**Appendix 1:**

Table 1. Counterbalancing of number of syllables, semantic condition, and type of visual referent across lists 1 and 2. Based on the condition that the participant is assigned to, only one of the lists will receive feedback.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | *List 1*  *[Feedback = yes]* | | | *List 2*  *[Feedback = No]* | | |
| *item* | *Syllables* | *Fact* | *Type* | *Syllables* | *Fact* | *Type* |
| 1 | 2 | Observable | Plant | 2 | Observable | Plant |
| 2 | 2 | Observable | Plant | 2 | Observable | Plant |
| 3 | 2 | Observable | Animal | 2 | Observable | Animal |
| 4 | 2 | Abstract | Plant | 2 | Abstract | Plant |
| 5 | 2 | Abstract | Animal | 2 | Abstract | Animal |
| 6 | 2 | Abstract | Animal | 2 | Abstract | Animal |
| 7 | 3 | Observable | Plant | 3 | Observable | Plant |
| 8 | 3 | Observable | Animal | 3 | Observable | Animal |
| 9 | 3 | Observable | Animal | 3 | Observable | Animal |
| 10 | 3 | Abstract | Plant | 3 | Abstract | Plant |
| 11 | 3 | Abstract | Plant | 3 | Abstract | Plant |
| 12 | 3 | Abstract | Animal | 3 | Abstract | Animal |

**Appendix 2: Power analyses**

The different types of feedback and the domains of learning tested in previous studies make it difficult to estimate the effect size of providing feedback. A previous meta-analysis used a broad-brush-strokes analysis, and estimated that the effect of providing feedback relative to no feedback at all was 0.26 [1]. However, in this meta-analysis, effect sizes were found to be mediated by the study design and type of feedback. Notably, in this meta-analysis, and in one other study of vocabulary learning [2], evaluative feedback was found to be associated with small negative effect sizes. However, these tended to be in studies which offered no further opportunity to learn the material, and are not, therefore, a useful model for our study. To shed further light on effect sizes expected for vocabulary learning, we examined effect sizes for foreign language learning studies with explicit feedback that did not include the correct answer (reviewed in [3]). The mean effect sizes (Cohen’s *d*) for such studies ranged between 0.4 – 0.6, depending on the timing of the test relative to training. We also examined meta-analyses of studies where feedback was given following list learning [4]. The authors report the average effect size of providing feedback to be 0.34. In this context, we assumed that a small-medium effect size (Cohen’s *d* = 0.4 or an approximate Cohen’s *f* = 0.2) for the effect of feedback on learning was justifiable.

In order to power our analyses for hypotheses 2 & 3, we predict that the weakest effect would be the between-subject effect of semantic context on phonological learning. We consequently examined studies that examine how semantic context affects phonological learning. Here, we find that effect sizes are actually quite strong when examining explicit recall of words. Henderson et al. [5] showed a significant session x training condition (phonological/ semantic) effect, reporting η2p of 0.22 and 0.42. This corresponds to Cohen’s *f* of 0.53 – 0.85. Gladfelter and Goffman [6] also report a similar effect size (η2p = 0.19, *f* = 0.48). We therefore estimated a large effect size of *f* = 0.4 for interactions with training condition and main effects of training condition.

For the positive controls, we expected to see medium-large effects. In our previous word learning study [7] , we found a large effect associated with word length (f = 1.04, η2p = .52). These are consistent with classic reports of the item length effect in word learning. The contrast in these studies is larger (2 syllables vs 4 syllables; and a linear design of 2,3, and 4 syllables in our previous study), than proposed in the current study (2 vs. 3 syllables). Therefore, we believe it is reasonable to assume at least a medium effect in our design. We also assessed the effect sizes associated with learning observable vs. abstract facts. Van Hell et al. [8] report an effect size of *f* = 1.06, (η2p = 0.53) in their design. Again, this suggests that our estimate of a medium effect size is fairly conservative.

*A priori* analyses were conducted using G\*Power 3.1.9.2 [9]. We determined the sample size needed to detect medium effect sizes with power of 0.90 for each of our effects of interest. For repeated measures analyses, the anticipated correlation between repeated measures was set at 0.5. In our previous word learning study, correlations between different word lengths or learning conditions were > 0.7, and we would anticipate correlations in this study also being moderate-high. Table 1 shows the sample sizes needed and the corresponding power. The largest total sample size in these power analyses was 68 (34 in each training condition). We chose to use 80 participants (40 in each training condition). Consequently, our design with 80 participants should be more than adequately powered.

