# Element Interactivity and the Spacing Effect

Ewan Murray1, Aidan J. Horner1,2, and Silke M. Göbel1,3

1Department of Psychology, University of York, UK

2York Biomedical Research Institute, University of York, UK

3CREATE & Department of Special Needs Education, University of Oslo

# Author Note

Correspondence concerning this article should be addressed to Ewan Murray, Department of Psychology, University of York, UK, Email: [ewan.murray@york.ac.uk](mailto:ewan.murray@york.ac.uk)

# Abstract

Across two experiments I investigated whether element interactivity, the maximum number of items held in working memory at once during a task, affected the efficacy of spaced practice. The spacing schedule matched the one used in the chapter three experiments: either one massed session or three consecutive spaced sessions followed by a seven-day delay. In experiment three, I compared a linear category (low element interactivity) to a relational category (higher element interactivity). There was a significant difference in accuracy on the post-test for the low versus higher element interactivity material, however, no main effect of spacing or interaction. This was the first time I did not find a main effect of spacing during this PhD project. In experiment four I compared the linear category versus the procedure used in chapter 3 and participants completed either massed or spaced practice. This time there was a main effect of spacing, but no main effect of task nor an interaction between spacing and task. Both the procedure and category tasks benefitted from spacing. This suggested that there may have been high levels of interference between the two category types which masked the effect of spacing in Experiment 3. Neither experiment provided evidence that element interactivity moderates the spacing effect.

*Keywords*: Mathematics learning, Memory, Spacing effect, Distributed practice, Complexity, Element Interactivity

# Element Interactivity and the Spacing Effect

The two experiments presented in this chapter were designed to further investigate what aspects of task complexity affect spacing, in particular I focus on element interactivity. An overview of spacing and complexity was provided in Chapter 1 and briefly in chapters 2 and 3, in this next section I focused on element interactivity and how it differs from our previous measure of complexity.

## Spacing and Complexity: Cognitive Load and Element interactivity

Previously, I manipulated procedural complexity by varying the number of steps in an arithmetic procedure and participants completed either a massed or spaced practice schedule chapter three. I proposed that more steps meant higher procedural complexity. I found a main effect of spacing in both the two- versus three-step procedures and two- versus five-step procedure experiments, however there was no evidence of an interaction between spacing and procedural complexity. I concluded that the number of steps did not affect the efficacy of spaced practice. This may be because while learning a procedure participants can learn one step at a time and, once they have recalled and applied the step, they do not need to maintain it in working memory. Therefore, it is possible that when a task can be broken down easily into subtasks, the number of subtasks does not interact with the efficacy of spacing. In contrast to procedural complexity another suggested measure of task complexity, *element interactivity*, may affect spaced practice as it provides an alternative way to think about complexity as opposed to just more items to remember. *Element interactivity* is defined as the maximum number of elements a learner has to hold in working memory simultaneously in order to accomplish a task ([Chen et al., 2023](#ref-chen2023)). Recently, *element interactivity* has been more clearly operationalised as a measurement of task complexity ([Chen et al., 2023](#ref-chen2023)), after previous critiques suggested it was not sufficiently well defined ([Karpicke & Aue, 2015](#ref-karpicke2015)).

As described in full in chapter one, Chen et al. ([2024](#ref-chen2024)) suggest that element interactivity moderates the spacing effect. They propose that higher element interactivity material depletes working memory resources leading to a reduced quality of learning during massed practice and that spacing allows for working memory resources to be recovered during rest periods. However, they also found a spacing effect for tasks that did not deplete working memory resources, in which case they suggest thiswasbecause participants were able to rehearse. They manipulated element interactivity by comparing a novice group to an expert group on the same task, suggesting that the expert group’s prior knowledge means the material had lower element interactivity. In contrast, I aimed to improve on this design by manipulating element interactivity within subject using artificial material where prior knowledge should have a minimal effect and what I manipulated were the connections between the elements of the task.

In experiment 3, I manipulated the structure of categories. In contrast to the procedures task, when categorising a number, participants need to retain which steps they have previously checked alongside the current step they are checking, this may add an increased load to working memory and increase the interactivity between elements participants were required to learn. In the current study, using element interactivity as the complexity manipulation, I hypothesised that spacing would be less effective as element interactivity increased. I chose this hypothesis as Chen et al. ([2024](#ref-chen2024)) found conflicting results, where spacing was either found for low element interactivity material (where rehearsal could occur) or higher element interactivity (where spacing provided time for working memory resources to recover). In light of these conflicting results, I considered Donovan and Radosevich ([1999](#ref-donovan1999)) meta-analysis results to be stronger evidence, where overall task complexity was negatively correlated with the magnitude of the spacing effect. I had three hypotheses:

H1) Spaced practice will lead to a greater retention than massed retrieval

H2) Participants will have lower retention of higher element interactivity material

H3) There will be a significantly smaller effect of spacing in the higher element interactivity condition.

# Experiment three

## Method

## Participants

As I expected a similar pattern of results to the previous experiments in chapter three, I reused the simulated power analysis, which suggested I needed 34 participants per group to power for a large effect of spacing typical in the domain of spacing ( = 0.11). Thiswasonly for the main effects, however, as recent best practices have suggested much larger sample sizes for interaction effects ([Sommet et al., 2023](#ref-sommet2023)). Futhermore, I hypothesised increased complexity would attenuate the spacing effect, reducing its efficacy, but not making massed practice more effective than spaced practice. Attenuated interactions are harder to power for than a full cross over interaction ([Lakens & Caldwell, 2021](#ref-lakens2021)). The power analysis suggested that 218 participants (109 per group) were required to fully power for the expected interaction. As this material was new and time was limited at this stage in the project, I did not have the resources to fully power for the interaction. The final sample consisted of 80 undergraduate students at the University of York UK. The average age of the participants was 19.78 years (SD = 1.71 years), 68 participants identified as female, eleven as male and one as non-binary.

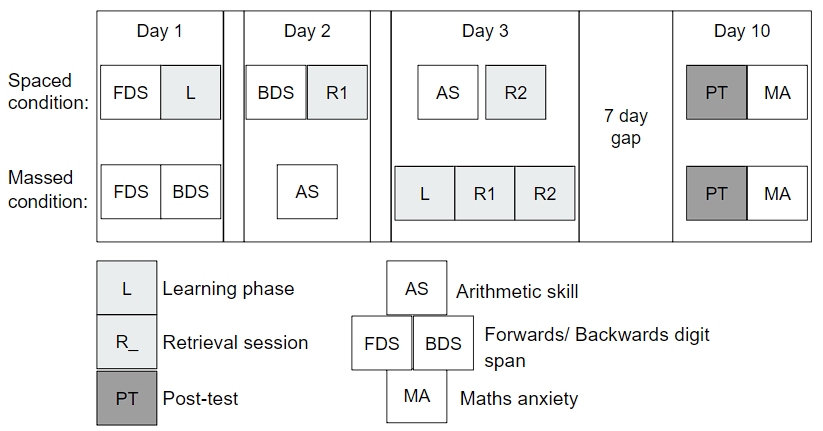
## Procedure

The University of York’s human participant pool was used for recruitment via Sona Systems (https://www.sona-systems.com/). Participants accessed the experiment through the experimental platform Gorilla (Anwyl-Irvine et al., 2020). Participants were granted course credit upon completion of the study. After signing up participants were asked their age and gender then chose when they would complete the study. Both Experiments 3 and 4 were approved by the University of York’s Ethics board, and all participants gave explicit consent to participate.

As I found large spacing effects across both experiments 1 and 2(see chapter 3) I used the same spaced practice schedule and only changed the material participants were tasked to learn. All participants who completed the experiment participated over four days: three consecutive and one session following a seven-day gap. The spacing manipulation was applied to the category learning task, and the individual differences tasks were used to pad the remaining time. As this task took longer than the main task in the Experiments 1 and 2 I omitted the maths anxiety test in Experiment 3.

Figure 1

Experimental Design (SAME AS chapter three FIG XX) DELETE AND REFER



*Note*. A diagram outlining the overall procedure of the experiment.

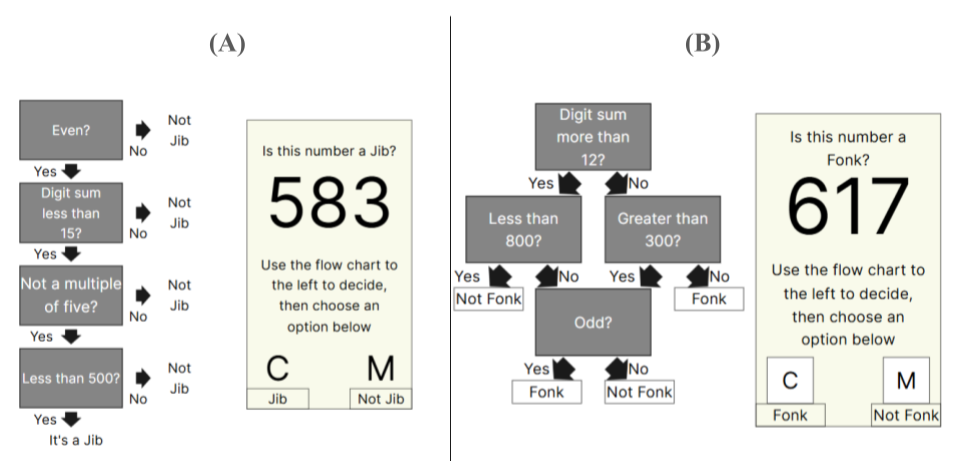
## **Material**

#### Explicit Category Learning.

During the main task, each participant learnt to categorise a number using one linear category (lower element interactivity) and one relational category (higher element interactivity) (see [Figure 2](#fig-myplot2)). Linear categories required participants to recall four constraints and, as long as all the constraints were true, the number was a member of the category (see [Figure 2](#fig-myplot2) - A). As participants were able to learn each constraint in isolation, this task was designed to have low element interactivity. Though participants were always shown the flowchart to categorise a number in the same order, the actual order they checked the constraints has no bearing mathematically. Additionally, when the target number was not a member of the category, to successfully apply the *linear* condition participants only needed to spot one constraint that did not hold (i.e., that the number was less than 500) at which point they could respond that it was not a category member. To be sure that it was a category member they needed to check all four constraints.

