# Appendix A. Results of a Meta-Analysis of Previous Studies

 In order to synthesize the results of existing studies on semantic clustering, we calculated a synthetized effect size (Cohen’s *d*) of semantically related sets over unrelated sets. The following tables summarize the results of the meta-analysis.

Table S1

*Synthetized Effect Sizes for Trials to Criterion During the Learning Phase*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Study name | Cohen’s *d* | SE | Variance | Lower limit | Upper limit | *Z* | *p* |
| Higa (1963) | 0.40 | 0.11 | 0.01 | 0.19 | 0.62 | 3.65 | .000 |
| Tinkham (1993) | 0.86 | 0.24 | 0.06 | 0.39 | 1.33 | 3.61 | .000 |
| Tinkham (1997) | 0.98 | 0.22 | 0.05 | 0.54 | 1.42 | 4.40 | .000 |
| Waring (1997) | 0.85 | 0.24 | 0.06 | 0.38 | 1.32 | 3.53 | .000 |
| Random model | 0.73 | 0.16 | 0.03 | 0.41 | 1.05 | 4.46 | .000 |

Table S2

*Synthetized Effect Sizes for Posttest Scores*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Study name | Cohen’s *d* | SE | Variance | Lower limit | Upper limit | *Z* | *p* |
| Hashemi & Gowdasiaei (2005) | 1.19 | 0.21 | 0.04 | 0.78 | 1.61 | 5.65 | .000 |
| Erten & Tekin (2008) | -1.29 | 0.30 | 0.09 | -1.88 | -0.71 | -4.32 | .000 |
| Papathanasiou (2009) | -0.56 | 0.17 | 0.03 | -0.90 | -0.22 | -3.26 | .001 |
| Hoshino (2010) | 0.26 | 0.08 | 0.01 | 0.10 | 0.43 | 3.10 | .002 |
| Wilcox & Medina (2013) | -1.58 | 0.41 | 0.17 | -2.38 | -0.78 | -3.87 | .000 |
| Ishii (2015) | -0.03 | 0.07 | 0.00 | -0.16 | 0.10 | -0.46 | .646 |
| Random model | -0.24 | 0.24 | 0.06 | -0.71 | 0.23 | -1.01 | .312 |

*Note.* Studies conducted by Schneider, Healy, and Bourne (1998, 2002) are not included in the meta-analysis because they do not report standard deviations of posttest scores, and it was not possible to calculate effect sizes. Finkbeiner and Nicol (2003) are not also included because they only report the response latency on a posttest, and descriptive statistics for posttest scores (accuracy) are not reported.

Table S1 summarizes the effect sizes of semantically related sets over unrelated sets when learning is measured by trials to criterion during the learning phase. In other words, these studies compared the number of trials needed to reach the criterion of correct recalls during the learning phase and did not measure long-term retention. Table S1 shows that the synthetized effect size (Cohen’s *d*) of semantically related sets over unrelated sets in the trials-to-criterion studies is 0.73 [0.41, 1.05] (the values inside the brackets indicate 95% confidence intervals). This indicates that semantically related items required more trials than unrelated items, producing a medium-sized effect, based on the criteria proposed by Plonsky and Oswald (2014): small (*d* = 0.4), medium (*d* = 0.7), and large (*d* = 1.0).

Table S2 summarizes the effect sizes of semantically related sets over unrelated sets when learning is measured by scores on posttests administered after the learning phase. Table S2 shows that, when learning is measured by posttests administered after the learning phase, the synthetized effect size of semantic clustering is -0.24 [-0.71, 0.23]. This suggests that, although posttest scores were generally lower for related items, the effect size did not reach the criterion of a small effect (*d* =

-0.40). The very small effect size and the 95% confidence interval that crosses zero suggest that, although semantic clustering might affect the initial rate of acquisition (i.e., trials to criterion), it may not necessarily influence subsequent retention as measured by posttests.

# Appendix B. Controlling Item Difficulty of the Semantically Related and Unrelated Words

In order to ensure that the semantically related and unrelated words were controlled for item difficulty, two types of items were matched for L2- and L1-related variables that might affect the learning burden. The following table describes these variables as well as the results of the Mann-Whitney *U*-tests. As the table below shows, the Mann-Whitney *U*-tests revealed no statistically significant differences between the related and unrelated items in any of the variables. Moreover, the effect sizes were small (*r* < .24). As a result, it was assumed that the two types of words were controlled for lexical factors other than semantic relatedness that might affect learning.

