 4 minutes to be too short, and 9 minutes too long, so we have set the lowest possible limit to 5 minutes, and the longest to 8 minutes.

Next, participants were familiarised with the interpolated activity they were going to perform during the main part of the procedure. The rereading group reread the practice text (this time with the time limit applied), the general knowledge test group answered four general knowledge questions, and the content related test group answered four questions based on the practice text.

Subjects in the rereading and general knowledge conditions also answered the four questions related to the practice text, so as to familiarise themselves with the scope and specificity level of questions that they will receive after reading the final text. All participants were told that there would be a cumulative test after the final text, examining their knowledge of the three texts following the practice text. In reality, the final test examined only the knowledge of the final text.

Participants assigned to the feedback condition also received feedback on their practice test achievement. Feedback was presented on a separate screen, which listed the questions, the participant's answers, and the correct answers in a tabular format. Incorrectly answered questions were highlighted in red, and correctly answered questions in green. After 40 seconds elapsed, a "Next" button appeared, allowing participants to proceed to the next text. By setting this cooldown period and by emphasising that there would be a cumulative test, we wanted to encourage our participants to carefully examine the feedback. The feedback was presented for maximally 60 seconds, after which the application proceeded to the next text.

After the practice round, participants proceeded to the main texts, engaging in the interpolated activities they were assigned. After the third text, all participatns were presented with twenty questions examining their knowledge of the third text. The computer recorded whether a participant correctly answered a question and whether the participant chose an intrusive distractor. This allowed us to compute our dependent variables — the total number of correct answers and the total number of intrusive distractors chosen.

Table 1: Descriptive statistics for the number of correct answers and chosen intrusors broken down by experimental condition.

Measure	Condition	n	M	SE	SD	min	max	skew	kurtosis
Total correct	Content, feedback	41	13.22	0.508	3.25	2	19	-0.800	1.503
	Content, no feedback	42	12.79	0.465	3.02	7	19	0.039	-0.775
	General, feedback	40	10.97	0.533	3.37	1	17	-0.481	0.462
	General, no feedback	40	10.47	0.449	2.84	5	16	-0.053	-0.986
	Rereading	40	10.88	0.443	2.80	4	17	-0.141	-0.253
Total intrusors	Content, feedback	41	3.15	0.258	1.65	0	7	0.292	-0.351
	Content, no feedback	42	3.38	0.257	1.67	0	7	0.203	-0.385
	General, feedback	40	4.17	0.318	2.01	0	8	0.024	-1.124
	General, no feedback	40	4.58	0.288	1.82	1	9	0.328	-0.484
	Rereading	40	4.62	0.350	2.21	1	10	0.272	-0.537

Results

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. There is another set of exclusion criteria based on the number of times the participants have read each of the three texts. These are used in robustness check analyses (see suplementary materials).

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, since there was no feedback option for the rereading group (n = 122). 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 -0.71 (95% CI: [-0.77, -0.63], $p = 4.793 \times 10^{-32}$). Boxplots for the groups in this analysis are shown in Figure 1.

Pillai's V for the analysis is 0.126, $p = 0.004$ (Wilks' $\Lambda = 0.875$, $p = 0.003$; Hotelling-Lawley's trace = 0.142, $p = 0.003$; Roy's largest root = 0.137, $p = 4.912 \times 10^{-4}$). The effect size, calculated as $\omega^2_{mult} = 0.109$ (bootstrap median[1] = 0.132, BC_α 95% CI = [0.012, 0.201]). To further inspect the relationship of the interpolated activities with our dependent variables, we have conducted a Roy-Bargmann stepdown analysis, as suggested by Tabachnick and Fidell (2012; 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. Therefore, we first conducted an ANOVA with interpolated activity type as the indepedent

[1] All bootstrap estimates taken from 10000 replications.