Figure 2

Screens during the Learning Phase



*Note*. Screens participants saw during the learning of linear (A) and relational (B) tasks.

In contrast, to apply the relational category simply checking all the steps was not sufficient. Each constraint was only useful when performed in the correct order and with knowledge of what the participant has previously checked. Due to the structure of the category (see [Figure 2](#fig-myplot2) - B), successful recall of any one constraint for the relational category would not allow participants to accurately categorise a number. This structure increased the number of elements required to be held in working memory at once, in turn making the relational category have higher element interactivity.

I ran a pilot study to check that each constraint had a similar response time, using this as a measure of difficulty and participants’ prior knowledge. I found only very small differences in response times between the constraints, except for working out the sum of the digits in a number, which took longer for participants to respond to. To minimise this issue, each category included either “Digit sum more than 10” or “Digit sum less than 15”. See [Table 1](#tbl-mytable1) for the full list of constraints.

Table 1

Constraints used to form categories

| Set one | Set two |
| --- | --- |
| Digit sum more than 12? | Digit sum less than 15? |
| Less than 800? | Even? |
| Greater than 300? | Not a multiple of five? |
| Odd? | Less than 500? |

*Note*. Table displaying the evaluative statements used to form the numerical categories, participants learned both sets, one in a linear structure one in a relational structure.

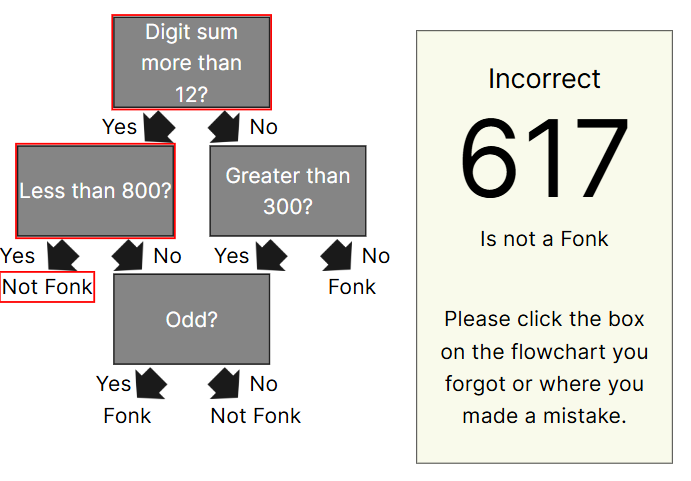
For the main task, I developed the categories to make them as similar as possible. First, it was necessary to ensure that half of the trials were members of the category and half were not, because participants either pressed a key to say the trial was a member of the category or another key to say that it was not. Therefore, answering randomly or only pressing one key throughout practice would result in 50% accuracy and I could maintain a 50% chance for both category types. Another important consideration was ensuring that if participants were performing the task as instructed, they would reach each potential end condition of the flowchart evenly. In the main task, across all trials participants checked an average of 3.25 constraints for the linear category and 2.5 constraints for the relational category, if they were performing the task as instructed.

During the practice trials participants saw the flowchart alongside the number they were required to check against (see [Figure 2](#fig-myplot2)). They then had to press C or M to categorise the number as either a member of the category or not. During each session participants completed 26 trials.

During the learning and retrieval phases of the experiment when participants got an answer incorrect they were asked to click where on the flowchart they made an error or forgot the constraint. I chose to do this as immediate feedback is a potential boundary condition of the spacing effect ([Emeny et al., 2021](#ref-emeny2021)).The main task of the post-test consisted of 24 additional trials, identical to practice, but with no feedback. Two additional tests were added for exploratory purposes, however, they provided no insights, so are only reported in the appendix (see APENDIX XX).

Figure 3

Feedback when answer was incorrect



*Note*. Example of feedback seen by participants when incorrectly categorising a number in the relational task.

#### Individual difference measures.

To enable comparisons with our previous experimental work (see chapter 3) I used the same *working memory*, *arithmetic fluency*, and *mathematics anxiety* tasks to measure participants’ individual differences. I briefly summarised them again, but please see chapter three for a full description. To measure participants’ working memory I employed a forwards and backwards digit span task based on the Wechsler memory scale - third edition (WMS-III) ([Wechsler, 1997](#ref-wechsler1997wechsler)). Arithmetic fluency was measured using the Math4Speed task ([Loenneker et al., n.d.](#ref-loenneker)) and participants answered as many addition, subtraction, multiplication and division questions as possible (two minutes per arithmetic type). Mathematics anxiety was measured using the Mathematics Anxiety Scale–UK (MAS-UK) ([Hunt et al., 2011](#ref-hunt2011)), a scale designed to measure mathematics anxiety for undergraduate students in the UK.

## Analysis

I preregistered the analysis plan (<https://osf.io/yf4nm>) and planned to run a two (spacing: massed versus spaced) by two (complexity: lower versus higher element interactivity) between-within subject Mixed ANOVA on accuracy on the post-test. I planned to run the analysis using both frequentist hypotheses testing methods and using the BayesFactor R package (Morey & Rouder, 2023) with the default Jeffreys priors to calculate a Bayes Factor for each effect. The inference criteria for the frequentist statistics were that the p-value must be less than 0.05 for the effect to be significant.

# Results

## Descriptive Statistics

This experiment had an attrition rate of 32%. Initially, 117 participants completed the first session of the experiment and 80 finished overall. Twenty participants in the spaced condition did not finish, while seventeen participants in the massed condition did not finish.

Table 2

Experiment three descriptive statistics

| Spacing | Complexity | Overall Score | Retrieval Accuracy | Working Memory | Arithmetic Fluency |
| --- | --- | --- | --- | --- | --- |
| Massed | Linear | 0.765 (0.226) | 0.872 (0.143) | 20.732 (3.860) | 78.366 (25.483) |
| Relational | 0.639 (0.203) | 0.737 (0.172) |
| Spaced | Linear | 0.762 (0.231) | 0.857 (0.136) | 19.564 (4.309) | 74.308 (24.628) |
| Relational | 0.629 (0.216) | 0.744 (0.148) |

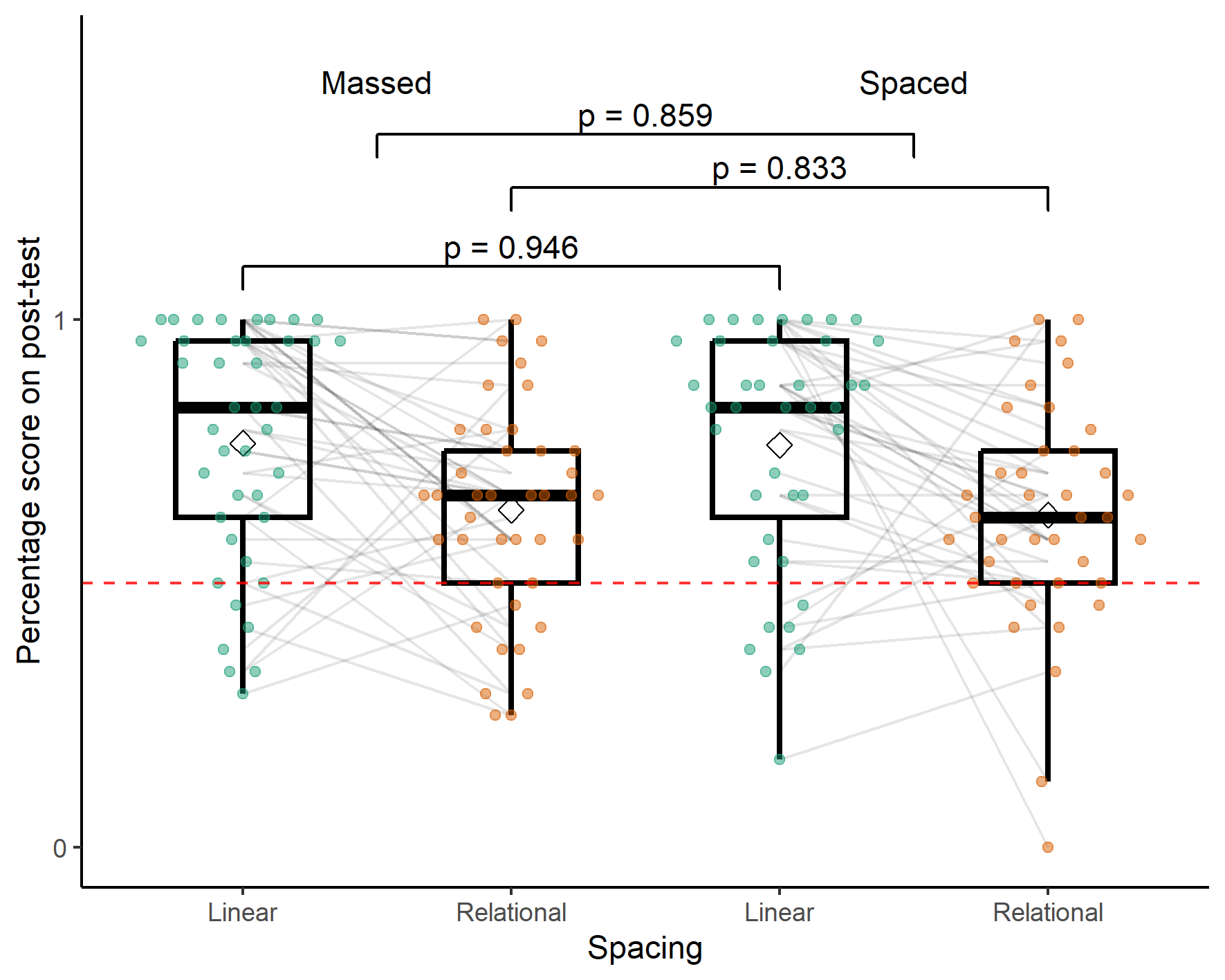
*Note*. Table displaying the mean and standard deviation for the percentage score on the post-test, retrieval accuracy (the mean percentage accuracy across both practice retrieval sessions) and the scores on the working memory, and arithmetic fluency tasks

## Main analysis

A two (spaced/massed) by two (linear/relational) mixed ANOVA was performed, with percentage accuracy on the post-test as the dependent variable. There was no significant main effect of spacing, *F*(1,78) = 0.032, *p* = 0.859, < 0.001 (BF₁₀ = 0.222). The percentage overall score on the post-test was not significantly different in the spaced (*M* = 0.696, *SD* = 0.232) than the massed condition (*M* = 0.702, *SD* = 0.223). There was a significant main effect of complexity, *F*(1,78) = 17.180, *p* < 0.001, = 0.082 (BF₁₀ = 483.040). The percentage overall score on the post-test was significantly higher in the Linear (*M* = 0.764, *SD* = 0.227) than the Relational condition (*M* = 0.634, *SD* = 0.208). There was no evidence of an interaction between spacing and complexity, *F*(1,78) = 0.011, *p* = 0.918, < 0.001 (BF₁₀ = 0.413).