We also verified sample sizes for the hypotheses of experimental interest by running simulations based on expected means and standard deviations in R. We estimated that there would be a ~10% difference over the feedback and no feedback conditions. To estimate the size of the effect between training condition and recall condition, we collected pilot data from 4 participants. This suggested an enhancement of approximately 10% for the attended condition when attention was directed to semantic vs phonological conditions, which are consequently simulated in our model means. The simulations indicated we had >90% power for the contrasts of interest. This code is available on the Open Science Framework (osf.io/bdfqy). These simulations were used to generated the figure that shows our predicted pattern of results (Figure 2).



Figure 1 depicts our predicted interaction between training condition, cued recall condition, and feedback. Accuracy is on the Y axis. The X axis shows simulated performance in the phonological and semantic cued recall conditions, grouped by the phonological and semantic training conditions. Bars in green show performance for items where evaluative feedback was provided, and those in purple show performance when no feedback was provided. The black line in each box shows the median of each condition, with the box bounds representing the 1st and 3rd quartiles.

**Table 2.** Power analyses

|  | f | No. of groups | No. of measurements | Total sample size | Actual Power |
| --- | --- | --- | --- | --- | --- |
| Mixed ANOVA (2 x 2 x 2, training condition x cued recall condition x feedback**)**  DV: cued recall scores  **Main effect of feedback** | 0.2 | 2 | 2 | 68 | 0.902 |
| Mixed ANOVA (2 x 2 x 2, training condition x cued recall condition x feedback**)**  DV: cued recall scores  **Within-between interaction of training condition\* and cued recall condition** | 0.4 | 2 | 2 | 20 | 0.922 |
| **Main effect of training condition** | 0.4 | 2 | 2 | 52 | 0.904 |
| **Within-between interaction of training condition and feedback** | 0.25 | 2 | 2 | 46 | 0.912 |
| Positive control: Repeated measures ANOVA (2 x 2, training condition x word length)  DV: cued recall of phonology scores  **Main effect of word length** | 0.25 | 2 | 2 | 46 | 0.912 |
| Positive control: Repeated measures ANOVA (2 x 2, training condition x observability)  DV: cued recall of semantic features scores  **Main effect of observability** | 0.25 | 2 | 2 | 46 | 0.912 |

*\* training condition is a between-subjects factor*

**Appendix 3: Results of pilot testing**

*Method*

Participants had to learn pseudowords and semantic facts about 24 novel visual stimuli they had not encountered before, drawn from 2 matched lists comprising 12 items each. Eight participants (between the ages of 18-40 years and native English speakers) gave informed consent and completed pilot testing (four were assigned to the semantic training condition and four to the phonological training conditions). In the semantic condition, participants repeated or recalled the semantic fact they heard, whereas in the phonological condition, they were asked to repeat or recall the pseudoword. Evaluative right/wrong was given for items from one of the lists, the other items did not receive any feedback. Across participants, list order and feedback assigned to each list were fully counterbalanced. Participants were exposed to eleven blocks of training, alternating between “Repeat” and “Retrieve” blocks. However, experimenters had the option to stop after block 7 and 9 if the participants were fatigued or were at ceiling. Participants in the phonological training condition completed all 11 blocks, however, 2 participants in the semantic training condition completed only 9 blocks (they did not do the last retrieve and repeat block). After training, participants were immediately tested on their recall of both phonological forms and semantic facts, regardless of the training condition they were exposed to. The order in which cued recall of phonology and semantics was tested was also counterbalanced across participants.

*Results and discussion*

During the course of training, participants in the phonological training condition showed learning of pseudowords, and those in the semantic training condition showed learning of semantic facts (Figure 1). Participants received one further reproduction block after block 10, before they attempted the final tests. Note that two participants in the semantic condition did not complete block 10 and 11.