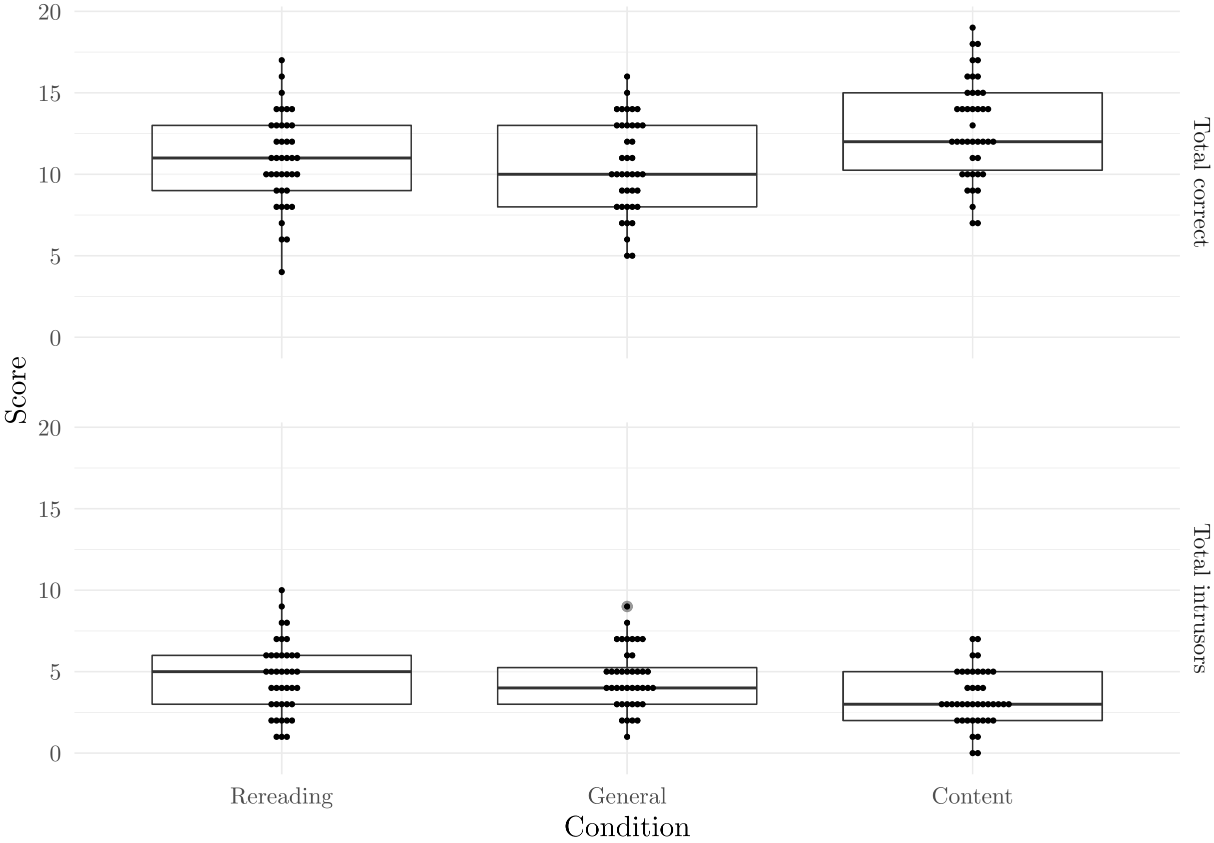

Figure 1: Boxplots broken down by experimental conditions included in the first MANOVA, and dependent variable, with overlayed raw scores.

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 $F(2,119) = 7.541$, $p = 8.254 \times 10^{-4}$. 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 ($F(1,118) = 79.674$, $p = 6.873 \times 10^{-15}$), but after taking into acount the number of correct answers, there is no evidence for an effect of interpolated activity on the total number of chosen intrusors ($F(2,118) = 0.844$, $p = 0.433$. 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 ($t = 1.355$, $p = 0.178$, $g_s = 0.19$, 95% CI = [-0.19, 0.57], Cohens's $U_{3,g_s} = 57.6\%$, probability of superiority = 55.39%). However, there is a difference between the two test groups ($t = 3.62$, $p = 4.34 \times 10^{-4}$, $g_s = 0.66$, 95% CI = [0.21, 1.1], Cohens'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 ($n = 163$). Boxplots for these groups are shown in Figure 2. 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.