Figure 4

Experiment three: Post-test performance after massed versus spaced practice by complexity condition



*Note*. A boxplot showing the percentage score on the post-test by spacing condition and complexity (linear versus relational category). The dashed line denotes chance at 50%.

## Exploratory analysis

### Individual difference measures

There were no significant differences between the two groups that may have biased the final results (see [Table 3](#tbl-ind1)).

Table 3

Experiment three exploratory variables t-test

|  | Statistic | df | p | p.adj |
| --- | --- | --- | --- | --- |
| Arithmetic Fluency | 0.72 | 77.98 | 0.47 | 1.00 |
| Retrieval Accuracy | 0.47 | 78.00 | 0.64 | 1.00 |
| Working Memory | 1.27 | 76.06 | 0.21 | 0.62 |

*Note*. Table displaying the results of t-tests to investigate if there were any significant differences between massed and spaced groups for experiment three. The p values were adjusted using the Bonferroni correction.

## Discussion

I found no evidence of a spacing effect. This contrasts with the results in chapter three where participants learnt procedures and benefited from an identical spaced practice schedule. For the linear condition, participants appeared to be performing the task properly. Performance was high and above chance. For the relational condition, participants may have not been performing the task properly or may not have understood the task’s instructions completely. There was a ceiling effect in the linear condition, however the distributions across the spacing condition were very similar. I would have hypothesised a much stronger ceiling effect in the spaced condition given the massed data.

I also ran exploratory analyses to check for relationships between the accuracy in either the cued recall or recognition tasks(see Appendix XX). No significant effects were seen in these analyses. There were no significant between group differences for working memory, retrieval accuracy or working memory. Exploratory correlation analyses suggested no significant realationships between any individual difference measure and post-test performance, as these analyses provided little theoretical insight they were not included.

If higher element interactivity is a boundary condition of the spacing effect then one reason I may have not gotten a spacing effect in either condition could be that both the linear and relational categories were too high in element interactivity. Both tasks require the participant to hold constraints in working memory while checking other constraints and could have put too high a load on working memory, in contrast to the procedures used in chapter three.

Alternatively, the constraints used in the two categories were very similar, which could have led to high levels of interference across the two conditions. In our meta-analysis, isolated material had a larger mean effect of spacing than material embedded in a course, in which recall of one item may have interfered with the recall of another (see chapter 2). If this was the case this could be an important boundary condition of the spacing effect and impact interleaving techniques, which has not previously been considered a boundary condition in the mathematics learning and spacing domain ([Emeny et al., 2021](#ref-emeny2021)).

In experiment four I compared a category learning task to a procedure learning task. If the linear category in experiment 3 was too high in element interactivity for spacing to be effective then I would expect the linear category to not benefit from spacing when it is learnt alongside a procedure in experiment 4. In contrast, if I did not get a spacing effect because of high interference then the linear category may benefit from the spacing effect in experiment four.

# Experiment Four

## Method

I used an identical experimental design to experiment three, but I changed the material participants were taught and how it was displayed. The reasons for this change are explained in the materials section below. Participants learnt one five-step procedure (adapted from experiment two) and an adapted five-constraint linear category employed in the previous experiment.

## Participants

I used the same experimental design as before, so planned to recruit a minimum of 68 participants (34 per condition) to achieve adequate power to detect the main effects but would require more than 200 participants to detect a medium interaction which was not feasible at this stage in the project. The average age of the participants was 27 years old (SD = 9.48 years). Forty-six participants identified as female and twenty-four as male. Participants were recruited through prolific and gave informed consent.

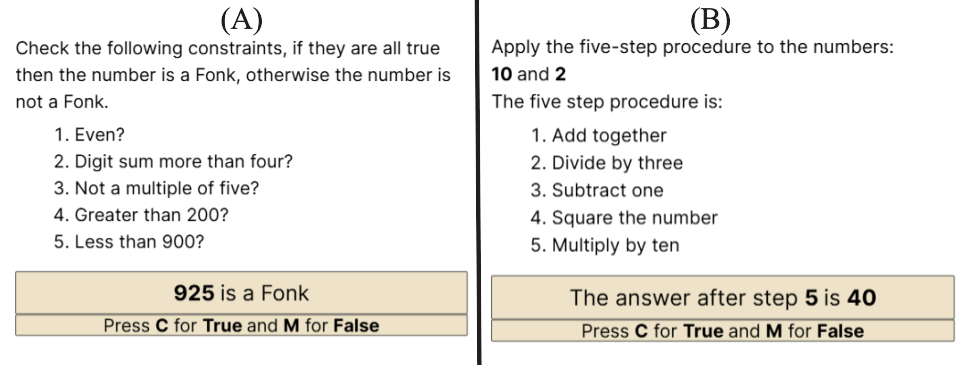
## **Material**

### Explicit Category Learning

I moved from a four-constraint category to a five-constraint category to make the task more difficult, to reduce performance, and enable any spacing effect to be viewed more clearly. In experiment three, for the four-constraint category each session had 24 trials, however, in experiment four the five-constraint category had 20 trials. I expected that decreasing the number of practice trials would reduce performance in the category condition.

Figure 5

Experiment four: Learning Phase



*Note*. Examples of the screens participants may have seen during the learning phase. In the retrieval phases and post-test participants would see the same screen without the steps/constraints available

## Learning arithmetic procedures

I used the five-step procedure from experiment two (see [Figure 5](#fig-exp2_task) (right)). It involved participants learning to apply five arithmetic steps to a pair of numbers. During the experiments in chapter three participants answered each trial by typing their answer as a number. As the category learning task has a chance level of 50%, I wanted to change how participants answered the learning arithmetic procedures task to reflect that. In the current experiment I provided participants with a number, which was either the correct or incorrect answer, and participants had to provide a binary response whether it was the correct answer or not. This change allowed for both tasks to consist of a binary choice and minimised the differences between the two tasks (see [Figure 5](#fig-exp2_task)).

Participants were asked whether the answer was correct after a certain step or after the completion of all the steps. The number of steps and constraints needed to be checked was matched across the two tasks. This allowed us to match the number of steps required to check the answer to the number of constraints participants would be required to check (if they followed the checking procedure in the order provided) and use that to classify the number without having to check in the original order.

# Results

This experiment had an attrition rate of 23%. Initially, 90 participants completed the first session of the experiment and 69 finished overall. Twelve participants in the spaced condition did not finish, while nine participants in the massed condition did not finish.

## Descriptive Statistics

Table 4

Experiment four descriptive statistics

| Spacing | Task | Overall Score | Retrieval Accuracy | Working Memory | Mathematics Anxiety | Arithmetic Fluency |
| --- | --- | --- | --- | --- | --- | --- |
| Massed | Procedure | 0.858 (0.102) | 0.921 (0.069) | 19.800 (5.556) | 46.743 (18.454) | 82.886 (27.931) |
| Category | 0.792 (0.2) | 0.893 (0.142) |
| Spaced | Procedure | 0.913 (0.083) | 0.832 (0.169) | 22.057 (4.518) | 49.829 (17.154) | 78.857 (31.746) |
| Category | 0.913 (0.121) | 0.861 (0.098) |

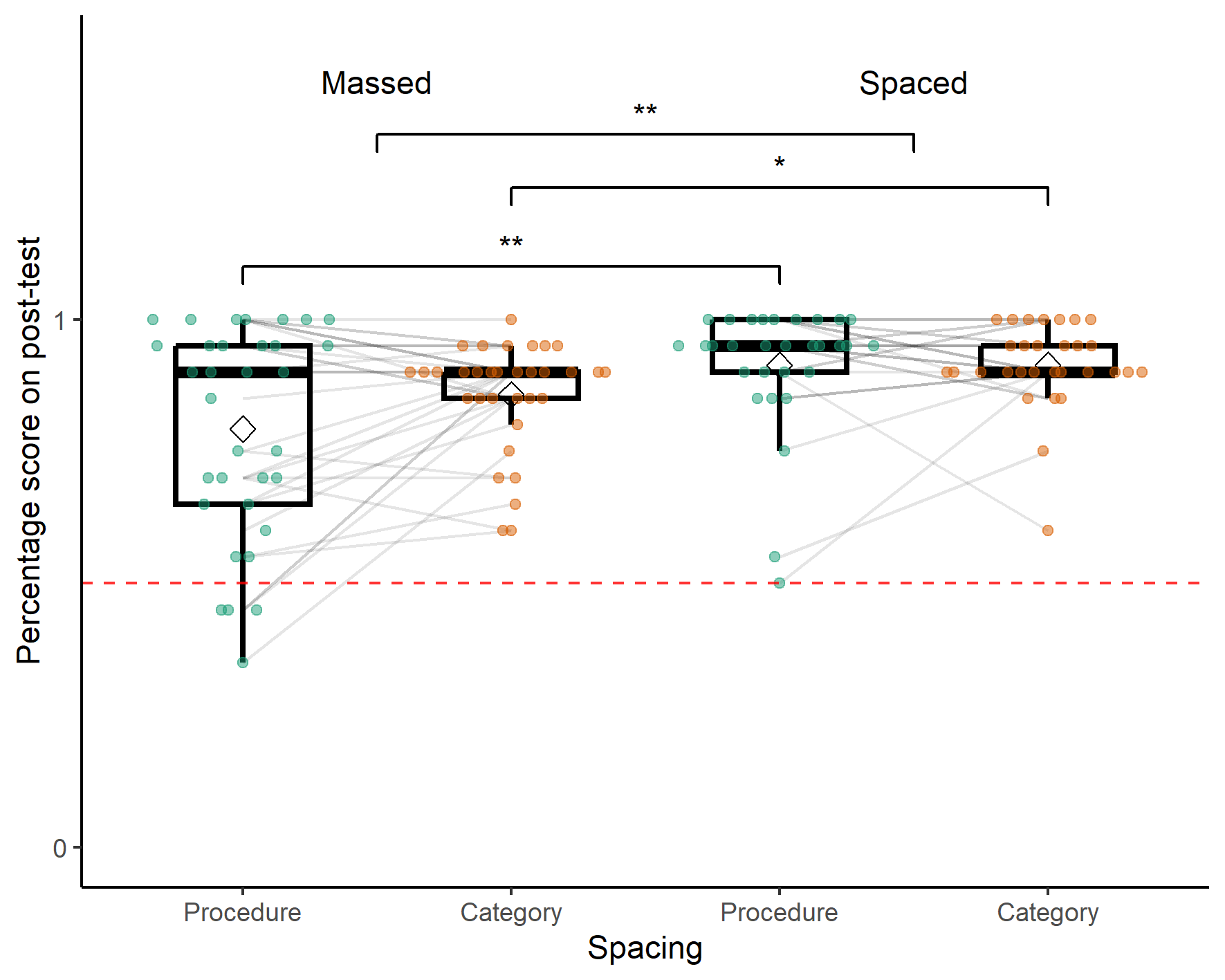
*Note*. Table displaying the mean and standard deviation for the percentage score on the post-test, retrieval accuracy (the mean percentage accuracy across both practice retrieval sessions) and the scores on the working memory, mathematics anxiety and arithmetic fluency tasks

