# Appendix S1: Learning phase performance

 There were one, three, five, and seven retrieval attempts for each word pair during the treatment in the Retrieval 1, 3, 5, and 7 groups, respectively. Table S1 summarises the number of correct responses in the four groups. For instance, the table shows that in the Retrieval 7 group, the average number of correct responses was 5.54, 8.69, 11.23, 13.50, 13.88, 14.96, and 14.58 out of 16 for each retrieval attempt. Because all four groups underwent identical experimental procedures during the period up to the first retrieval attempt (i.e., initial presentation + first retrieval), we may be able to expect all groups to have produced a similar level of retrieval success on the first attempt as long as they did not differ in their ability to learn second language (L2) vocabulary. Similarly, the Retrieval 3, 5, and 7 groups should have yielded similar levels of retrieval success on the second and third retrieval attempts, and the Retrieval 5 and 7 groups should have performed equally well on the fourth and fifth retrieval attempts provided that they were equivalent in their ability to learn L2 vocabulary. In order to test the above assumptions, three separate analyses were conducted. First, the number of correct responses obtained by the four retrieval frequency groups on the first retrieval attempt was submitted to a one-way ANOVA with the retrieval frequency level (1 / 3 / 5 / 7) as an independent variable. No statistically significant difference was found among the four groups, *F* (2, 97) = 0.21, *p* = .890, and a very small effect size was found (η2 < .01). The result suggests that the four groups did not differ significantly from each other in their degree of success on the first retrieval attempt.

 Second, for the Retrieval 3, 5, and 7 groups, the number of correct responses on the second and third retrieval attempts was submitted to a two-way 3 (retrieval frequency: 3 / 5 / 7) x 2 (retrieval attempt: 2 / 3) ANOVA. Neither the main effect of retrieval frequency, *F* (1, 71) = 0.34, *p* = .712, ηp2 = .01, nor the interaction between the retrieval frequency and retrieval attempt was significant, *F* (2, 71) = 0.61, *p* = .549, ηp2 = .02. Third, for the Retrieval 5 and 7 groups, the number of correct responses on the fourth and fifth retrieval attempts was submitted to a two-way 2 (retrieval frequency: 5 / 7) x 2 (retrieval attempt: 4 / 5) ANOVA. The ANOVA detected a significant interaction between the retrieval frequency and retrieval attempt, *F* (1, 48) = 6.35, *p* = .015, ηp2 = .12. The main effect of retrieval frequency was not significant, *F* (1, 48) = 1.28, *p* = .263, ηp2 = .03. Due to the significant interaction, the simple main effect of retrieval frequency was tested. The simple main effect of retrieval frequency was not significant on the fourth, *F* (1, 48) = 0.24, *p* = .628, and fifth retrieval attempts, *F* (1, 48) = 3.62, *p* = .063, indicating that the difference between the Retrieval 5 and 7 groups was not significant on either retrieval. Taken together, the results suggest that all four groups produced roughly similar levels of retrieval success during the period when they underwent identical experimental procedures. Thus it may be reasonable to assume that the four retrieval frequency groups did not differ significantly from each other in their ability to learn L2 vocabulary.

Table S1

*Average Number of Correct Responses During the Learning Phase*

|  |  |
| --- | --- |
| Retrieval frequency  | Retrieval attempts |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 1(*n* = 24) | 5.13  |  |  |  |   |   |   |
| *2.82*  |  |  |  |   |   |   |
| 3(*n* = 24) | 5.00  | 8.08  | 10.96  |  |  |   |   |
| *3.38*  | *3.98*  | *3.95*  |  |  |   |   |
| 5(*n* = 24) | 4.79  | 8.83  | 12.04  | 13.92  | 15.17  |  |  |
| *2.93*  | *3.41*  | *3.52*  | *2.64*  | *1.69*  |  |  |
| 7(*n* = 26) | 5.54  | 8.69  | 11.23  | 13.50  | 13.88  | 14.96  | 14.58  |
| *4.37*  | *4.58*  | *4.32*  | *3.34*  | *2.88*  | *2.18*  | *2.25*  |

*Note.* Standard deviations in italics. The maximum score for each cell is 16. Responses were scored with the strict scoring procedure.

