**The effect of order-label training on generalization, a replication study of Ramscar et al., (2010)**

Eva Viviani

Michael Ramscar

Elizabeth Wonnacott

21 July 2020

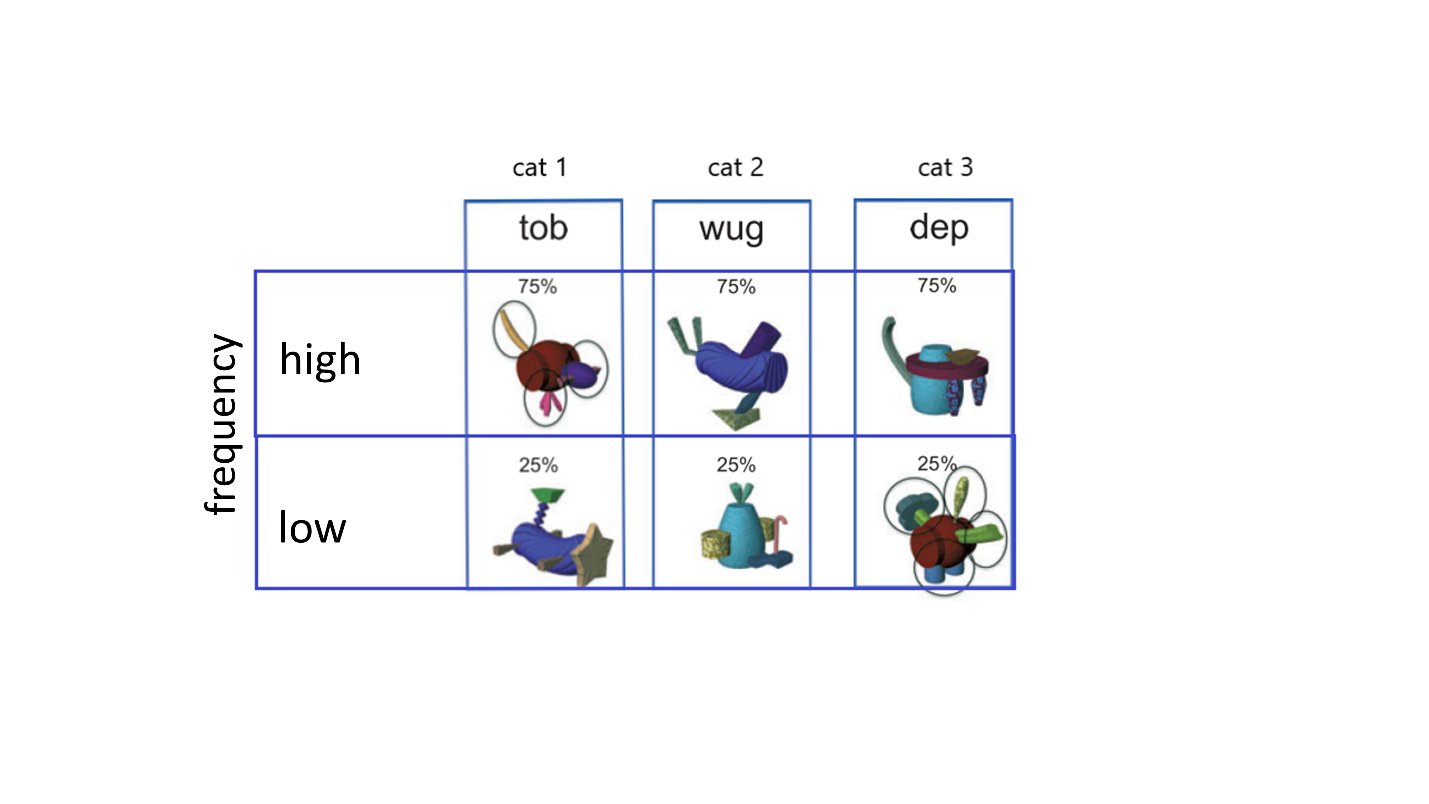
# Study information

In this preregistration we describe a detailed plan to replicate Ramscar et al. (2010) paper about the effect of the order of presentation of feature-label pair association on symbolic learning. We attempt an exact replication of the original study except (1) that the participants are recruited and tested online and (2) a further task is added at the end of the experiment. We will also looking at the data from the original two test tasks separately, as well as combined, in order to further elucidate the nature of these tasks.