## Main analysis

When I analysed post-test performance seven participants (five spaced, two massed) were identified as extreme outliers with low performance (SD > 3) and were removed as planned in the pre-registration. Additionally, there were concerns with heterogeneity of variance and normality of the underlying distribution. To address these concerns, I performed the planned standard mixed ANOVA (with outliers removed), followed by a robust ANOVA on the trimmed means (using the total dataset), which as an analysis is more robust to a lack of normality, heterogeneity, and outliers. First, I performed a two (spaced/massed) by two (category/procedure) mixed ANOVA, with percentage accuracy on the post-test as the dependent variable. There was a significant main effect of spacing, *F*(1,61) = 9.578, *p* = 0.003, = 0.099 (BF₁₀ = 11.617). The percentage overall score on the post-test was significantly higher in the spaced condition (*M* = 0.913, *SD* = 0.103) than the massed condition (*M* = 0.825, *SD* = 0.161). There was no significant main effect of task, *F*(1,61) = 2.997, *p* = 0.088, = 0.015 (BF₁₀ = 0.815). The percentage overall score on the post-test was numerically lower in the category (*M* = 0.850, *SD* = 0.177) than the procedure condition (*M* = 0.884, *SD* = 0.097). There was no significant interaction between spacing and task, *F*(1,61) = 2.997, *p* = 0.088, = 0.015 (BF₁₀ = 0.920). The model with the strongest Bayesian evidence was the model that included both the Spacing and Spacing:Task interaction (BF₁₀ = 12.067).

Figure 6

Experiment four: Post-test performance after massed versus spaced practice by task condition



*Note*. A boxplot showing the percentage score on the post-test by spacing and task conditions. The dashed line denotes chance at 50%. The significance stars represent: \* p < .05, \*\* p < .01, \*\*\* p < .001

Second, I performed the Robust ANOVA using the WRS2 ([Wilcox, 2012](#ref-wilcox2012introduction)) R package. These results mirrored the traditional ANOVA, showing a significant main effect of Spacing [Table 5](#tbl-mytable4).

Table 5

Experiment four - Robust ANOVA

|  | df1 | df2 | Q | p |
| --- | --- | --- | --- | --- |
| Spacing | 1 | 32.066 | 4.650 | 0.039 |
| Task | 1 | 31.790 | 1.458 | 0.236 |
| Interaction | 1 | 31.790 | 3.376 | 0.076 |

*Note*. Table displaying the results of the robust ANOVA using 20% trimmed means via the WRS2 R package

### Individual difference measures

I again tested for differences between the two groups for the individual difference measures. In contrast to experiment three, retrieval accuracy was significantly higher in the massed group during practice than the spaced group (see [Table 6](#tbl-mytable6)).

Table 6

Experiment four exploratory variables t-test

|  | Statistic | df | p | p.adj |
| --- | --- | --- | --- | --- |
| Arithmetic Fluency | 0.56 | 66.92 | 0.57 | 1.00 |
| Mathematics Anxiety | -0.72 | 67.64 | 0.47 | 1.00 |
| Retrieval Accuracy | 2.90 | 44.95 | 0.01 | 0.02 |
| Working Memory | -1.86 | 65.28 | 0.07 | 0.27 |

*Note*. Table displaying the results of t-tests to investigate if there were any significant differences between massed and spaced groups for experiment four. The p values were adjusted using the Bonferroni correction.

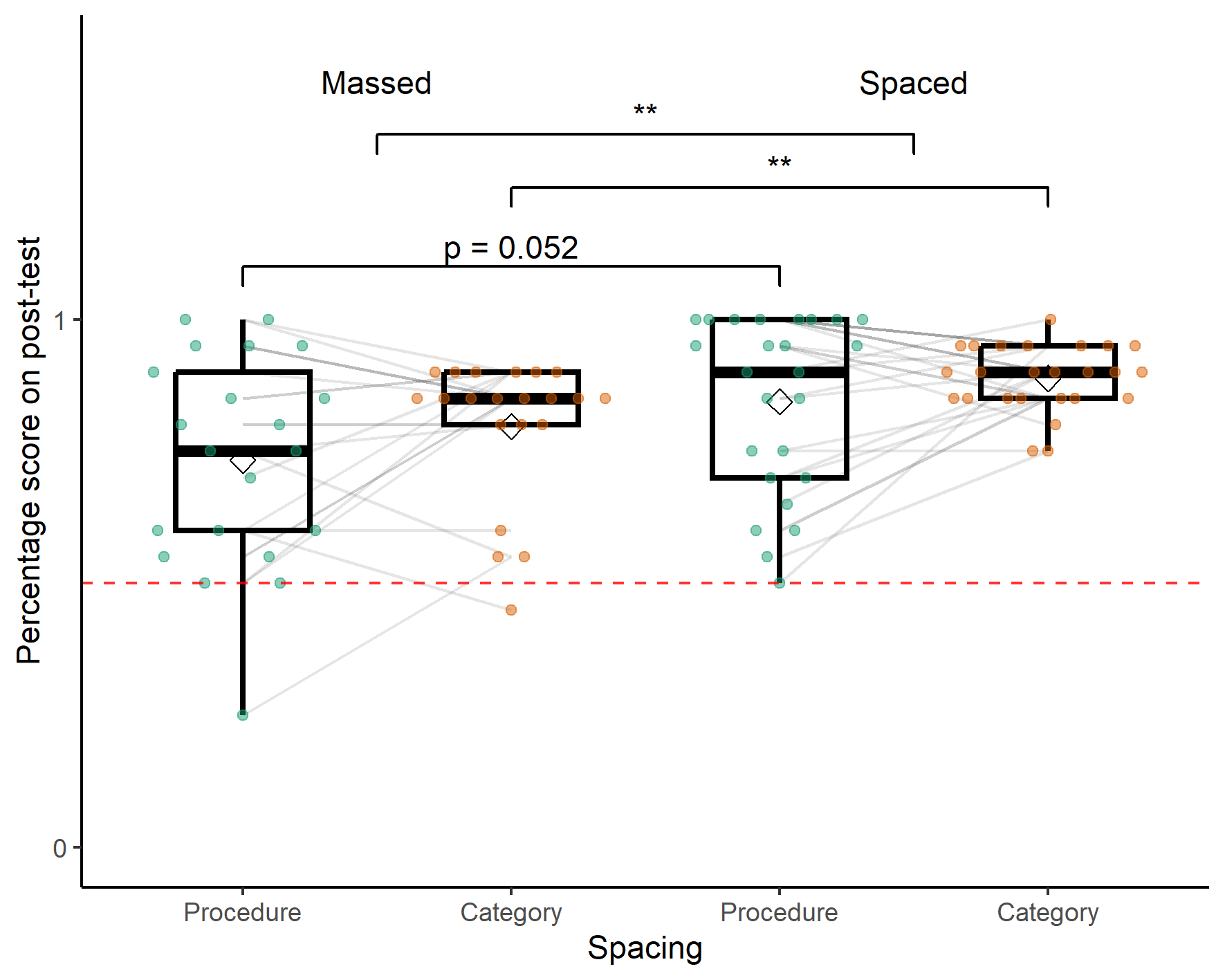
### Follow up analysis

Due to the ceiling effect in experiment four I decided to gain ethical approval to run a follow up study to track performance after a longer timescale. Two weeks after each participant finished the initial post-test they were presented with another post-test with the exact same procedure, but different numbers. Participants were told that it was optional and that they did not have to take part.Twenty-one participants originally assigned to the massed condition returned alongside twenty-five participants in the spaced condition. This group of participants had a mean age of 28.48 years (SD = 10.38) and twelve of the participants identified as male.

I ran the same analyses as I did for the initial post-test. A two (spaced/massed) by two (category/procedure) mixed ANOVA was performed, with percentage accuracy on the post-test as the dependent variable. There was a significant main effect of spacing, *F*(1,44) = 7.908, *p* = 0.007, = 0.106 (BF₁₀ = 8.057). The percentage overall score on the post-test was significantly higher in the spaced (*M* = 0.866, *SD* = 0.127) than the massed condition (*M* = 0.765, *SD* = 0.173). There was a significant main effect of task, *F*(1,44) = 4.470, *p* = 0.040, = 0.033 (BF₁₀ = 1.654). The percentage overall score on the post-test was significantly higher in the category (*M* = 0.793, *SD* = 0.190) than the procedure condition (*M* = 0.847, *SD* = 0.111). There was no evidence of an interaction between spacing and task, *F*(1,44) = 0.157, *p* = 0.694, = 0.001 (BF₁₀ = 1.541).