# Appendix S2: Results of multiple comparisons for posttest scores

 Table S2 and Table S3 summarise the results of the multiple comparisons for posttest scores. For instance, Table S2 shows that with strict scoring on the productive posttest, the difference between the Retrieval 1 and 3 groups was not statistically significant (*p* = .747), and a very small effect size was observed (*d* = 0.33). The table also shows that the 95% confidence interval (hereafter, CI) of the effect size was [-0.25, 0.92], and the 95% CI of difference in the mean was [-7%, 24%].

Table S2

*Results of Multiple Comparisons on the Productive Posttest*

|  |  |  |
| --- | --- | --- |
|  |  | Strict scoring |
| Retention | Retrieval | 1 |  | 3 |  | 5 |
| Interval | frequency | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |
| Immediate | 3 | .747 | 0.33 [-0.25, 0.92] | [-7%, 24%] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 2.04 [1.33, 2.76] | [27%, 49%] |  | <.001 | 1.52 [0.86, 2.18] | [18%, 40%] |  |  |  |  |
|  | 7 | <.001 | 1.88 [1.20, 2.56] | [25%, 47%] |  | <.001 | 1.38 [0.75, 2.02] | [16%, 39%] |  | 1.000 | -0.20 [-0.77, 0.37] | [-7%, 3%] |
| 1 week | 3 | 1.000 | 0.07 [-0.52, 0.65] | [-12%, 15%] |  |  |  |   |  |  |  |  |
|  | 5 | .013 | 0.94 [0.33, 1.55] | [8%, 35%] |  | .026 | 0.88 [0.27, 1.49] | [7%, 33%] |  |  |  |  |
|  | 7 | <.001 | 1.30 [0.67, 1.92] | [18%, 45%] |  | <.001 | 1.25 [0.62, 1.87] | [16%, 44%] |  | .770 | 0.44 [-0.13, 1.02] | [-3%, 23%] |
| 4 weeks | 3 | 1.000 | 0.05 [-0.53, 0.63] | [-12%, 14%] |  |  |  |   |  |  |  |  |
|  | 5 | .007 | 0.93 [0.32, 1.54] | [9%, 39%] |  | .011 | 0.93 [0.32, 1.54] | [9%, 37%] |  |  |  |  |
|  | 7 | .004 | 1.00 [0.40, 1.61] | [11%, 39%] |  | .006 | 1.01 [0.40, 1.61] | [10%, 37%] |  | 1.000 | 0.03 [-0.54, 0.60] | [-14%, 16%] |

*Note.* Effect sizes (*d*) of 0.40, 0.70, and 1.00 indicate small, medium, and large effects, respectively (Plonsky & Oswald, 2014). [ ] indicate 95% confidence intervals for effect sizes. CI of diff. = 95% confidence intervals of difference in the means.

|  |  |  |
| --- | --- | --- |
|  |  | Sensitive scoring |
| Retention | Retrieval | 1 |  | 3 |  | 5 |
| interval | frequency | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |
| Immediate | 3 | 1.000 | 0.23 [-0.36, 0.81] | [-9%, 19%] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 1.57 [0.91, 2.23] | [17%, 38%] |  | <.001 | 1.30 [0.66, 1.94] | [12%, 32%] |  |  |  |  |
|  | 7 | <.001 | 1.53 [0.88, 2.18] | [17%, 37%] |  | <.001 | 1.25 [0.63, 1.88] | [12%, 31%] |  | 1.000 | -0.15 [-0.72, 0.42] | [-4%, 2%] |
| 1 week | 3 | 1.000 | 0.01 [-0.57, 0.59] | [-14%, 14%] |  |  |  |  |  |  |  |  |
|  | 5 | .003 | 1.04 [0.42, 1.66] | [11%, 38%] |  | .004 | 1.04 [0.42, 1.66] | [11%, 37%] |  |  |  |  |
|  | 7 | <.001 | 1.20 [0.59, 1.82] | [15%, 43%] |  | <.001 | 1.21 [0.59, 1.83] | [15%, 42%] |  | 1.000 | 0.21 [-0.36, 0.78] | [-8%, 18%] |
| 4 weeks | 3 | 1.000 | 0.06 [-0.52, 0.64] | [-12%, 15%] |  |  |  |  |  |  |  |  |
|  | 5 | .001 | 1.06 [0.44, 1.68] | [12%, 42%] |  | .002 | 1.13 [0.51, 1.76] | [13%, 39%] |  |  |  |  |
|  | 7 | <.001 | 1.15 [0.53, 1.76] | [15%, 43%] |  | .001 | 1.23 [0.61, 1.85] | [15%, 41%] |  | 1.000 | 0.07 [-0.50, 0.64] | [-13%, 16%] |