The experiment is a word learning experiment in which participants learn associations between pictures of novel objects and written labels (e.g. wug). Critically, there are two learning-conditions – in the *feature-label* condition pictures are presented before labels, in the *feature-label* condition labels are presented before pictures. In both conditions there are three labels which are associated with six classes of fribbles, with each subcategory of fribbles having a set of discriminating features (Figure 1, circled features). However, critically, 1.1. 2.1 and 3.1 (*high-frequency* items*)* are encountered three times as often as subcategories 1.2 2.2 and 3.2 (*low frequency* items) (there is also a fourth “easy” control category of blue fribbles). Due to the way that body shape is shared across subcategories, it can be seen that for low frequency items, body shape is a high salience feature which has been strong associated with an alternative label to the correct label. According to discriminative learning theory, participants in the *feature-label* condition should be better able to ignore shape and learn the discriminating features which determine which labels should be used.

Following Ramscar et al., (2010) after exposure to the language, we include 2AFC tests testing generalization where participants had to choose which of four novel fribbles went with a label/ which of four labels went with a novel fribble. We also follow this with a new contingency judgment task that links the theoretical predictions about the resulting learning weights from error-driven learning, to the participants’ ratings behaviour of the feature-label association plausibility. The participants’ judgment is predicted to roughly reflect the learning weights of the model when learning is completed.

We expect that this pre-registration will establish (1) a robust paradigm that can be used as starting point for future studies online, (2) a sensitive task capturing estimates of learning weights across participants, providing a direct comparison to the predictions based on the RW model (Rescorla, Wagner, and others (1972)). We base our predictions on Ramscar et al., (2010), although our predicted effect sizes are based on data from an unpublished study that replicates Ramscar et al., (2010) effects[[1]](#footnote-2), and – for the additional contingency task – from our own pilot data.



Figure

Our primary research question is as follows:

Does Feature-Label-Ordering affect discrimination learning? In particular, when participants are confronted with novel exemplars, is there an advantage in terms of generalization of predictive cues for feature-label over label-feature in the low frequency condition?

# Hypotheses

**2AFC tasks**

We translate our predictions in terms of average accuracy estimated in two 2FAC tests (as previously implemented by Ramscar et al. (2010)), *2AFC – four pictures* and *2AFC – four labels* tasks. These tasks require participants to select the correct label-feature association among four alternatives from a set of novel exemplars (built with the same cues present in the stimuli in the learning phase). We get a measure of predicted effect size from the data from study that replicates Ramscar et al. (2010).

For (1) combined data across the two tasks (following Ramscar et al., 2010) and (2) for each 2AFC task separately (in case this effect in fact holds in only one task[[2]](#footnote-3)) we will test the following hypotheses,

1. **Higher accuracy in the low frequency condition for the feature-label learning group compared to label-feature learning group, as evidenced by an effect of training-condition for low-frequency test items.** We need a measure of predicted effect size for the BF analyses (see below). To get this, we ran a logistic mixed model over the study that replicates Ramscar et al., (2010) data (combining results across two 2AFC tasks), and fit separate slopes for the effect of learning at each level of frequency. We obtained the following coefficient statistics for the effect of learning in the low frequency condition: [ = 1.17, SE = 0.15, z value = 7.44, *p* < 0.0001]. We take the estimate **1.17** as our predicted effect size.
2. **A greater benefit of feature-label over label-feature for low frequency than for high frequency items, as evidenced by an interaction between frequency and learning-condition.** To get the predicted effect size, we ran a logistic mixed model (combining results across two 2AFC tasks from the study that replicates Ramscar et al., 2010) and obtained the following coefficient statistics for the relevant “learning by frequency” interaction: [ = 1.02, SE = 0.26, z value = 3.92, *p* = 0.0001]. We take the estimate **1.02** as our predicted effect size to be used in BF calculations described below.

Secondary Hypotheses:

1. **Overall higher accuracy in the *feature-label* learning condition, compared to *label-feature* learning-condition, as evidenced by a main effect of learning-condition.** To get our predicted effect size, we extracted the coefficient for the main effect of learning-condition from the model described above, this was [ = 0.66, SE = 0.14, z value = 4.87, *p* 0.0001] so our predicted effect size is **0.66**.

Note that this hypothesis is less strong – and is thus a secondary hypothesis - since although this was found in Ramscar et al., (2010) and his replication, this main effect *not* found in similar studies by Vujovic et al., (under review) and the prediction is less clear cut in terms of the computational model.