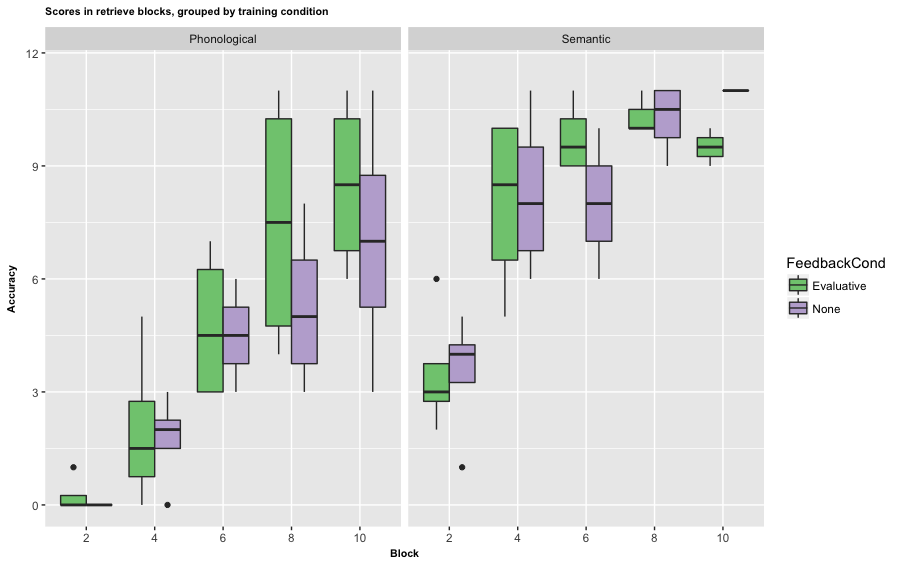


Figure 1. Boxplot showing accuracy in the retrieve blocks during training, grouped by training condition and whether feedback was provided. The black lines show median scores in each block, the upper and lower bounds of the box denote the 75th and 25th quartile respectively and the whiskers denote the maximum/minimum in the condition excluding outliers.

At the end of training, participants completed tests assessing their cued recall of semantics and phonology. This data conformed to the pattern we predicted, and also suggested that there were no ceiling effects in semantic scores. Due to the small sample size, we did not run statistical tests to assess differences.

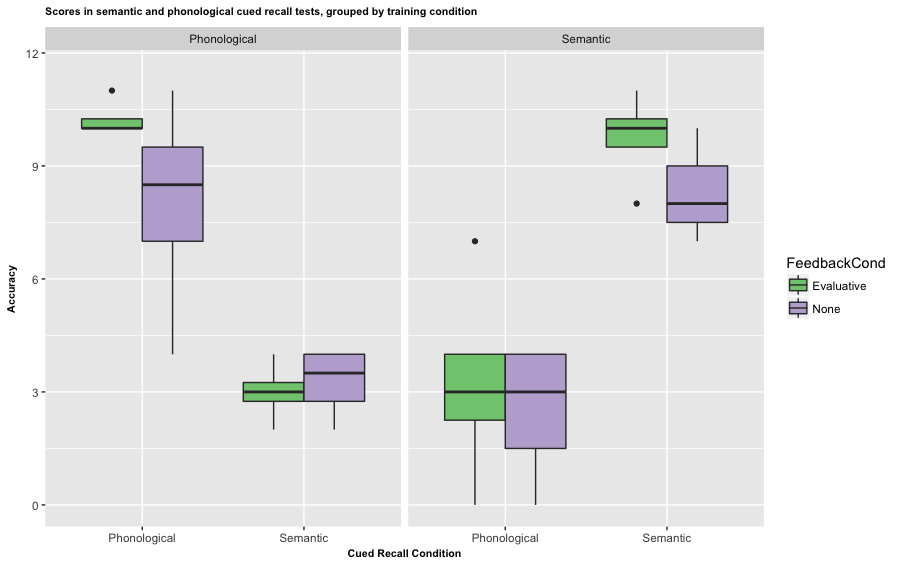


Figure 2 depicts performance in the final recall tests of phonology and semantics, grouped by training condition and whether the participants received evaluative feedback or not.