Figure 7

Procedure versus Category - Main effect of Spacing (two weeks later)



*Note*. A boxplot showing the percentage score on the post-test by spacing condition and number of steps in the procedure. The dashed line denotes chance at 50%. The significance stars represent: \* p < .05, \*\* p < .01, \*\*\* p < .001

Again, the homogeneity of variance assumption to run the ANOVA were not met, so I followed up with a robust ANOVA using the WRS2 R package (see [Table 7](#tbl-mytable5)).

Table 7

Experiment four- follow up - Robust ANOVA

|  | df1 | df2 | Q | p |
| --- | --- | --- | --- | --- |
| Spacing | 1 | 23.105 | 5.189 | 0.032 |
| Task | 1 | 21.636 | 2.614 | 0.120 |
| Interaction | 1 | 21.636 | 1.529 | 0.230 |

*Note*. Table displaying the results of the robust ANOVA using 20% trimmed means via the WRS2 R package

# Discussion

In experiment four I compared spaced versus massed practice across two tasks. Participants learnt a procedure, similar to the one used in experiment two (with the answer format changed from short answer to whether a provided answer was correct) and a version of the linear category used in the previous experiment (with an extra constraint). I found a significant main effect of spacing, but no significant effect of task nor a significant interaction, however, the interaction between spacing and complexity was closer to significance than in all other experiments in this thesis. Due to monetary, time and resource constraints inherent to a PhD project and this experiment’s placement at the end of the PhD project I did not power for the interaction term. This result supports the previous findings that spacing can be a useful technique to improve the learning of mathematics procedures. I first ran the preregistered analysis; however, two assumptions were violated, homogeneity of variance and the presence of outliers. To minimise the effect of these issues, I ran a robust ANOVA on the trimmed means. These results supported each other, and I have reported them both for transparency. This analysis also aimed to negate the effect of the ceiling effect by removing values at ceiling.

I set prolific up so that participants were supposed to be similarly matched to with the sample in experiment three, however, while all participants claimed to be current UK based undergraduate students, over 18 years of age, the mean age was around 27 years old, in contrast to experiment three, where the participants had a mean age closer to 20. There was also a much larger spread of ages with the standard deviation of 9.48 years for experiment four as opposed to 1.71 for experiment three. It was possible that participants on prolific were more likely to be mature students, however, this may also be an example of inaccurate screening on prolific where participants did not remove their student status after graduation. There is mixed evidence that age may affect the efficacy of the spacing effect, with some studies finding that it doesn’t interact ([Bercovitz et al., 2017](#ref-bercovitz2017); [Kornell et al., 2010](#ref-kornell2010)) and others finding evidence that age may reduce the magnitude of the spacing effect in verbal materials ([Balota et al., 1989](#ref-balota1989); [Simone et al., 2013](#ref-simone2013)). As I found a spacing effect in both conditions and had high performance this was unlikely to be an issue, however, perhaps I may have found larger spacing effects in a younger population which in turn may have made finding an interaction more likely.

I found that participants’ average performance across all practice trials was significantly better in the massed condition than in the spaced condition. This trade-off for lower performance during initial practice that leads to poorer performance after a delay is typical in spaced practice studies ([Walsh et al., 2018](#ref-walsh2018)) , indeed sometimes massed practice is sometimes still more effective for short intervals such as a week instead of four weeks ([Rohrer & Taylor, 2006](#ref-rohrer2006)). This phenomenon was not observed in experiment three or either of chapter three’s studies where practice performance did not significantly differ across groups. Prior studies looking at mathematics and spacing found that the increased fluency participants in the massed group felt during practice led to overconfidence in how well they would perform on a later test, with respect to participants in the spaced condition ([Emeny et al., 2021](#ref-emeny2021)). Participants in the massed group in this experiment could have been overconfident, leading to them paying less attention to the later questions or perhaps rehearsing less once the practice has finished. I found no significant differences between the massed and spaced group for the other individual difference measures: Arithmetic Fluency, Mathematics Anxiety or Working Memory.

The results were hindered by a ceiling effect. I tried to avoid this during development of this material by reducing the number of practice trials and adding an additional constraint to the category learning task, however, it was not sufficient. Alternatively, for the procedure, the ceiling effect could be due to the changes in how the material was presented. In chapter three I required a short answer from participants after they completed each trial. This had a much lower chance of making a correct guess. I chose to do this to make the task more comparable to learning the category.

The follow up study, two weeks after the initial post-test, found significant effects of spacing and task, but no interaction. However, this analysis was also potentially unreliable due to heterogeneous variance between conditions. In contrast, the robust ANOVA found a significant main effect of spacing, but no significant effect of task nor a significant interaction. This follow up study was intended to allow for forgetting and reduce performance. Many participants were still close to ceiling, particularly in the spaced condition. Interestingly the main effect of spacing was retained, but analyses suggested that this was no longer significant for the procedure task, but just for the category task. An important clarification was that the massed condition was no longer truly massed in this case as they took the second post-test two weeks after their first, so both had some spaced practice, albeit without feedback.

# General Discussion

The aim of the two experiments presented in this chapter was to investigate the effect of task complexity on spaced practice in mathematics learning. While the previous experiments in chapter three manipulated the number of steps in a procedure, varying the total number of elements required to be recalled, in experiment three and four I sought to manipulate element interactivity. I did this by ensuring that the number of elements needed to be recalled was constant; and instead varied how the material was structured. Experiment three compared the structure of two different categories, while experiment four compared a category learning task to a procedure learning task.

In experiment three, I taught participants two categories. The rules were available to participants during practice and feedback was provided. I found no effect of spacing in either the linear or relational category condition. This was particularly surprising in the linear condition as, using the same spacing schedule (see chapter three), I found large effects of spacing for procedures of similar length. This led us to think that perhaps even the lower complexity condition may have been too high in element interactivity for spacing to be effective or that there was too much interference between the two category structures. To investigate whether interference or task complexity was the cause of the lack of a spacing effect I next compared the linear category task to the five-step procedure used in chapter three.

In experiment four, I compared a five-step procedure (previously used in experiment two) to a five-constraint linear category. I found a significant main spacing effect, but no significant effect of task nor a significant interaction. The spacing effect was significant in the procedure condition providing further evidence that spacing was effective for learning mathematics procedures, however, there was also a significant spacing effect in the category condition in contrast to experiment three. The overall pattern of the data hints that spacing may have been more effective for the procedure than the category, however, this was not significant and there was no Bayesian evidence to support this. The interaction effect was closer to significance in this experiment than any other during the PhD, suggesting that this may be the best avenue to pursue in the future.

There were three changes to the linear category task from experiment three to experiment four. First, in experiment three the linear category was learnt alongside a relational category task, both tasks used very similar constraints making interference between the two very likely, while in experiment four the steps in the procedure were distinct from the constraints needed to be checked in the category, making interference less likely. Second, the constraints were presented as a list instead of a flowchart. Perhaps the list was easier to integrate as the items were spatially closer together and not in separate boxes. Third, the number of constraints increased from four to five.

There were two factors that may have diminished the ability to find the spacing by task interaction. First, the experiment had low power to detect the interaction. Recent experimental design best practices suggest that to adequately power for an interaction requires hundreds of participants, and I did not predict that the effect would fully reverse but instead merely attenuate which further increases power requirements ([Button et al., 2013](#ref-button2013); [Sommet et al., 2023](#ref-sommet2023)). I did not have the resources to run a fully powered study at this stage of the PhD. Second, the ceiling effects present across both experiments. Despite the attempts to reduce the chance of too high performance participants were still close to ceiling in all conditions, this may have hidden the interaction effect. As I needed 20 trials per session for category learning counterbalancing, I had 20 trials for the learning arithmetic procedures task, for parity. This was double the number of trials participants completed in chapter three, however, as participants were no longer required to input a number and instead choose correct or not, I thought the additional trials would provide a better granularity of performance as chance changed from close to zero to 50%. This may however have led to increased performance. In the future, with more time and resources, it would be better to do more extensive piloting to find the minimum number of trials to retain reasonable performance in the post-test, without being at ceiling. Alternatively, practice to criterion could be used. I chose to control the exact amount of practice used in these studies, rather than to criterion, as the deficient processing account of the spacing effect suggested that the amount of practice trials would affect the efficacy of spacing as massed practice would not benefit from further practice once a certain level of performance was reached. I decided to err on the side of too much practice as I believed this would be the best chance to see a spacing effect, considering the deficient processing account ([Delaney et al., 2010](#ref-delaney2010); [Greene, 1989](#ref-greene1989)), and capture any interaction. While the ceiling effect caused problems with the statistical analysis there is also some evidence that spacing is less effective for high-performing participants, instead benefiting medium performers the most ([Nazari & Ebersbach, 2019](#ref-nazari2019)), there was not however enough evidence to support this in the meta-analysis in chapter 2.

Alternatively, spacing may not interact with task complexity when operationalised as element interactivity. To have more confidence in this claim, further research could focus on reducing interference between the tasks, perhaps by maintaining the structure of the relational versus linear categories but varying the surface features of the task. For example, one category could categorise a number, as participants were required to do in this study, while the other category could define whether a shape was a member of the group. An example constraint could be whether the shape has a right angle, in contrast to the constraints used in this study such as “is the number even?”. This would reduce the effect of interference while retaining the structure of the categories and could be counterbalanced to control for the effect of these surface features. I initially considered implementing this manipulation, however, I considered it to be too far removed from the experiments in chapter three.

## Conclusion

I tested two novel manipulations of task complexity. Experiment three was the first time I failed to find a main effect of spacing during the PhD project. Neither the linear nor the relational category learning tasks benefited from spaced practice versus massed practice. However, in experiment four the linear category did benefit from spaced practice. The linear category in the second experiment varied slightly from the first as it was designed to better match the procedure task, however the largest difference was that it was learnt alongside a procedure instead of another similar category. In both experiments the complexity or task manipulation did not interact with the effect of spacing, however, experiment four was closer to significance than any other manipulation. Future research should explore the differences between the effectiveness of spacing for low and higher element interactivity material, but should ensure that the surface features of the tasks are distinct, limiting the potential interference.

