# **\SPH\APPENDIX S1**

# **\AP1\SPACING IN THE THREE GROUPS (EXPERIMENT 1)**

The purpose of Experiment 1 was to compare the effects of part and whole learning that have equivalent spacing. This appendix explains how the trials in the whole, four-item part, and 10-item part learning groups were arranged to ensure that the treatments in the three groups would be matched for spacing.

# **\AP2\Whole Learning Group**

In the whole learning group, the 20 target word pairs were repeated in a block of 20 items and encountered five times throughout the treatment (Table 1, top). In other words, there were five cycles of 20 items, and each item was encountered only once in each cycle. As a result, encounters of a given item were separated by 19 trials on average in this group. As noted in the Experiment 1 “Treatment” section, the item order was randomized anew for each cycle. Although randomization of the item order changed spacing for individual items, it did not affect the average spacing as a whole because possible differences from the mean spacing (19 trials) were cancelled out across items in each cycle. For instance, if one item in a given cycle had spacing of 24 (19 + 5) trials, another item in the same cycle had spacing of 14 (19 – 5) trials. Consequently, the average spacing remained 19 trials even after randomization. This was also true for the four- and 10-item part learning groups below.

# **\AP2\10-Item Part Learning Group**

In the 10-item part learning group, target words were repeated in two blocks of 10 items (items 1−10 and 11−20) except in the final review, where they were practiced once in a block of 20 items. As in the whole learning group, the target word pairs were encountered five times throughout the treatment including the final review. Table 1 (middle) illustrates the item order and spacing in the 10-item part learning group. As shown in Table 1, the treatment in the 10-item part learning group consisted of the following: primacy buffers, eight cycles of 10 target items, filler cycle, final review, and recency buffers. In Cycles 1, 2, 5, and 6, items in the first block (items 1−10) were studied, and in Cycles 3, 4, 7, and 8, items in the second block (items 11−20) were studied. The arrangement of the trials in this group is based on that of the 108A group in Crothers and Suppes (1967, Experiment 9), whereby the trials for the first block were intervened by trials for the second block and vice versa.

In Table 1, the column with the heading “Spacing” indicates the average spacing for all items in a given cycle. For instance, in the 10-item part learning group, the first and second encounters of items 1 to 10 were separated by nine trials on average. Therefore, “9” is given for the spacing column in Cycle 1 (Table 1, middle). Note that the mean spacing between the second and third encounters of target words is longer (34 trials) than the spacing between the first and second encounters (nine trials). This is because the second and third encounters of the first block (Cycles 2 and 5) were separated by Cycle 3 (10 trials), Cycle 4 (10 trials), and the filler cycle (five trials), and the second and third encounters of the second block (Cycles 4 and 7) were intervened by the filler cycle (five trials), Cycle 5 (10 trials), and Cycle 6 (10 trials). As a result, 34 (9 + 10 + 10 + 5 = 34) is given for the spacing column in Cycles 2 and 4. Similarly, the mean spacing between the fourth and fifth encounters of target words was longer (29 and 19 trials for the first and second blocks, respectively) than the spacing between the third and fourth encounters (nine trials).

The average spacing in the 10-item part learning group can be calculated by averaging the mean intervening trials of Cycles 1 to 8, that is, (9 + 34 + 9 + 34 + 9 + 29 + 9 + 19) / 8 = 19 trials, which is exactly the same as in the whole learning group. As matching the average number of intervening trials is a common method of controlling spacing (e.g., Karpicke & Roediger, 2007; Logan & Balota, 2008; Pyc & Rawson, 2007), the 10-item part and 20-item whole groups are regarded as having equivalent spacing. Note that filler items are used to manipulate spacing in the 10-item part learning group. Specifically, the filler cycle after Cycle 4 intervened between the second and third encounters of target items (Table 1, middle). Without the five filler trials in the filler cycle, the average spacing between the second and third encounters would be 29 (34 – 5 = 29) trials, instead of 34. Consequently, without the filler cycle, the average spacing in the 10-item group would be (9 + 29 + 9 + 29 + 9 + 29 + 9 + 19) / 8 = 17.75 trials and no longer match that of the whole learning group (19 trials). The final review was also used to manipulate spacing in the 10-item part learning group. Without the final review, the mean spacing in this group would be shorter ([9 + 34 + 9 + 34 + 9 + 9] / 6 = 17.33 trials) than in the whole learning group (19 trials).