Hypotheses for control checks:

1. **Higher accuracy in the high vs low frequency condition as evidenced by a main effect of frequency.** To get our predicted effect size, we extracted the coefficient for the main effect of frequency condition from the model described under (B): [ = 1.7, SE = 0.24, z value = 6.85, *p* 0.0001]. Therefore our predicted effect size is **1.7**.

Note that this effect is predicted to be quite large and our other manipulations depending on an effect of frequency, hence this is a control check hypothesis.

1. **We also predict from Ramscar et al., (2010) that average accuracy in the condition where the association between fribble and label is 100% will be at ceiling for the majority of the participants (around 98%), regardless of the learning condition.** This condition will used mainly as sanity check over the whole experimental procedure, and in the analysis later on as exclusion criterion (see analysis plan).

**Contingency judgment task:**

In this task participants are given a label + novel fribble and are required to express their confidence on a scale from -100 to +100 regarding their association: “Use a scale from -100 to +100, where +100 means you are really sure IT IS the correct label, -100 means you are really sure IT IS NOT the correct label and 0 means you don’t know”. The possible label-fribble combinations are of two types:

1. match: label-fribble pair presented is 100% correct for both high and low frequency conditions.
2. mismatch: Items label is incorrect for this fribble type but the fribble has a high salience feature (shape) which *is* with the label. See figure 2 below for an example of both match and mismatch-type scenarios.

Correct performance is to give match items positive ratings and mismatch-items low ratings.