Although we initially planned to include only 6 blocks during training (with blocks 2, 4 & 6 being retrieve blocks), we find that this would lead to substantial difference in accuracy between the phonological and semantic training conditions (38.54% and 78.13% respectively). Indeed, we conducted the piloting we originally proposed in the revision with 2 participants (4 interleaved repeat and recall blocks) only to find that participants were nearly at floor, which led us to use this longer training program. However, by the 10th block, we find that participants reach a relatively similar and high level of accuracy across the two conditions (79.17% and 87.5% for the phonological and semantic training condition respectively). We therefore propose using the 11 block regime for training, with blocks 2,4,6,8, and 10 being retrieve blocks. This would allow participants 6 exposures to learn the words/ facts during training.

After 11 blocks of training, scores in the final recall tests remain high for the condition participants are trained for (76.04% and 79.17% for phonological and semantic recall in the phonological and semantic training condition respectively). However, we believe that we are unlikely to encounter ceiling effects as our final tests will be conducted one week after training. Testing participants one week later is likely to be associated with a fall of ~20% in accuracy/

*\*\* Pilot data and analysis scripts are available at* [***https://osf.io/bdfqy/***](https://osf.io/bdfqy/)

**Appendix 4: Exploratory data analyses**

1. ***Effects of error rates on the feedback x recall condition***

Given the overall difference in performance in the semantic and phonological recall conditions at final test, *F*(1,78)=74.3, *p*<.001, we assessed whether the effect of feedback was modulated by higher error rates in the phonological training condition. As we found this to be the case, we assessed whether the same pattern of learning was observed at the final test stage when only including the easier phonological learning and harder semantic learning conditions, that is, learning two syllable words and learning abstract facts alone. Here, we did not observe significant main effects of training condition, *F*(1,78)=2.1, *p*=.148, or recall, *F*(1,78)=2.5, *p*=.11, suggesting that accuracy was matched across the two conditions. We continued to observe an interaction between training condition and recall, *F*(1,78)=183.9, *p*<.001 indicating that directing attention to a specific aspect of learning did influence learning performance. Importantly, the interaction between feedback condition and recall was not significant, *F*(1,78)=0.5, *p*=.485, neither was the interaction between feedback condition and training, *F*(1,78)=0.73, *p*=.40, and we also did not observe a main effect of feedback, *F*(1,78)=.26, *p*=.61. This suggests that higher error rates during phonological learning could be a factor driving the beneficial effect of feedback. However, this dataset is reduced by half in terms of the number of trials, so it might a less powerful means of assessing the effect of feedback. We therefore conduct two follow up analyses to check whether the two levels of difficulty interact with feedback (see exploratory analyses in the main manuscript).

**References**

1. Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A. & Morgan, M. 1991 The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research* **61**, 213–238. (doi:10.3102/00346543061002213)

2. Pashler, H., Cepeda, N. J., Wixted, J. T. & Rohrer, D. 2005 When does feedback facilitate learning of words? *J Exp Psychol Learn Mem Cogn* **31**, 3–8. (doi:10.1037/0278-7393.31.1.3)

3. Li, S. 2010 The Effectiveness of Corrective Feedback in SLA: A Meta-Analysis. *Language Learning* **60**, 309–365. (doi:10.1111/j.1467-9922.2010.00561.x)

4. Kulik, J. A. & Kulik, C. 1988 Timing of feedback and verbal learning. *Review of Educational Research* **58**, 79–97.

5. Henderson, L., Weighall, A. & Gaskell, G. 2013 Learning new vocabulary during childhood: Effects of semantic training on lexical consolidation and integration. *Journal of Experimental Child Psychology* **116**, 572–592. (doi:10.1016/j.jecp.2013.07.004)

6. Gladfelter, A. & Goffman, L. 2017 Semantic richness and word learning in children with autism spectrum disorder. *Dev Sci* **59**, e12543–15. (doi:10.1111/desc.12543)

7. Krishnan, S., Watkins, K. E. & Bishop, D. V. In press. The effect of recall, reproduction, and restudy on word learning: a pre-registered study. *BMC Psychol*

8. van Hell, J. G. & Mahn, A. C. 1997 Keyword Mnemonics Versus Rote Rehearsal: Learning Concrete and Abstract Foreign Words by Experienced and Inexperienced Learners. *Language Learning* **47**, 507–546.

9. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. 2007 G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* **39**, 175–191. (doi:10.3758/BF03193146)