Table S3

*Item Characteristics of Semantically Related and Unrelated Words*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | L2-related variables (English target word) |  | L1-related variables (Japanese translation) |
|  |  | Frequency | Length | Pronounceability |  | Frequency | Length | Familiarity | Imageability |
|  |  | BNC /COCA | BNC | Letters | Syllables | Biphoneme | Triphoneme | Positional |  | Raw | Log | Letters | Mora |
| Related　 | *M* | 9.08 | 10.00 | 6.63 | 2.25 | 0.003  | 0.0001 | 0.051 |  | 4.05 | 0.50 | 3.67 | 2.29 | 5.00 | 4.57 |
|  | *SD* | 2.80 | 3.49 | 1.41 | 0.79 | 0.003  | 0.0001 | 0.018 |  | 5.72 | 0.42 | 0.96 | 1.12 | 0.87 | 0.66 |
| Unrelated　 | *M* | 8.38 | 9.67 | 6.58 | 2.21 | 0.004  | 0.0002 | 0.052 |  | 4.21 | 0.65 | 3.33 | 2.33 | 5.17 | 4.42 |
|  | *SD* | 2.18 | 2.32 | 2.12 | 0.98 | 0.003  | 0.0002 | 0.019 |  | 2.93 | 0.26 | 0.64 | 0.76 | 0.68 | 0.91 |
| Mann–  | *U* | 246.50 | 283.00 | 274.50 | 266.50 | 248.00 | 246.00 | 283.00 |  | 209.50 | 209.00 | 260.00 | 235.00 | 270.00 | 243.00 |
| Whitney | *Z* | -0.87 | -0.10 | -0.28 | -0.48 | -0.82 | -0.87 | -0.10 |  | -1.62 | -1.63 | -0.63 | -1.18 | -0.37 | -0.93 |
|  | *p* | .39 | .92 | .78 | .63 | .41 | .39 | .92 |  | .11 | .10 | .53 | .24 | .71 | .35 |
|  | *r* | .12 | .01 | .04 | .07 | .12 | .13 | .01 |  | .23 | .24 | .09 | .17 | .05 | .13 |

*Note*. *n* = 24 for related and unrelated items. BNC/COCA = frequency level in BNC/COCA list (Nation, 2012); BNC = frequency level in BNC list (Nation, 2006); Biphoneme = average biphoneme probability (Vaden, Hickok, & Halpin, 2009); Triphoneme = average triphoneme frequency (Vaden et al., 2009); Positional = average positional probability (Vaden et al., 2009); Raw = raw frequency of the Japanese translation equivalent per million (Amano & Kondo, 1999); Log = log-transformed frequency per million; *r* = an effect size for Mann-Whitney *U*-tests.