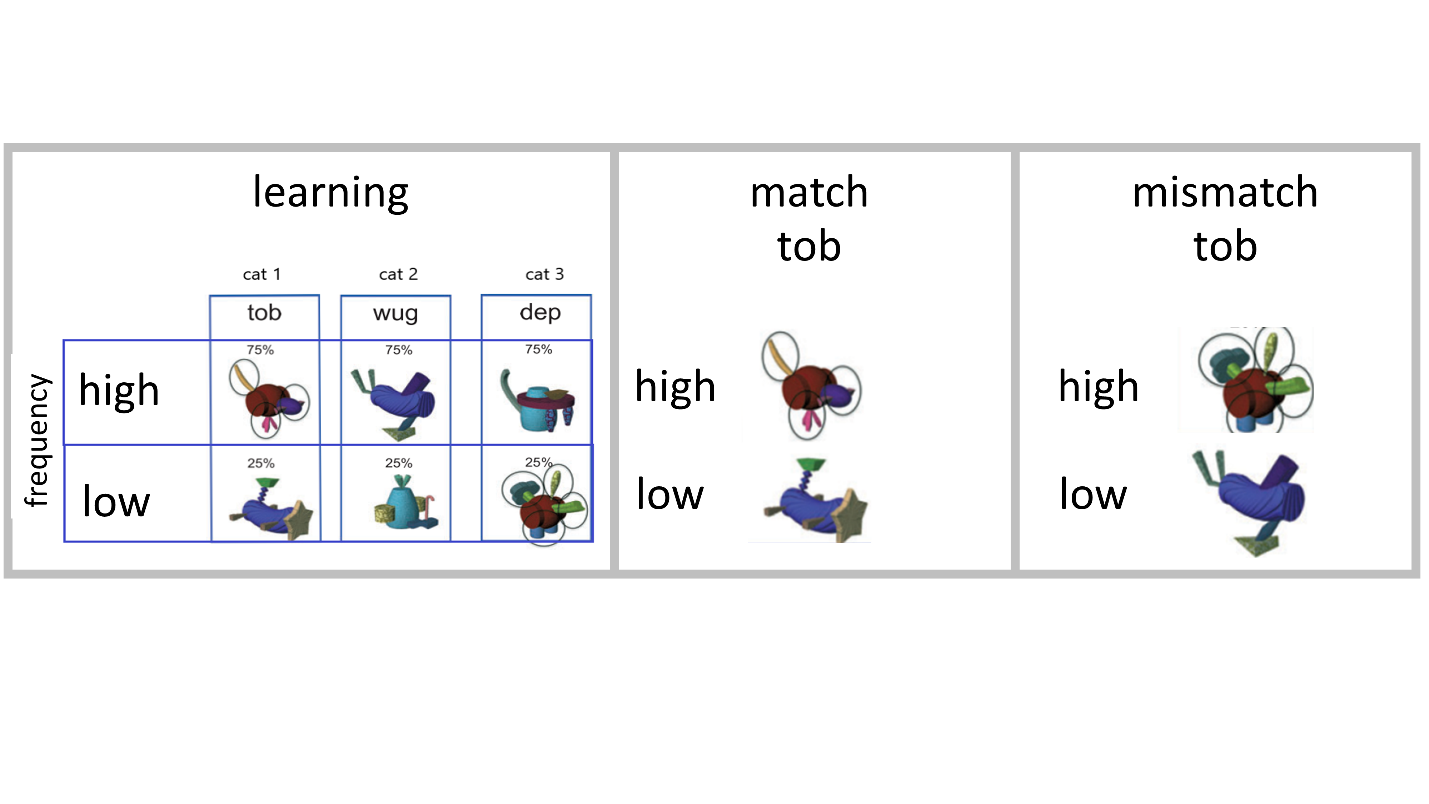


Figure 2

Based on the weights in the computational model, there are specific predictions about which of LF and FL training will lead to better performance for each of match-high Frequency, match-low Frequency, mismatch-high Frequency and mismatch-low Frequency. The weights of the model are shown in figure 3 below.

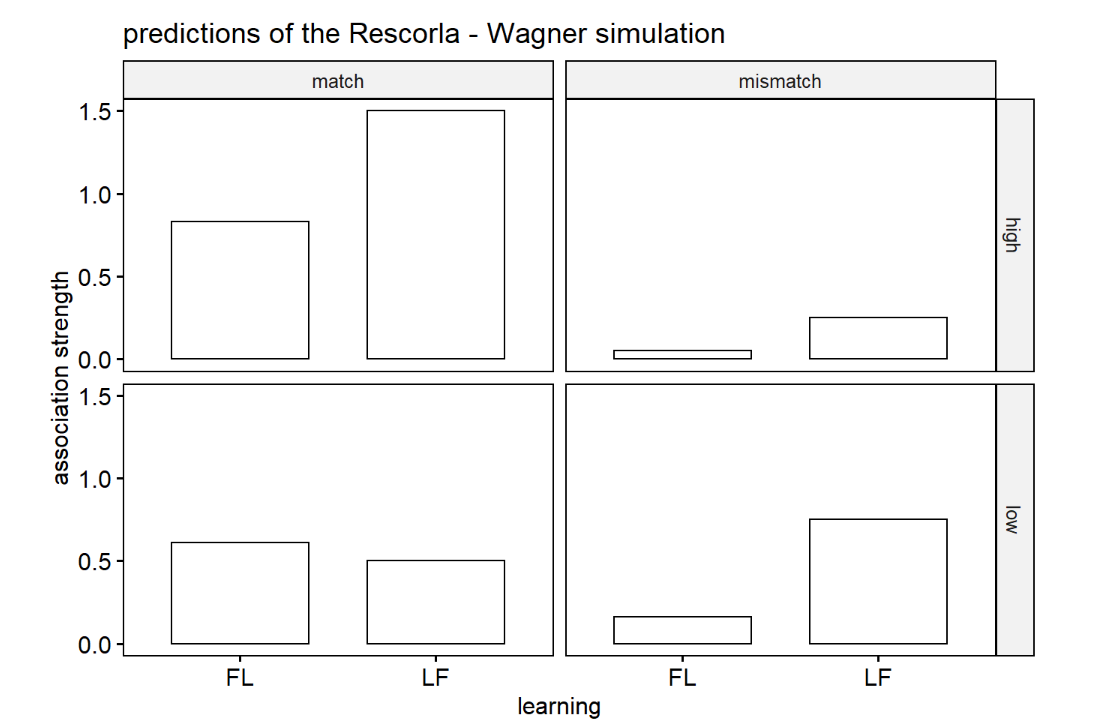


Figure 3

We translate the weights of the model into four predictions about which of FL/LF has *higher* ratings for each frequency-match combination.

These predictions will be tested in a linear mixed effect model where we have learning-condition, frequency condition and the match-type as factors where we fit separate slopes for learning-condition for each level of frequency condition by match-type.

C.1) High frequency – match: **stronger learning- in label-feature** (evidenced by positive effect of learning-condition i.e. higher ratings in label feature than feature label))

C.2) Low frequency – match: **stronger learning in feature-label** (evidenced by negative effect of learning-condition i.e. higher ratings in feature-label feature).