Table S3

*Results of Multiple Comparisons on the Receptive Posttest*

|  |  |  |
| --- | --- | --- |
|  |  | Strict scoring |
| Retention | Retrieval | 1 |  | 3 |  | 5 |
| interval | frequency | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |
| Immediate | 3 | 1.000 | -0.02 [-0.60, 0.57] | [-14%, 13%] |  |  |  |  |  |  |  |  |
|  | 5 | .004 | 1.07 [0.45, 1.69] | [8%, 28%] |  | .003 | 1.00 [0.38, 1.62] | [8%, 30%] |  |  |  |  |
|  | 7 | .001 | 1.21 [0.59, 1.83] | [11%, 31%] |  | <.001 | 1.13 [0.52, 1.74] | [11%, 32%] |  | 1.000 | 0.23 [-0.34, 0.81] | [-3%, 8%] |
| 1 week | 3 | 1.000 | 0.00 [-0.58, 0.58] | [-16%, 16%] |  |  |  |  |  |  |  |  |
|  | 5 | .003 | 1.08 [0.46, 1.70] | [10%, 33%] |  | .003 | 0.99 [0.37, 1.60] | [9%, 34%] |  |  |  |  |
|  | 7 | <.001 | 1.32 [0.69, 1.94] | [15%, 37%] |  | <.001 | 1.21 [0.59, 1.83] | [14%, 38%] |  | 1.000 | 0.37 [-0.20, 0.95] | [-2%, 12%] |
| 4 weeks | 3 | 1.000 | -0.01 [-0.59, 0.58] | [-14%, 14%] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 1.36 [0.72, 2.01] | [15%, 36%] |  | <.001 | 1.25 [0.61, 1.89] | [14%, 37%] |  |  |  |  |
|  | 7 | .001 | 1.08 [0.47, 1.69] | [11%, 36%] |  | .001 | 1.02 [0.42, 1.63] | [11%, 37%] |  | 1.000 | -0.10 [-0.66, 0.47] | [-12%, 9%] |

|  |  |  |
| --- | --- | --- |
|  |  | Sensitive scoring |
| Retention | Retrieval | 1 |  | 3 |  | 5 |
| interval | frequency | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |
| Immediate | 3 | 1.000 | -0.02 [-0.60, 0.56] | [-14%, 13%] |  |  |  |  |  |  |  |  |
|  | 5 | .001 | 1.20 [0.57, 1.83] | [10%, 30%] |  | .001 | 1.07 [0.45, 1.69] | [9%, 31%] |  |  |  |  |
|  | 7 | <.001 | 1.35 [0.72, 1.98] | [13%, 31%] |  | <.001 | 1.20 [0.59, 1.82] | [12%, 33%] |  | 1.000 | 0.22 [-0.35, 0.79] | [-3%, 6%] |
| 1 week | 3 | 1.000 | -0.04 [-0.62, 0.54] | [-17%, 14%] |  |  |  |  |  |  |  |  |
|  | 5 | .001 | 1.19 [0.56, 1.83] | [12%, 35%] |  | <.001 | 1.16 [0.53, 1.79] | [12%, 37%] |  |  |  |  |
|  | 7 | <.001 | 1.33 [0.70, 1.96] | [15%, 38%] |  | <.001 | 1.29 [0.67, 1.92] | [15%, 40%] |  | 1.000 | 0.22 [-0.35, 0.79] | [-4%, 9%] |
| 4 weeks | 3 | 1.000 | -0.05 [-0.63, 0.53] | [-15%, 13%] |  |  |  |  |  |  |  |  |
|  | 5 | .001 | 1.35 [0.71, 1.99] | [14%, 36%] |  | <.001 | 1.31 [0.67, 1.95] | [15%, 39%] |  |  |  |  |
|  | 7 | .001 | 1.06 [0.46, 1.67] | [11%, 37%] |  | <.001 | 1.06 [0.45, 1.67] | [12%, 39%] |  | 1.000 | -0.08 [-0.65, 0.49] | [-12%, 9%] |

# Appendix S3: Results of multiple comparisons for efficiency scores

Table S4 and Table S5 summarise the results of the multiple comparisons for efficiency scores. For instance, Table S4 shows that with strict scoring on the productive posttest, the difference between the Retrieval 1 and 3 groups was statistically significant (*p* < .001), and a large effect size was observed (*d* = 1.16). The table also shows that the 95% CI of the effect size was [0.53, 1.78], and the 95% CI of difference in the mean was [0.27, 0.83].