# Appendix C. Translation Accuracy Rates (%) During Learning Phase and Posttests

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Massed |  | Spaced |
| 　 | 　 | *M* | *SD* | Min | Max | 95% CI | 　 | *M* | *SD* | Min | Max | 95% CI |
| Related Items | Pretest | 0.00 | 0.00 | 0.00 | 0.00 | - |  | 0.39 | 1.46 | 0.00 | 8.33 | [-0.01, 0.79] |
|  | Retrieval 1 | 54.23 | 24.31 | 0.00 | 91.67 | [48.06, 60.41] |  | 29.63 | 16.14 | 4.17 | 66.67 | [25.22, 34.04] |
|  | Retrieval 2 | 72.72 | 24.26 | 0.00 | 100.00 | [66.56, 78.88] |  | 32.72 | 21.04 | 4.17 | 87.50 | [26.97, 38.46] |
|  | Retrieval 3 | 81.18 | 21.84 | 0.00 | 100.00 | [75.64, 86.73] |  | 48.15 | 21.86 | 4.17 | 91.67 | [42.18, 54.12] |
|  | Immediate Posttest | 54.17 | 21.19 | 4.17 | 95.83 | [48.78, 59.55] |  | 53.24 | 23.30 | 8.33 | 91.67 | [46.88, 59.60] |
|  | Delayed Posttest | 17.74 | 13.49 | 0.00 | 50.00 | [14.31, 21.17] |  | 24.15 | 16.52 | 0.00 | 70.83 | [19.64, 28.66] |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| Unrelated Items  | Pretest | 0.34 | 1.37 | 0.00 | 8.33 | [-0.01, 0.68] |  | 0.15 | 0.79 | 0.00 | 4.17 | [-0.06, 0.37] |
|  | Retrieval 1 | 55.31 | 24.99 | 8.33 | 100.00 | [48.96, 61.66] |  | 27.70 | 17.68 | 0.00 | 70.83 | [22.87, 32.53] |
|  | Retrieval 2 | 76.01 | 22.69 | 4.17 | 100.00 | [70.25, 81.77] |  | 35.73 | 20.83 | 4.17 | 83.33 | [30.04, 41.41] |
|  | Retrieval 3 | 85.28 | 20.93 | 16.67 | 100.00 | [79.97, 90.60] |  | 51.00 | 22.34 | 4.17 | 95.83 | [44.90, 57.10] |
|  | Immediate Posttest | 54.10 | 23.60 | 4.17 | 95.83 | [48.11, 60.09] |  | 58.87 | 23.46 | 8.33 | 95.83 | [52.47, 65.28] |
|  | Delayed Posttest | 15.99 | 14.63 | 0.00 | 58.33 | [12.28, 19.71] |  | 29.09 | 18.68 | 0.00 | 79.17 | [23.99, 34.19] |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| Total | Pretest | 0.17 | 0.69 | 0.00 | 4.17 | [-0.01, 0.34] |  | 0.27 | 0.91 | 0.00 | 4.17 | [0.02, 0.52] |
|  | Retrieval 1 | 54.77 | 23.26 | 6.25 | 95.83 | [48.86, 60.68] |  | 28.67 | 15.96 | 8.33 | 64.58 | [24.31, 33.02] |
|  | Retrieval 2 | 74.36 | 22.87 | 2.08 | 100.00 | [68.55, 80.17] |  | 34.22 | 20.29 | 6.25 | 83.33 | [28.68, 39.76] |
|  | Retrieval 3 | 83.23 | 20.53 | 8.33 | 100.00 | [78.02, 88.45] |  | 49.58 | 21.30 | 6.25 | 91.67 | [43.76, 55.39] |
|  | Immediate Posttest | 54.13 | 21.20 | 4.17 | 91.67 | [48.75, 59.52] |  | 56.06 | 22.66 | 10.42 | 93.75 | [49.87, 62.24] |
| 　 | Delayed Posttest | 16.87 | 13.16 | 0.00 | 54.17 | [13.53, 20.21] | 　 | 26.62 | 16.73 | 0.00 | 75.00 | [22.05, 31.19] |

*Note*. *n* = 62 for the massed group and *n* = 54 for the spaced group. CI = confidence interval. Six participants who scored zero for both semantically related and unrelated items on at least one of the following were excluded from the analysis: Retrieval 2 during the learning phase, Retrieval 3 during the learning phase, and immediate posttest.

# Appendix D. Logistic Mixed-Effects Model of Translation Accuracy (Learning Phase)

To determine if spacing and semantic relatedness affected learning-phase performance, the translation accuracy rates during the learning phase were analyzed using a logistic mixed-effects model (mixed logit model), implemented through the lme4 software package in R (Bates, Mächler, Bolker, & Walker, 2015). The dependent variable was a binary response (correct / incorrect). Fixed effect predictors were group (massed vs. spaced) and semantic relatedness (related vs. unrelated). These fixed-effect factors were centered using deviation coding (–0.5, 0.5) to match the inferences drawn from ANOVA (Linck & Cunnings, 2015). The retrieval position (retrieval 1, 2, and 3) was also included as a fixed-effect factor with retrieval 1 as the reference level (dummy coding). To control for possible differences in translation latency, translation latency was added as a covariate in the model. The scores on the paired-associate section of the LABJT (Language Aptitude Battery for the Japanese; Sasaki, 1993) were also included as a covariate. Participants and items were treated as random effects. The models included the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013). The following table summarizes the results of the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 　 | Estimate | *SE* | *z* | *p* |
| Intercept | -0.51 | 0.15 | -3.54 | .00 |
| Group | -1.39 | 0.22 | -6.25 | .00 |
| Relatedness | -0.13 | 0.21 | -0.62 | .54 |
| Retrieval Position [Retrieval 2] | 0.82 | 0.05 | 15.64 | .00 |
| Retrieval Position [Retrieval 3] | 1.61 | 0.06 | 26.71 | .00 |
| Translation Latency | 0.68 | 0.07 | 9.41 | .00 |
| LABJT | 0.31 | 0.10 | 2.99 | .00 |
| Group × Relatedness | 0.11 | 0.20 | 0.53 | .60 |
| Group × Retrieval Position [Retrieval 2] | -0.86 | 0.09 | -9.07 | .00 |
| Group × Retrieval Position [Retrieval 3] | -0.70 | 0.10 | -6.97 | .00 |
| Relatedness × Retrieval Position [Retrieval 2] | -0.21 | 0.10 | -2.05 | .04 |
| Relatedness × Retrieval Position [Retrieval 3] | -0.29 | 0.12 | -2.42 | .02 |
| Group × Relatedness × Retrieval Position [Retrieval 2] | -0.07 | 0.19 | -0.39 | .69 |
| Group × Relatedness × Retrieval Position [Retrieval 3] | 0.23 | 0.20 | 1.16 | .25 |