# References

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, *4*(1), 3–9. <https://doi.org/10.1037/0882-7974.4.1.3>

Bercovitz, K. E., Bell, M. C., Simone, P. M., & Wiseheart, M. (2017). The spacing effect in older and younger adults: does context matter? *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, *24*(6), 703–716. <https://doi.org/10.1080/13825585.2016.1251552>

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. <https://doi.org/10.1038/nrn3475>

Chen, O., Kai Yin Chan, B., Anderson, E., O’sullivan, R., Jay, T., Ouwehand, K., Paas, F., & Sweller, J. (2024). The effect of element interactivity and mental rehearsal on working memory resource depletion and the spacing effect. *Contemporary Educational Psychology*, *77*, 102281. <https://doi.org/10.1016/j.cedpsych.2024.102281>

Chen, O., Paas, F., & Sweller, J. (2023). A cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educ. Psychol. Rev.*, *35*(2), 63. <https://doi.org/10.1007/s10648-023-09782-w>

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). *Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature* (Vol. 53, pp. 63–147). <https://doi.org/10.1016/S0079-7421(10)53003-2>

Donovan, J. J., & Radosevich, D. J. (1999). *A meta-analytic review of the distribution of practice effect: Now you see it, now you don’t*.

Emeny, W. G., Hartwig, M. K., & Rohrer, D. (2021). Spaced mathematics practice improves test scores and reduces overconfidence. *Appl. Cogn. Psychol.*, *35*(4), 1082–1089. <https://doi.org/10.1002/acp.3814>

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *J. Exp. Psychol. Learn. Mem. Cogn.*, *15*(3), 371–377. <https://doi.org/10.1037/0278-7393.15.3.371>

Hunt, T. E., Clark-Carter, D., & Sheffield, D. (2011). The development and part validation of a u.k. Scale for mathematics anxiety. *Journal of Psychoeducational Assessment*, *29*(5), 455–466. <https://doi.org/10.1177/0734282910392892>

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educ. Psychol. Rev.*, *27*(2), 317–326. <https://doi.org/10.1007/s10648-015-9309-3>

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*(2), 498–503. <https://doi.org/10.1037/a0017807>

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920951503. <https://doi.org/10.1177/2515245920951503>

Loenneker, H., Cipora, K., Artemenko, C., Soltanlou, M., Bellon, E., Smedt, B. D., Orza, J. G., Giannouli, V., Gutiérrez-Cordero, I., Lipowska, K., Dijck, J.-P. van, Yao, X., Nuerk, H.-C., & Huber, J. (n.d.). *Math4Speed - a freely available normed measure of arithmetic fluency*. <https://doi.org/10.31219/osf.io/8mtpe>

Nazari, K. B., & Ebersbach, M. (2019). Distributed practice in mathematics: Recommendable especially for students on a medium performance level? *Trends Neurosci Educ*, *17*, 100122. <https://doi.org/10.1016/j.tine.2019.100122>

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Appl. Cogn. Psychol.*, *20*(9), 1209–1224. <https://doi.org/10.1002/acp.1266>

Simone, P. M., Bell, M. C., & Cepeda, N. J. (2013). Diminished but not forgotten: Effects of aging on magnitude of spacing effect benefits. *The Journals of Gerontology: Series B*, *68*(5), 674–680. <https://doi.org/10.1093/geronb/gbs096>

Sommet, N., Weissman, D. L., Cheutin, N., & Elliot, A. J. (2023). How Many Participants Do I Need to Test an Interaction? Conducting an Appropriate Power Analysis and Achieving Sufficient Power to Detect an Interaction. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231178728. <https://doi.org/10.1177/25152459231178728>

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. *Cogn. Sci.*, *42 Suppl 3*, 644–691. <https://doi.org/10.1111/cogs.12602>

Wechsler, D. (1997). Wechsler memory scalethird edition (WMS-III). *San Antonio, TX: The Psychological Corporation*.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic press.

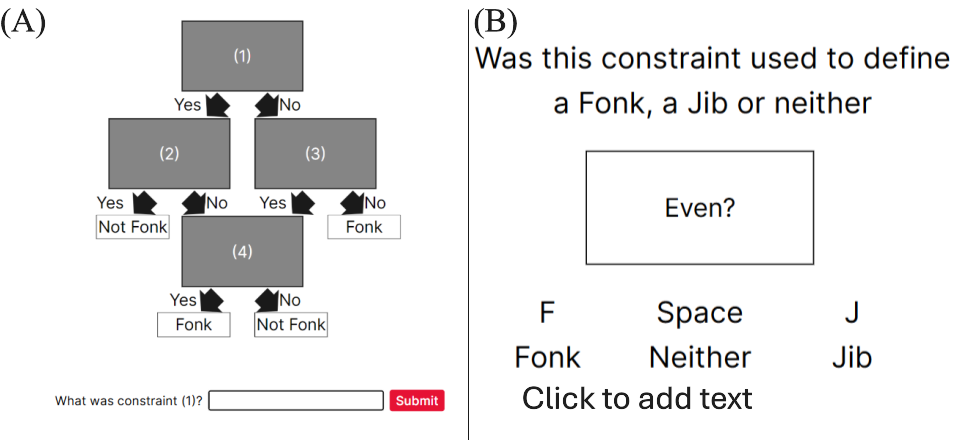
# Appendix

# Experiment three Secondary Post-tests

Participants additionally completed a task where they were asked to retrieve each constraint using a labelled flowchart as a cue, the *cued-recall* task (see [Figure A1](#fig-post-test) - A) and a recognition task which asked them to choose whether a cue belonged to the fonk category, jib category or was a new constraint not used in the task (see [Figure A1](#fig-post-test) - B).

Figure A1

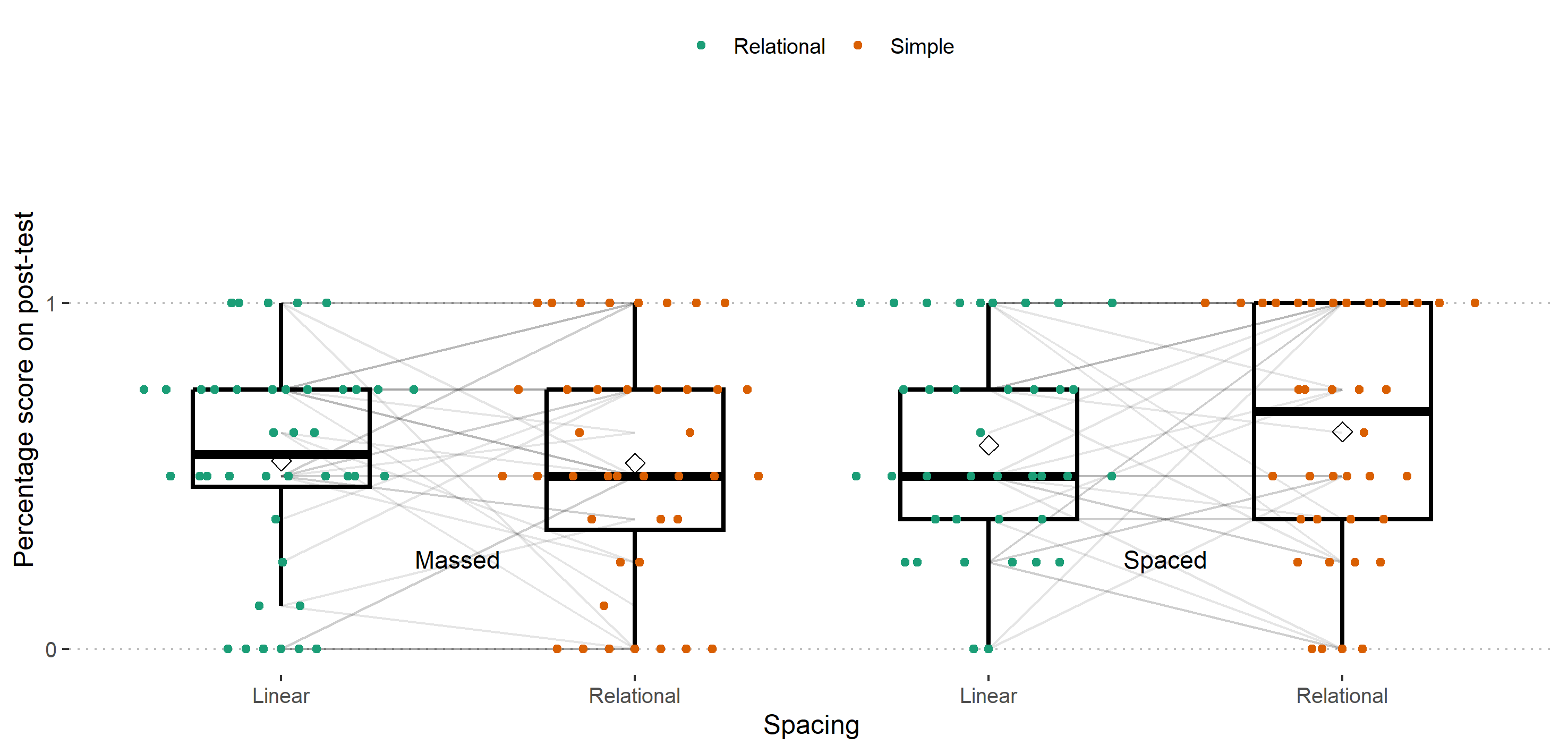
Participants completed additional trials identical to their practice trials, and a cued recall of constraints task (A) and recognition of each constraint task (B)



For the cued recall task, participants had to provide a short answer when prompted with either the linear or relational flowchart. Participants were given one point if the constraint belonged to the category and one point if it was in the correct location. Each category consisted of four constraints for a total of 8 points. There was no significant effect of Spacing (*F*(1, 76) = 1.188, *p* = .279 , = 0.011, BF₁₀ = 0.406), Complexity (*F*(1, 76) = 0.173, *p* = .678 , < 0.001, BF₁₀ = 0.192) or an interaction between spacing and complexity (*F*(1, 76) = 0.328, *p* = .569, < 0.001, BF₁₀ = 0.274).