Table S4

*Results of Multiple Comparisons for Efficiency Scores on the Productive Posttest*

|  |  |  |
| --- | --- | --- |
|  |  | Strict scoring |
| Retention | Retrieval | 1 |  | 3 |  | 5 |
| interval | frequency | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |
| Immediate | 3 | <.001 | 1.16 [0.53, 1.78] | [0.27, 0.83] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 1.38 [0.74, 2.03] | [0.34, 0.82] |  | 1.000 | 0.11 [-0.47, 0.70] | [-0.12, 0.17] |  |  |  |  |
|  | 7 | <.001 | 1.84 [1.16, 2.52] | [0.53, 1.00] |  | .182 | 0.85 [0.26, 1.44] | [0.07, 0.36] |  | .357 | 1.57 [0.92, 2.22] | [0.12, 0.25] |
| 1 week | 3 | .059 | 0.58 [-0.02, 1.17] | [0.00, 0.55] |  |  |  |  |  |  |  |  |
|  | 5 | .162 | 0.54 [-0.05, 1.13] | [-0.02, 0.49] |  | 1.000 | -0.15 [-0.73, 0.43] | [-0.19, 0.11] |  |  |  |  |
|  | 7 | .132 | 0.55 [-0.03, 1.13] | [-0.01, 0.49] |  | 1.000 | -0.13 [-0.7, 0.44] | [-0.19, 0.12] |  | 1.000 | 0.02 [-0.55, 0.59] | [-0.11, 0.12] |
| 4 weeks | 3 | .097 | 0.57 [-0.02, 1.16] | [0.00, 0.46] |  |  |  |  |  |  |  |  |
|  | 5 | .816 | 0.36 [-0.22, 0.95] | [-0.09, 0.37] |  | 1.000 | -0.34 [-0.92, 0.25] | [-0.24, 0.06] |  |  |  |  |
|  | 7 | .134 | 0.57 [-0.01, 1.15] | [0.00, 0.43] |  | 1.000 | -0.06 [-0.63, 0.50] | [-0.16, 0.12] |  | 1.000 | 0.32 [-0.25, 0.89] | [-0.06, 0.2.00] |

*Note.* Effect sizes (*d*) of 0.40, 0.70, and 1.00 indicate small, medium, and large effects, respectively (Plonsky & Oswald, 2014). [ ] indicate 95% confidence intervals for effect sizes. CI of diff. = 95% confidence intervals of difference in the means.

|  |  |  |
| --- | --- | --- |
|  |  | Sensitive scoring |
| Retention | Retrieval | 1 |  | 3 |  | 5 |
| interval | frequency | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |
| Immediate | 3 | <.001 | 1.61 [0.94, 2.28] | [0.47, 1.00] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 2.11 [1.38, 2.83] | [0.62, 1.10] |  | 1.000 | 0.58 [-0.01, 1.17] | [0.00, 0.26] |  |  |  |  |
|  | 7 | <.001 | 2.59 [1.81, 3.36] | [0.81, 1.27] |  | .007 | 1.39 [0.75, 2.02] | [0.18, 0.44] |  | .311 | 1.79 [1.11, 2.46] | [0.12, 0.24] |
| 1 week | 3 | .001 | 0.89 [0.29, 1.50] | [0.15, 0.71] |  |  |  |  |  |  |  |  |
|  | 5 | .001 | 0.92 [0.31, 1.54] | [0.15, 0.66] |  | 1.000 | -0.09 [-0.68, 0.49] | [-0.18, 0.13] |  |  |  |  |
|  | 7 | .000 | 1.09 [0.48, 1.70] | [0.23, 0.72] |  | 1.000 | 0.16 [-0.41, 0.73] | [-0.11, 0.20] |  | 1.000 | 0.36 [-0.21, 0.94] | [-0.04, 0.18] |
| 4 weeks | 3 | .002 | 0.83 [0.22, 1.43] | [0.11, 0.61] |  |  |  |  |  |  |  |  |
|  | 5 | .018 | 0.71 [0.11, 1.31] | [0.06, 0.54] |  | 1.000 | -0.24 [-0.83, 0.34] | [-0.20, 0.08] |  |  |  |  |
|  | 7 | .001 | 0.96 [0.36, 1.56] | [0.16, 0.63] |  | 1.000 | 0.14 [-0.43, 0.71] | [-0.10, 0.17] |  | 1.000 | 0.44 [-0.13, 1.02] | [-0.03, 0.22] |