*Note*. Intraclass correlation coefficients on the random effect variances (Subject and Item) were .26 and .07, respectively.

# Appendix E. Within-Set Error Rates (%) During Learning Phase and Posttests

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 　 | 　 | Massed | 　 | Spaced |
| 　 | 　 | *M* | *SD* | Min | Max | 95% CI | 　 | *M* | *SD* | Min | Max | 95% CI |
| Related Items | Retrieval 1 | 8.74 | 8.69 | 0.00 | 37.50 | [6.53, 10.94] |  | 2.31 | 2.99 | 0.00 | 12.50 | [1.50, 3.13] |
|  | Retrieval 2 | 7.93 | 8.39 | 0.00 | 29.17 | [5.80, 10.06] |  | 2.55 | 3.83 | 0.00 | 16.67 | [1.50, 3.59] |
|  | Retrieval 3 | 8.00 | 10.38 | 0.00 | 45.83 | [5.36, 10.63] |  | 3.32 | 4.31 | 0.00 | 20.83 | [2.14, 4.49] |
|  | Immediate Posttest | 3.83 | 4.67 | 0.00 | 20.83 | [2.64, 5.02] |  | 2.70 | 3.25 | 0.00 | 12.50 | [1.81, 3.59] |
|  | Delayed Posttest | 6.32 | 5.10 | 0.00 | 20.83 | [5.02, 7.61] |  | 3.70 | 4.26 | 0.00 | 16.67 | [2.54, 4.87] |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| Unrelated Items | Retrieval 1 | 3.23 | 4.30 | 0.00 | 16.67 | [2.13, 4.32] |  |  |  |  |  |  |
|  | Retrieval 2 | 3.43 | 5.77 | 0.00 | 29.17 | [1.96, 4.89] |  |  |  |  |  |  |
|  | Retrieval 3 | 1.41 | 3.11 | 0.00 | 16.67 | [0.62, 2.20] |  |  |  |  |  |  |
|  | Immediate Posttest | 0.54 | 1.92 | 0.00 | 12.50 | [0.05, 1.03] |  |  |  |  |  |  |
|  | Delayed Posttest | 0.47 | 1.70 | 0.00 | 8.33 | [0.04, 0.90] |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| Total | Retrieval 1 | 5.98 | 5.73 | 0.00 | 22.92 | [4.53, 7.44] |  |  |  |  |  |  |
|  | Retrieval 2 | 5.68 | 6.01 | 0.00 | 29.17 | [4.15, 7.21] |  |  |  |  |  |  |
|  | Retrieval 3 | 4.70 | 6.14 | 0.00 | 25.00 | [3.14, 6.26] |  |  |  |  |  |  |
|  | Immediate Posttest | 2.18 | 2.54 | 0.00 | 10.42 | [1.54, 2.83] |  |  |  |  |  |  |
| 　 | Delayed Posttest | 3.39 | 2.62 | 0.00 | 10.42 | [2.73, 4.06] | 　 |  |  |  |  |  |

*Note*. *n* = 62 for the massed group and *n* = 54 for the spaced group. The frequency of within-set errors was not calculated for the unrelated items in the spaced group. This is because, for a given unrelated word, 23 words in the spaced group can be classified as within-set errors, whereas only five words in the massed group can be classified as such. Therefore, comparing the within-set errors for unrelated items in these two groups might be misleading (for details, see the Scoring and Data Analysis section in the main text).