Figure A2

Experiment three: Cued recall task performance after massed versus spaced practice by complexity condition

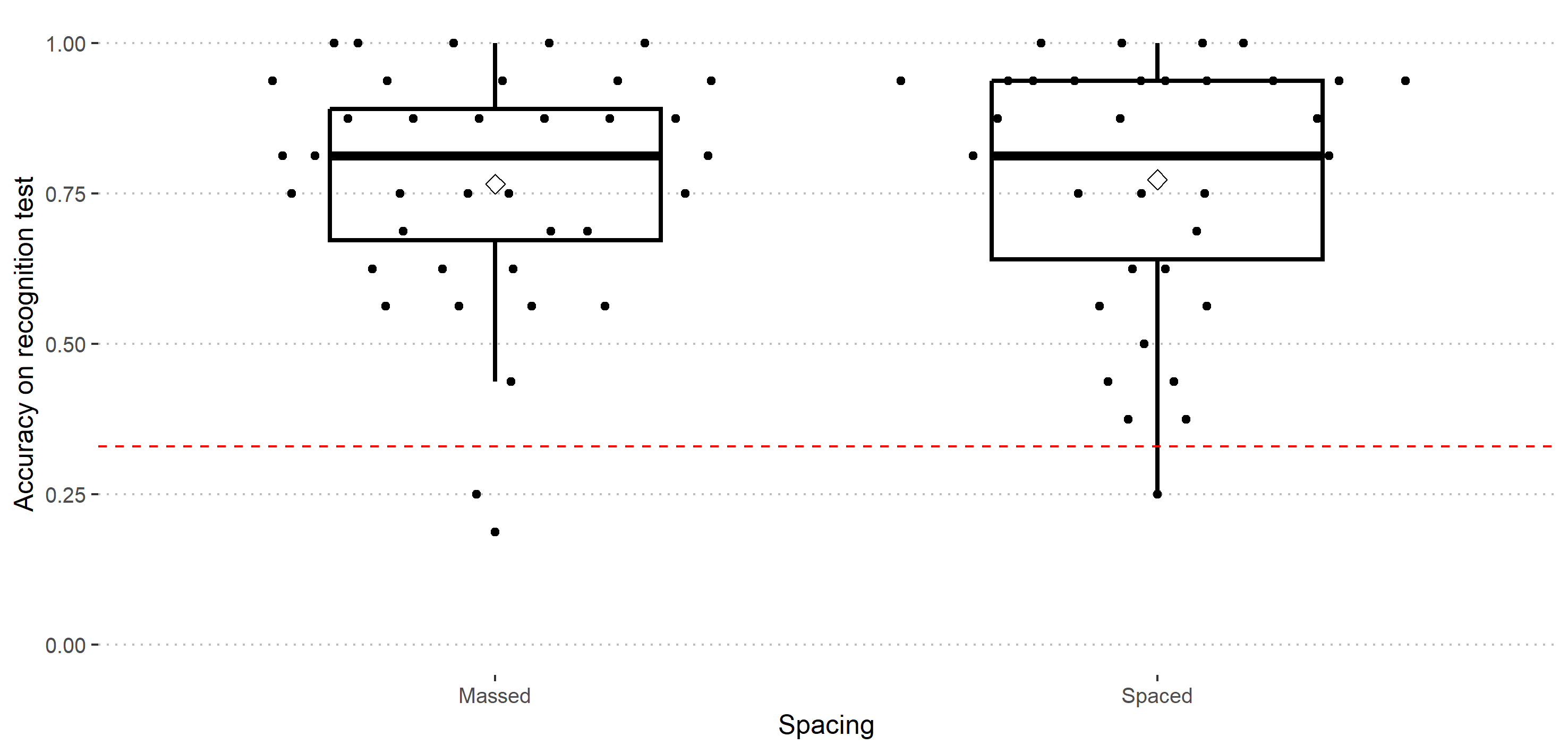


*Note*. A boxplot showing the percentage score on the cued recall task by spacing condition and complexity (linear versus relational category).

Generally accuracy on the recognition task was high (see [Figure A3](#fig-recog)) and there were no significant differences in performance between the massed and spaced groups (*t* = -0.163, , df = 75.11, *p* = 0.870, 95% CI = [-0.097, 0.082]). I also separated the types of mistakes participants could have made. Mistakes were either category based (i.e. they stated that a constraint from the linear category was from the relational category) or old/new mistakes (i.e. they responded that a constraint they had never seen before i.e. “Multiple of three?” was used in a category they did learn) (see [Figure A4](#fig-recog2)). The types of mistakes participants made across spaced versus massed conditions were very similar and not insightful.

Figure A3

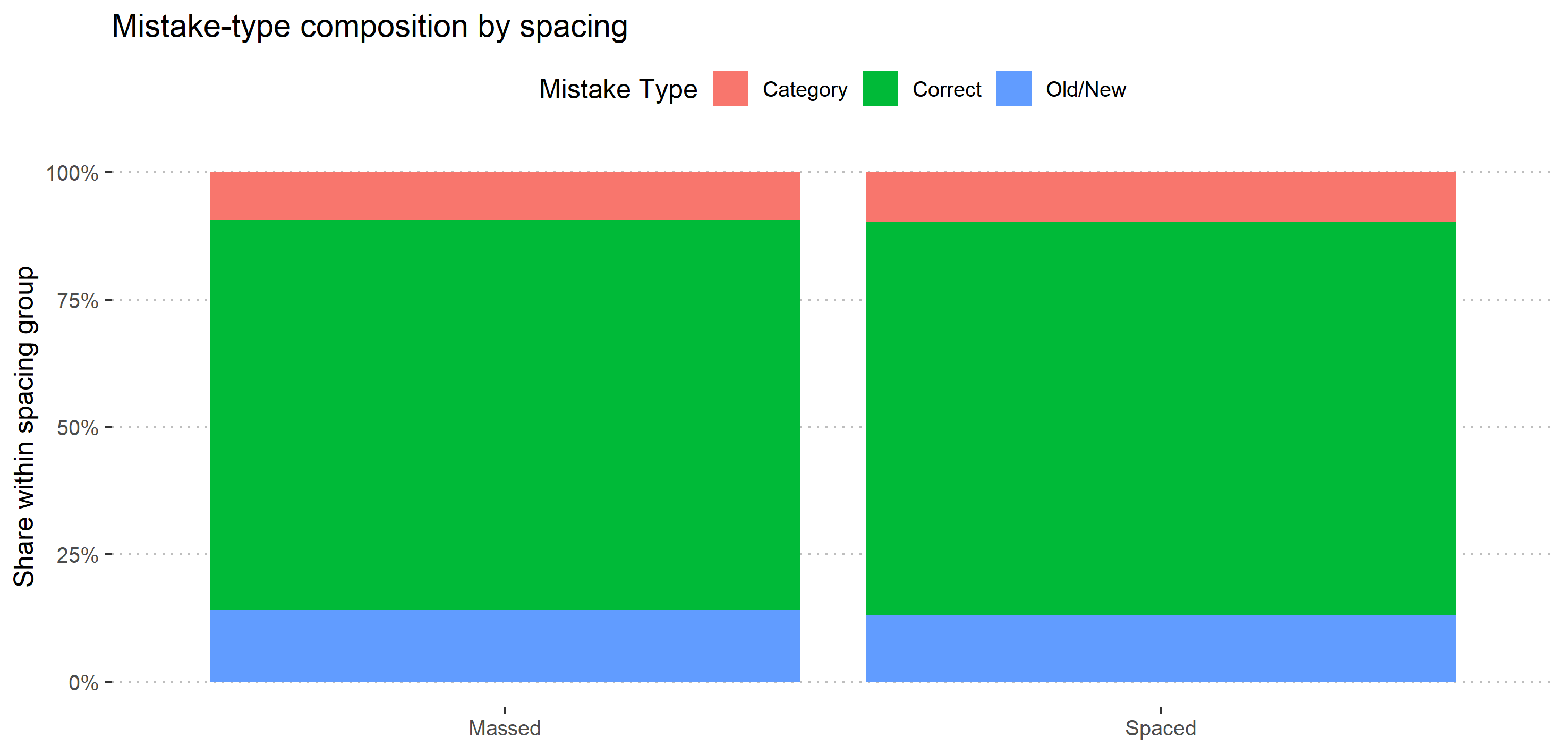
Experiment three: Accuracy on the recognition task by Spacing condition



*Note*. A boxplot showing the percentage score on the recognition test by spacing condition.

Figure A4

Experiment three: breakdown of correct answers verus the types of mistakes participants made during recognition task by Spacing condition



*Note*. A boxplot showing the percentage score on the recognition test by spacing condition.

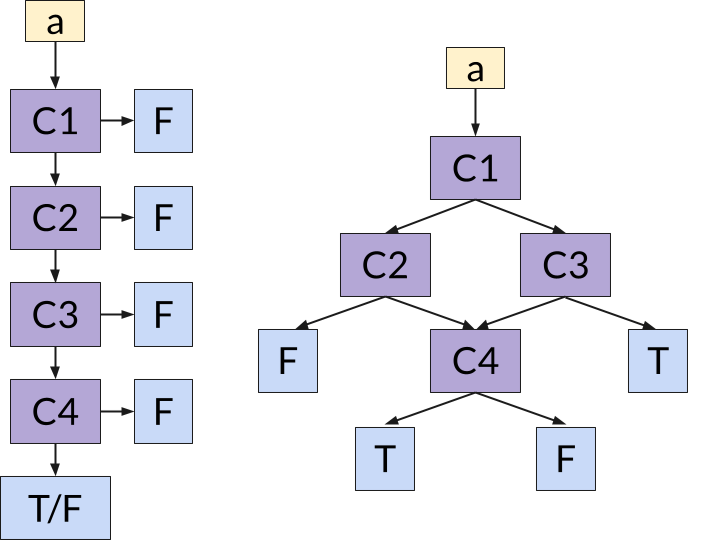
# Feedback Analysis

### Feedback Analysis Experiment three

I believed that the lack of spacing effect was interesting if participants were performing the task as instructed. I explored participants’ feedback to assess their participation.