C.3) High frequency – mismatch-type: **stronger learning in feature-label** (evidenced by positive effect of learning-condition i.e. higher ratings in label-feature than feature-label)

C.4) Low frequency – mismatch-type: **stronger learning in feature-label** (evidenced by positive effect of learning-condition i.e. higher ratings in label-feature than feature-label).

To estimate predicted effect sizes: we have conducted a small pilot experiment where we included this test. Although the differences between conditions were all in the predicted direction, we had very small samples, (due to an error during data collection which meant that not all participants received every test). We therefore don’t think that the specific estimate obtained for each contrast in this pilot is likely very precise/reliable. Nevertheless, across analyses we saw an average difference between conditions of approximately 16 points, and in absence of other data we take this estimate as our predicted difference between conditions i.e. 16 points average for C.1, C.3.C.4 and - 16 for C.4.

**Hypotheses for control checks:**

D) **Higher ratings for match compared to the mismatch trials.** To get the predicted effect size, we ran a linear mixed model over our pilot data and obtained the following coefficient statistics for the main effect of type: [β = 53.78, SE = 6.24, z value = 8.60, p < 0.0001]. We take the estimate **53.78** as our predicted effect size.

E) **Higher ratings for high frequency match- than low frequency match trials, but lower ratings for high frequency mismatch than low frequency mismatch,** as evidenced by an interaction between frequency condition and type. To get the predicted effect size, from the linear mixed model ran over our pilot data we obtained the following coefficient statistics for the interaction between frequency condition and type: [β = 83.61, SE = 12.81, z value = 6.52, p < 0.0001]. Therefore we take the estimate **83.61** as our predicted effect size.

Note that these are relative large effects; finding these effects is necessary in order to be able to see the effects of our key hypotheses.

# Sampling plan

## Existing data

Registration prior to creation of data.

## 

## Data collection procedures

1. Participants will be tested via Gorilla (<https://gorilla.sc/>), and recruited through Prolific (<https://gorilla.co/>). Participants will complete the experiment online independently.
2. All adults who consent to participate will be tested, however some exclusion criteria will apply for the analysis. In line with Ramscar et al., (2010), those that will score less or equal to 80% on the control condition will be removed from the analysis.
3. Payment per session will be made to each adult participant (£7 per hour). Adult participants will be offered the amount at the end of the session only.

## 

## Sample size

This study aims to test a maximum of 200 adults English native speakers. The choice of the sample size is informed by an a priori power analysis based on data of the study that replicates Ramscar et al., (2010). Note that we simulate power for the 2AFC tests only, as we do not have appropriate data we can use for the contingency judgment test (our pilot was run between rather than within participants).

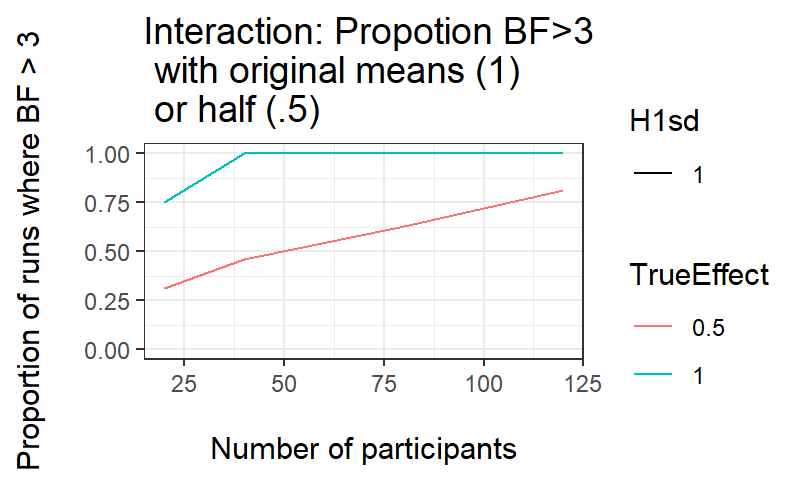
We ran logistic mixed effect models using values and sample variance extracted from that data set, with sample size varied from 20 to 120 participants. We also ran simulations where the fixed effects of the interaction was half the size as in the original, given that effect sizes are often lower in replications. Importantly, we have specified 48 trials (control trials are omitted) per participant in order to obtain power for a single 2AFC test (i.e. 4 picture test and 4 ltest separately), therefore, since only half of participants do each of these tests (following Ramscar et al., 2010) the sample size has to be doubled in order to get sample size for the whole experiment and power for each test separately.