# Appendix F. Logistic Mixed-Effects Model of Within-Set Error Rate (Learning Phase)

To determine if semantic relatedness caused interference, the proportion of within-set errors during the learning phase was analyzed using a mixed-effects logit model. Two models were constructed: Model A compared the within-set error rates of related items between the massed and spaced groups, while Model B compared the within-set error rates between related and unrelated items in the massed group. The models included the maximal random effects structure justified by the design (Barr et al., 2013). The following tables summarize the results of the models.

Model A: Comparison of related items in massed and spaced groups

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 　 | Estimate | *SE* | *z* | *p* |
| Intercept | -3.71 | 0.18 | -20.15 | .00 |
| Group | -1.56 | 0.29 | -5.39 | .00 |
| Retrieval Position [Retrieval 2] | -0.02 | 0.16 | -0.10 | .92 |
| Retrieval Position [Retrieval 3] | 0.07 | 0.17 | 0.41 | .68 |
| Translation Latency | 0.79 | 0.12 | 6.79 | .00 |
| LABJT | -0.21 | 0.10 | -2.05 | .04 |
| Group × Retrieval Position [Retrieval 2] | 0.31 | 0.28 | 1.10 | .27 |
| Group × Retrieval Position [Retrieval 3] | 0.48 | 0.27 | 1.78 | .07 |

*Note*. Intraclass correlation coefficients on the random effect variances (Subject and Item) were .16 and .06, respectively.

Model B: Comparison of related and unrelated items in massed group

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 　 | Estimate | *SE* | *z* | *p* |
| Intercept | -3.30 | 0.16 | -20.12 | .00 |
| Relatedness | 0.73 | 0.27 | 2.71 | .01 |
| Retrieval Position [Retrieval 2] | -0.31 | 0.18 | -1.75 | .08 |
| Retrieval Position [Retrieval 3] | -1.08 | 0.26 | -4.13 | .00 |
| Translation Latency | 0.88 | 0.15 | 6.00 | .00 |
| LABJT | -0.38 | 0.15 | -2.52 | .01 |
| Group × Retrieval Position [Retrieval 2] | -0.04 | 0.30 | -0.13 | .90 |
| Group × Retrieval Position [Retrieval 3] | 1.20 | 0.38 | 3.17 | .00 |

*Note*. Intraclass correlation coefficients on the random effect variances (Subject and Item) were .13 and .04, respectively.

References

Amano, S., & Kondo, T. (1999). *Nihongo-no Goitokusei [Lexical properties of Japanese: Word familiarity].* Tokyo, Japan: Sanseido.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Erten, I. H., & Tekin, M. (2008). Effects on vocabulary acquisition of presenting new words in semantic sets versus semantically unrelated sets. *System*, *36*, 407–422.

Finkbeiner, M., & Nicol, J. (2003). Semantic category effects in second language word learning. *Applied Psycholinguistics*, *24*, 369–383.

Hashemi, M. R., & Gowdasiaei, F. (2005). An attribute-treatment interaction study: Lexical-set versus semantically-unrelated vocabulary instruction. *RELC Journal*, *36*, 341–361.

Higa, M. (1963). Interference effects of intralist word relationships in verbal learning. *Journal of Verbal Learning and Verbal Behavior*, *2*, 170–175.

Hoshino, Y. (2010). The categorical facilitation effects on L2 vocabulary learning in a classroom setting. *RELC Journal*, *41*, 301–312.

Ishii, T. (2015). Semantic connection or visual connection: Investigating the true source of confusion. *Language Teaching Research*, *19*, 712–722.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*, 185–207.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59–82.

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved March 3, 2017, from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Papathanasiou, E. (2009). An investigation of two ways of presenting vocabulary. *ELT Journal*, *63*, 313–322.

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

Sasaki, M. (1993). Relationships among second language proficiency, foreign language aptitude, and intelligence: A structural equation modeling approach. *Language Learning*, *43*, 313–344.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy & L. E. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 77–90). Mahwah, NJ: Erlbaum.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*, 419–440.

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System*, *21*, 371–380.

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research*, *13*, 138–163.

Vaden, K. I., Hickok, G. S., & Halpin, H. R. (2009). Irvine Phonotactic Online Dictionary, Version 2.0. Retrieved October 1, 2017, from http://www.iphod.com

Waring, R. (1997). The negative effects of learning words in semantic sets: A replication. *System*, *25*, 261–274.

Wilcox, A., & Medina, A. (2013). Effects of semantic and phonological clustering on L2 vocabulary acquisition among novice learners. *System*, *41*, 1056–1069.