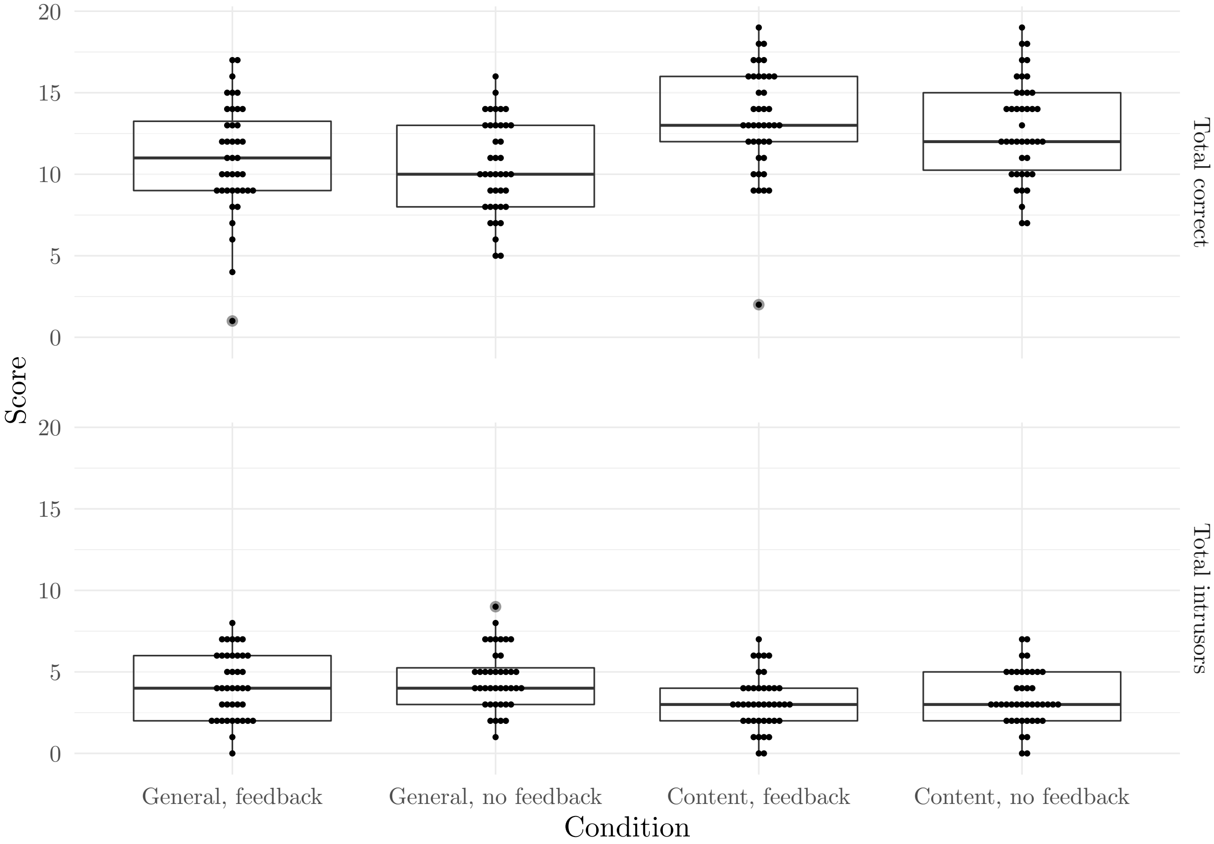

Figure 2: Boxplots broken down by experimental conditions included in the second MANOVA, and dependent variable, with overlayed raw scores.

Pillai's V for the interpolated activity effect (calculated with type III sums of squares) is 0.071, $p = 0.003$ (Wilks' $\Lambda = 0.929$, $p = 0.003$; Hotelling-Lawley's trace = 0.08, $p = 0.003$; Roy's largest root = 0.08, $p = 0.003$) confirming the main effect of interpolated activity type. The effect size $\omega^2_{mult} = 0.065$ (bootstrap median = 0.072, BC_α 95% CI = [0.007, 0.139]).

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 = 0.003, $p = 0.8$ (Wilks' $\Lambda = 0.997$, $p = 0.8$; Hotelling-Lawley's trace ≈ 0, $p = 0.8$; Roy's largest root ≈ 0, $p = 0.8$). The effect size is ω^2_{mult} = -0.003 (bootstrap median = 0.003[2]).