We then used the same method which we will use for our actual experimental data (see below) to calculate Bayes Factors for the interaction between learning-condition and frequency condition (choosing this contrast as interactions require the largest samples). We looked at the sample required for 80% power to both accept H1. We also ran a model with the fixed effect set to 0 and looked for the sample required for 80% power to reject H0.

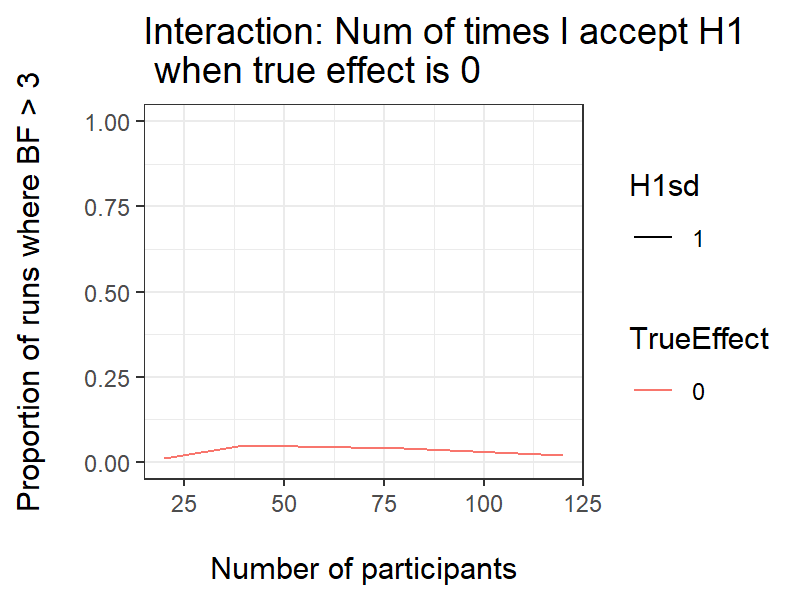
**Interaction between frequency and learning conditions**

Running these simulations, we see that there’s about a 80% probability of finding support for the H1 with a BF >3 with a sample size of around 30 participants *if the true effect for the interaction between frequency and learning perfectly matches the ones of the study that replicates Ramscar et al., (2010)*, i.e., =1. Therefore the minimum sample size that has a chance of supporting the H1 if those conditions holds is 30 participants per task (60 participants, blue bold line).

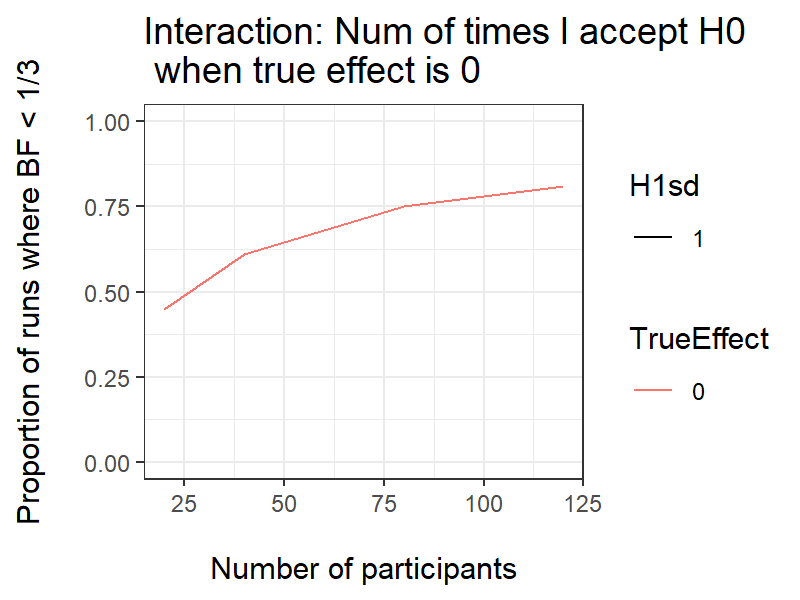
However, a more conservative approach is advised considering that effect sizes are generally smaller in replications. With the interaction set to be half of the original size effect i.e. ( =0.5) the minimum sample size that allows to achieve 80% probability to support H1 is around 100 participants (red bold line).