Table S5

*Results of Multiple Comparisons for Efficiency Scores on the Receptive Posttest*

|  |  |  |
| --- | --- | --- |
|  |  | Strict scoring |
| Retention | Retrieval | 1 |  | 3 |  | 5 |
| interval | frequency | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |
| Immediate | 3 | <.001 | 1.93 [1.23, 2.63] | [0.56, 1.05] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 2.63 [1.84, 3.42] | [0.74, 1.15] |  | .717 | 0.58 [-0.01, 1.18] | [0.00, 0.28] |  |  |  |  |
|  | 7 | <.001 | 3.03 [2.20, 3.87] | [0.88, 1.29] |  | .010 | 1.16 [0.54, 1.77] | [0.14, 0.42] |  | .631 | 1.16 [0.55, 1.78] | [0.07, 0.21] |
| 1 week | 3 | <.001 | 1.40 [0.75, 2.05] | [0.41, 0.99] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 1.85 [1.15, 2.54] | [0.55, 1.04] |  | 1.000 | 0.34 [-0.24, 0.93] | [-0.07, 0.26] |  |  |  |  |
|  | 7 | <.001 | 2.12 [1.41, 2.84] | [0.66, 1.15] |  | .272 | 0.74 [0.15, 1.33] | [0.05, 0.37] |  | 1.000 | 0.90 [0.30, 1.49] | [0.04, 0.18] |
| 4 weeks | 3 | <.001 | 1.56 [0.89, 2.22] | [0.41, 0.91] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 1.95 [1.24, 2.65] | [0.49, 0.91] |  | 1.000 | 0.16 [-0.42, 0.74] | [-0.11, 0.19] |  |  |  |  |
|  | 7 | <.001 | 2.33 [1.59, 3.07] | [0.65, 1.07] |  | .163 | 0.74 [0.15, 1.33] | [0.05, 0.36] |  | .466 | 0.98 [0.38, 1.59] | [0.07, 0.25] |

|  |  |  |
| --- | --- | --- |
|  |  | Sensitive scoring |
| Retention | Retrieval | 1 |  | 3 |  | 5 |
| interval | frequency | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |  | *p* | *d* | CI of diff. |
| Immediate | 3 | <.001 | 1.99 [1.28, 2.70] | [0.60, 1.09] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 2.68 [1.88, 3.48] | [0.76, 1.19] |  | .902 | 0.53 [-0.06, 1.12] | [-0.01, 0.27] |  |  |  |  |
|  | 7 | <.001 | 3.15 [2.30, 4.01] | [0.93, 1.34] |  | .008 | 1.19 [0.57, 1.81] | [0.15, 0.43] |  | .427 | 1.44 [0.80, 2.08] | [0.10, 0.22] |
| 1 week | 3 | <.001 | 1.45 [0.80, 2.10] | [0.44, 1.03] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 1.83 [1.14, 2.52] | [0.54, 1.05] |  | 1.000 | 0.23 [-0.36, 0.81] | [-0.10, 0.22] |  |  |  |  |
|  | 7 | <.001 | 2.16 [1.44, 2.88] | [0.69, 1.18] |  | .334 | 0.71 [0.13, 1.30] | [0.04, 0.36] |  | 1.000 | 1.15 [0.54, 1.76] | [0.07, 0.21] |
| 4 weeks | 3 | <.001 | 1.60 [0.93, 2.27] | [0.45, 0.96] |  |  |  |  |  |  |  |  |
|  | 5 | <.001 | 1.95 [1.25, 2.66] | [0.51, 0.95] |  | 1.000 | 0.12 [-0.47, 0.70] | [-0.12, 0.18] |  |  |  |  |
|  | 7 | <.001 | 2.33 [1.59, 3.07] | [0.68, 1.12] |  | .229 | 0.71 [0.12, 1.29] | [0.04, 0.35] |  | .467 | 1.01 [0.41, 1.62] | [0.07, 0.26] |

# Reference

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.