Figure A5

Diagram outlining how constraints were labelled in experiment three



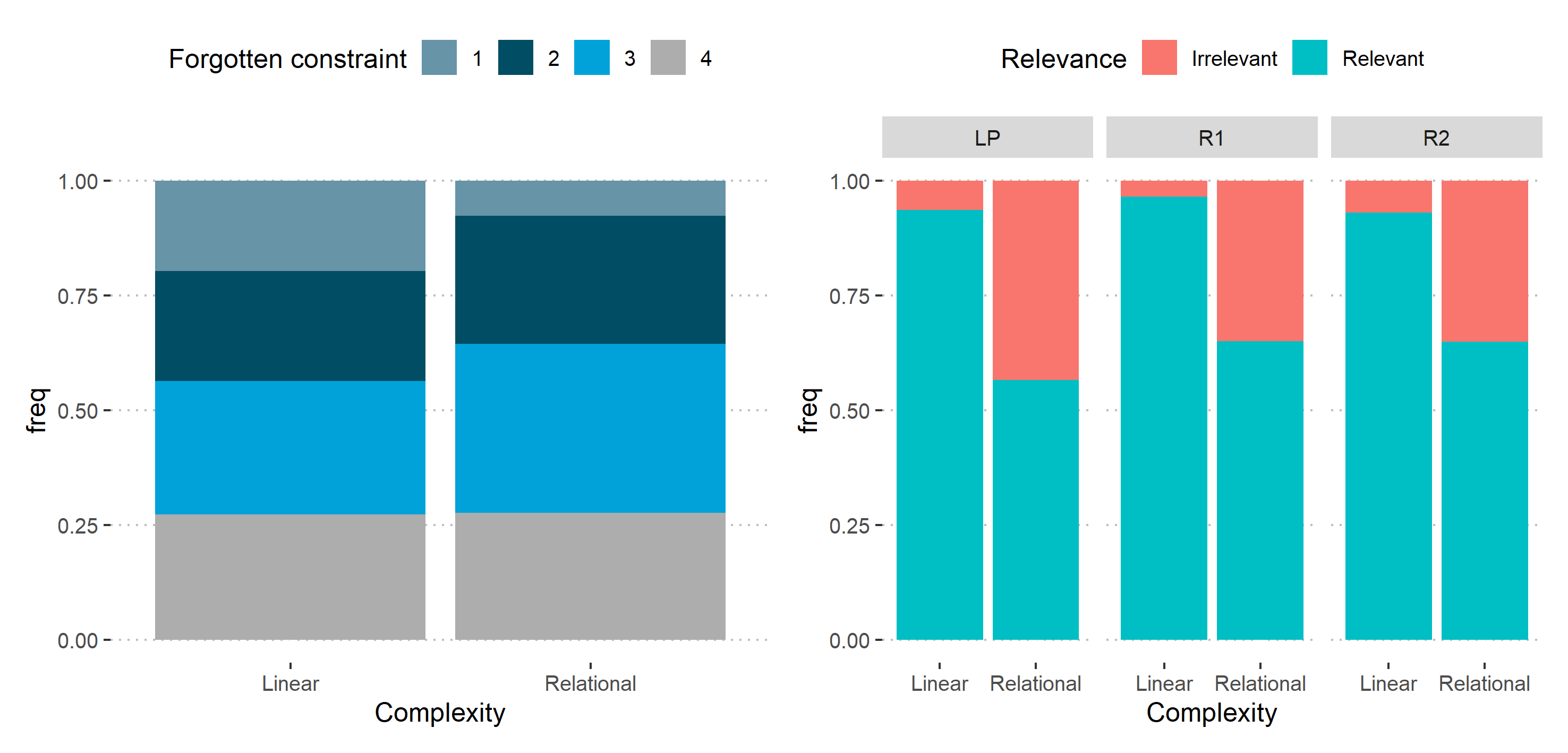
*Note*. The four constraints were labelled C1 to C4 and were arranged as shown in the diagram. (Left) The linear category structure, (Right) The relational category structure.

During the main task if participants’ response was incorrect, participants were required to select which constraint was forgotten that lead them to the incorrect answer, to continue. I reviewed participants’ selections to better understand how they were using the feedback. First, participants did not seem to be choosing any step in either the linear or relational constraint (see [Figure A6](#fig-exp3_fb) (left)). If they were simply trying to get through the task they may always have clicked step one or four. This was important when considering the relevance of their selection as the first step was always relevant.

I defined a *relevant* feedback selection as a response that would make sense for the participant to choose if they looked at what they were required to do and decided where they made their error. In the linear condition, if the third constraint was false, then a participant who said they made their error as they forgot the fourth constraint made an *irrelevant* choice as if they were checking in the order that they learned the constraints they would not need to recall constraint four during that trial. In the relational category trials (see [Figure A5](#fig-fb1) (right)), participants followed a specific route through the flowchart, if they end after the second choice (i.e. C2 is true or C3 is False) then C4 was irrelevant. If they need to check C2 then C3 is irrelevant and vice versa. There was a much larger proportion of irrelevant feedback choices for the relational than the linear category task across all sessions (see [Figure A6](#fig-exp3_fb) (right)).

Figure A6

Experiment three: Feedback - Forgetting and Relevance

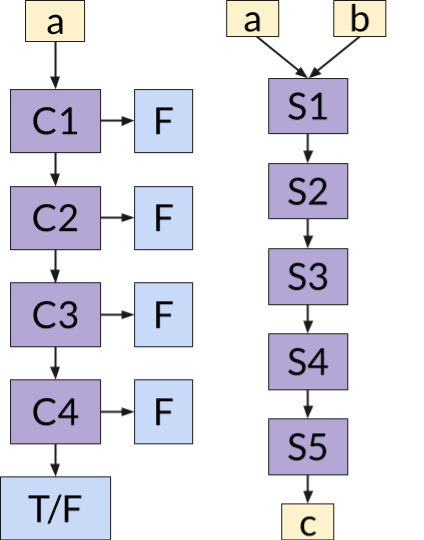


*Note*. (Left) A stacked bar chart the proportions of which constraints participants relayed that they forgot for the linear and relational categories. (Right) a stacked bar chart showing the proportions of relevant versus irrelevant forgetting

### Feedback Analysis Experiment four

Figure A7

Diagram describing how constraints were labelled in experiment three

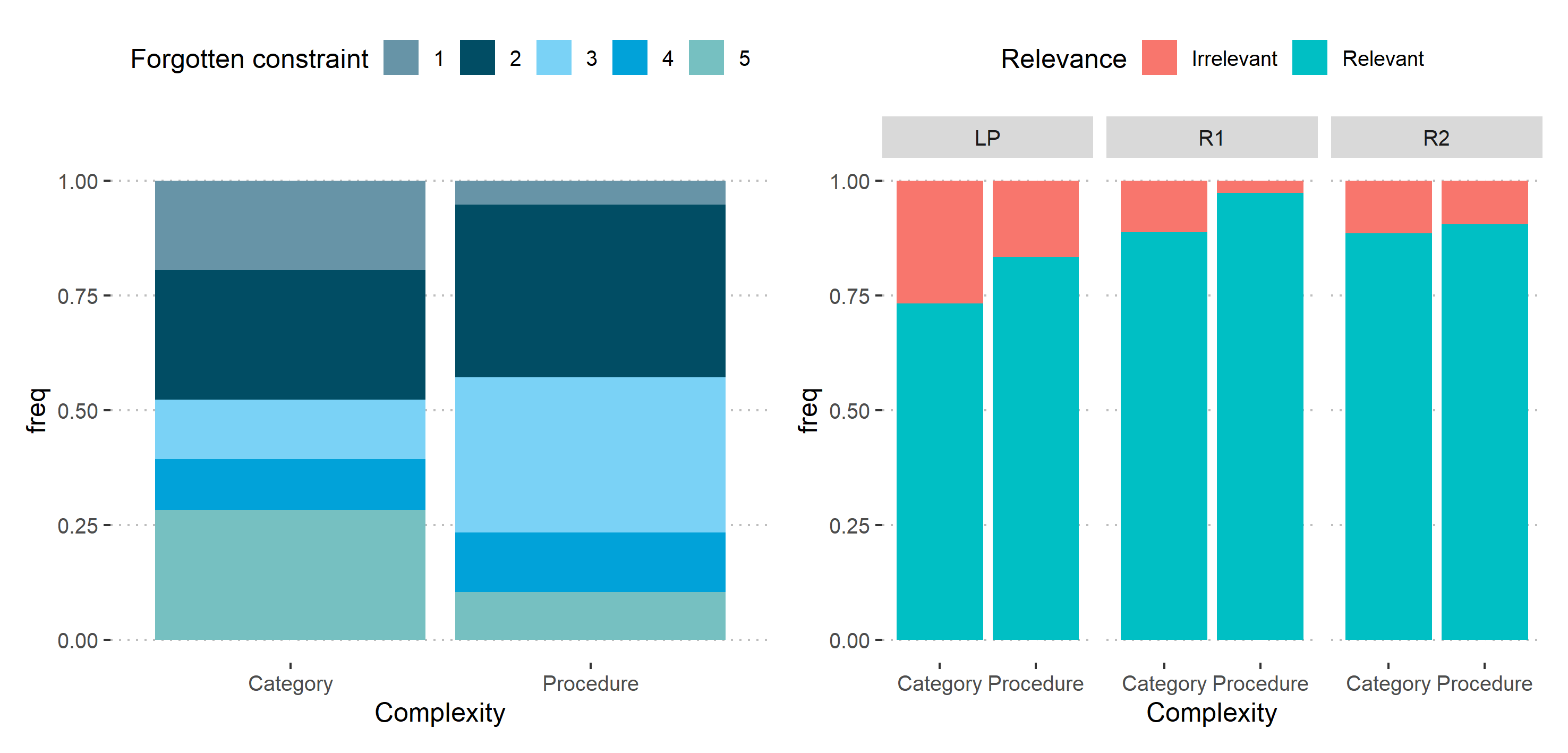


*Note*. The four constraints were labelled C1 to C4 and were arranged as shown in the diagram. (Left) The linear category structure, (Right) The relational category structure.

In experiment four, if incorrect, participants needed to first state whether they made an error or forgot a constraint/step. Then they make a selection as to which constraint/step was forgotten/mistaken that lead them to the incorrect answer. I added this to the feedback to gain more information about whether participants were making mistakes or forgetting.

Figure A8

Experiment three: Feedback - Forgetting and Relevance



*Note*. (Left) A stacked bar chart the proportions of which constraints participants relayed that they forgot for the linear and relational categories. (Right) a stacked bar chart showing the proportions of relevant versus irrelevant forgetting

Participants in the linear condition were choosing relevant feedback options most of the time. Participants frequently chose irrelevant feedback options.

and this provided another metric to analyse participants’ participation in the task. By recording participants’ indication of where they forgot or made an error it is possible to see whether that choice is relevant to the specific trial. For example, if they said they forgot the last step of the linear category procedure, but they did not need to recall it to succeed in that trial that would be an irrelevant choice.