# **\AP2\Four-Item Part Learning Group**

In the four-item part learning group, target words were repeated in five blocks of four items (items 1−4, 5−8, 9−12, 13−16, and 17−20) except in the final review, for which a block of 20 items was used. As in the other two groups, the target word pairs were encountered five times throughout the treatment including the final review. Table 1 (bottom) presents the item order and spacing in the four-item part learning group. As illustrated in Table 1, the treatment in this group consisted of the following: primacy buffers, 20 cycles of four target items, filler cycle, final review, and recency buffers. The arrangement of the trials in this group is an extension of that of the 108A group in Crothers and Suppes (1967, Experiment 9). Note that, as in the 10-item part learning group, spacing between the second and third (43 trials) and the fourth and fifth encounters of target items (19 to 35 trials) was longer than spacing between the first and second and the third and fourth encounters (three trials).

 The average spacing in the four-item group can be calculated by averaging the mean intervening trials of Cycles 1 to 20—that is, (3 + 43 + 3 + 43 + 3 + 43 + 3 + 43 + 3 + 43 + 3 + 35 + 3 + 31 + 3 + 27 + 3 + 23 + 3 + 19) / 20 = 19 trials, which is exactly the same as in the 10-item and whole groups. Because encounters of a given item were separated by 19 trials on average in all three groups, they are regarded as being matched in spacing. Note that filler items are also used in the four-item group to manipulate spacing. Without the eight filler trials in the filler cycle, the average spacing between the second and third encounters of target words would be 35 trials (43 – 8 = 35) instead of 43. As a result, without the filler cycle, the mean spacing in this group would be (3 + 35 + 3 + 35 + 3 + 35 + 3 + 35 + 3 + 35 + 3 + 35 + 3 + 31 + 3 + 27 + 3 + 23 + 3 + 19) / 20 = 17 trials, and spacing would no longer be equivalent in the three groups. The final review was also used to manipulate spacing as in the 10-item group. Without the final review, the average spacing in the four-item group would be shorter ([3 + 43 + 3 + 43 + 3 + 43 + 3 + 43 + 3 + 43 + 3 + 3 + 3 + 3 + 3] / 15 = 16.33 trials) than in the other two groups (19 trials). By matching the mean intervening trials in the three groups, the present experiment aimed to isolate the effects of the part-whole distinction and spacing.

 At the same time, because matching the average amount of time between repetitions is another common method of controlling spacing (e.g., Kang, Lindsey, Mozer, & Pashler, 2014; Storm, Bjork, & Storm, 2010), the average amount of time between encounters was also analyzed. The analysis indicated that the three groups had roughly equivalent spacing using time as the unit of spacing rather than trial (see the Experiment 1 “Results” section).

# **\SPH\APPENDIX S2**

# **\AP1\RESULTS OF MULTIPLE COMPARISONS FOR POSTTEST SCORES**

Table S1 summarizes the results of the multiple comparisons for posttest scores. For instance, the table shows that, with strict scoring on the immediate productive posttest, the difference between the control and the four-item groups was not statistically significant (*p* = 1.000), and a small effect size was observed (*d* = .13). Effect sizes (*d*) of .20, .50, and .80 indicate small, medium, and large effects, respectively (Cohen, 1992).

**\TT\Table S1.** Results of multiple comparisons for posttest scores (Experiment 2)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Control |  | Four-item |
| \TCH\Retention interval | Posttests | Scoring | Groups | *p* | *d* |  | *p* | *d* |
| \TB\Immediate | Productive | Strict | Four-item | 1.000 | 0.13 |  |  |  |
|  |  |  | Whole | .299 | 0.49 |  | .693 | 0.32 |
|  |  | Sensitive | Four-item | 1.000 | 0.08 |  |  |  |
|  |  |  | Whole | .653 | 0.35 |  | 1.000 | 0.26 |
|  | Receptive | Strict | Four-item | .470 | 0.40 |  |  |  |
|  |  |  | Whole | .081 | 0.60 |  | 1.000 | 0.24 |
|  |  | Sensitive | Four-item | .546 | 0.38 |  |  |  |
|  |  |  | Whole | .075 | 0.60 |  | 1.000 | 0.27 |
| Delayed | Productive | Strict | Four-item | .004 | 0.87 |  |  |  |
|  |  |  | Whole | .013 | 1.02 |  | 1.000 | 0.11 |
|  |  | Sensitive | Four-item | .013 | 0.84 |  |  |  |
|  |  |  | Whole | .006 | 0.96 |  | 1.000 | 0.06 |
|  | Receptive | Strict | Four-item | .180 | 0.56 |  |  |  |
|  |  |  | Whole | .132 | 0.57 |  | 1.000 | 0.04 |
|  |  | Sensitive | Four-item | .201 | 0.54 |  |  |  |
|  |  |  | Whole | .118 | 0.58 |  | 1.000 | 0.06 |