Furthermore, we find no evidence for an interaction effect between activity type and feedback — Pillai's V = 0.001, $p = 0.941$ (Wilks' $\Lambda = 0.999$, $p = 0.941$; Hotelling-Lawley's trace ≈ 0, $p = 0.941$; Roy's largest root

[2] The BC_α 95% CI for this estimate is $[-0.006, 0.004]$.

Table 2: Full ANOVA and ANCOVA models for the second Roy-Bargmann stepdown analysis.

Term	*SS*	*df*	*F*	*p*
ANOVA				
Activity	109.393	1	11.200	0.001
Feedback	3.904	1	0.400	0.528
Activity x Feedback	0.045	1	0.005	0.946
Residuals	1553.046	159		
ANCOVA				
Activity	0.301	1	0.175	0.676
Feedback	0.173	1	0.100	0.752
Total correct	63.216	1	36.760	0.000
Activity x Feedback	0.813	1	0.473	0.493
Activity x Total correct	0.862	1	0.501	0.480
Feedback x Total correct	0.130	1	0.075	0.784
Activity x Feedback x Total correct	1.229	1	0.715	0.399
Residuals	266.551	155		

≈ 0, $p = 0.941$). The effect size ω^2_{mult} = -0.005 (bootstrap median = 0.003[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 ($F(1, 159) = 11.2, p = 0.001$). 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 suprise 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 ($F(1, 155) = 0.175, p = 0.676$). The full models are shown in Table 2.

Additional analyses

Because it is theoretically interesting to see whether there is evidence for no 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` (Morey & Rouder, 2018) package. All posteriors obtained from 6000 simulations.

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. The mean of the posterior intercept distribution is 11.381 (95% highest density interval (HDI) = [10.863, 11.888]). The estimated mean of the *b* coefficient associated with the content-test condition is 1.254 (95% HDI = [0.553,

[3] The BC_α 95% CI = $[-0.006, -0.005]$. 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.

2.005]). The 95% highest density interval for the posterior indicates that there is a fair amount of uncertainty around the exact effect of content-related testing. However, most of the probability mass is quite far above the null value, implying that we can be certain that there really is a positive effect (given the used priors, of course). 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 the null value. However, 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 (89.8% of the posterior lies below 0).

Bayesian reanalysis of the second Roy-Bargmann procedure

Notes

Plots created using `ggplot2` (Wickham, 2016). Bootstrap conducted using the `boot` package (Canty & Ripley, 2017). Methods and analyses written using `rmarkdown` (Allaire et al., 2019) and `knitr` (Xie, 2019). The package `car` (Fox & Weisberg, 2011) was used to obtain type III sums of squares. `compute.es` (Re, 2013) was used to obtain effect sizes for contrasts. `kableExtra` was used to help generate tables (Zhu, 2019). Other utilities used are `tidyverse` (Wickham, 2017), `magrittr` (Bache & Wickham, 2014), `here` (Müller, 2017), `conflicted` (Wickham, 2018), `psych` (Revelle, 2018). Highest density intervals obtained using (Meredith & Kruschke, 2018).

References

Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., … Iannone, R. (2019). *Rmarkdown: Dynamic Documents for R.*

Bache, S. M., & Wickham, H. (2014). *Magrittr: A Forward-Pipe Operator for R.*

Canty, A., & Ripley, B. D. (2017). *Boot: Bootstrap R (S-Plus) Functions.*

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88-97.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression* (Second ed.). Thousand Oaks CA: Sage.

Meredith, M., & Kruschke, J. (2018). *HDInterval: Highest (Posterior) Density Intervals.*

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs.*

Müller, K. (2017). *Here: A Simpler Way to Find Your Files.*

Re, A. C. D. (2013). *Compute.es: Compute Effect Sizes.*

Revelle, W. (2018). *Psych: Procedures for Psychological, Psychometric, and Personality Research.* Evanston, Illinois: Northwestern University.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using Multivariate Statistics.* Pearson.

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wickham, H. (2017). *Tidyverse: Easily Install and Load the 'Tidyverse'.*

Wickham, H. (2018). *Conflicted: An Alternative Conflict Resolution Strategy.*

Xie, Y. (2019). *Knitr: A General-Purpose Package for Dynamic Report Generation in R.*

Zhu, H. (2019). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax.*