We looked also at the probability of an incorrect conclusion (observing support for H1 when H0 is true, Type I error) when the effect of the interaction in our simulations is set to 0. The simulations shows that the chances of accepting H1 when H0 is true are very low.



We computed also what are the chances of obtaining substantial support for the null hypothesis when the effect really is null. By running this simulation, we see there’s about a 80% probability of finding support for the null hypothesis with a sample size of around 100 participants with a BF< 1/3 (around 30% of the Bayes Factors are inconclusive evidence), if the true effect size is 0.



## Stopping rule

## We will first run 50 participants then will add data in batches of 20. We’re going to inspect the data and run the analysis we outlined in the analysis plan not earlier than having collected 50 participants (net of the exclusion criterion), and then after adding each batch of 20. We’re going to stop data collection before getting the estimated sample size (200 participants) only if we can find substantial evidence for accepting either H1 or H0 that is, if BF>3 for H1, or BF<1/3 for H1, for each of the hypotheses.

# Variables

## Manipulated variables

Learning-condition (between participants).

1. Label-to-Feature (LF) learning, i.e., label presented *before* the picture of the fribble.

2. Feature-to-Label (FL) learning, i.e., label presented *after* the picture of the fribble.

Frequency condition (within participants). The frequency of the subcategories is manipulated so that each labeled category drew 75% of its exemplars from one subcategory (high frequency condition) and 25% of its exemplars from another subcategory (low frequency condition). The two subcategories that make up each labeled category do not share any features. Finally, each low-frequency subcategory shared its high salient feature with the high-frequency exemplars of a different labeled category (see table 1).

## Measured variables

We have three testing tasks, two AFC tasks and one contingency judgement task. Participants will complete only one of the two AFC while the contingency judgement is completed by all participants. Following Ramscar et al., (2010), we use accuracy on the two speeded alternative forced choice test (AFC) where participants either:

A) Match an unseen exemplar to the four category labels (56 trials, among which 8 controls) **2AFC- 4 labels test**

B) Match a label with four unseen exemplars (56 trials, among which 8 controls) 2**AFC- 4 pictures test**

Note that HALF of participants will do one of these AFC tasks and half will do the other.

C) All participants will end their experiment with the contingency judgment task. Participants estimate the strength of the association between a label and an unseen exemplar on a scale from -100 to +100. Half of the trials will be match between label and the exemplars (match, 30 trials, 15 high frequency, 15 low frequency) and half mismatch between label and novel exemplars (mismatch, 30 trials, 15 high frequency, 15 low frequency). Total: 60.

# Design Plan

## Study type

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

## 

## Blinding

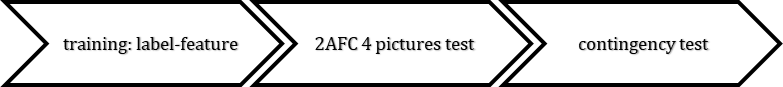
Participants will not know the learning group to which they have been assigned, nor that there is a frequency manipulation.

## 

## Study design

The experiment comprises a training session followed by two tests. Half of participants undergo the label-feature training and half the feature-label training and for each of these groups half take the 2AFC-four-pictures and half the 2AFC-four-labels test – all participants take the contingency test. This leads to four experiment paths as follows:

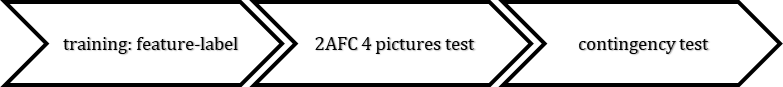
Group 1



Group 2



Group 3



Group 4



Assignment to each of the four participant groups will be random. The three tests will be analyzed separately meaning that we have the following design for each test:

**2AFC- 4 pictures test:** ONE between-participant factor *training-condition* with two levels (feature-label/label-feature) ONE within participant factor *frequency* (high-frequency/low-frequency)

**2AFC- 4 words test:** ONE between-participant factor *training-condition* with two levels (feature-label/label-feature) ONE within participant factor *frequency* (high-frequency/low-frequency)

*Note:* for these two AFC tests, there are items where a control category fribble is the target – these are removed from these main analyses; also note that, following Ramscar et al., (2010) we will also conduct an analysis with the combined data from the 2AFC tests, with the same factors as for the separate 2AFC tests.

**contingency test:** ONE between-participant factor *training-condition* with two levels (feature-label/label-feature) TWO within participant factor *item-frequency* (high-frequency/low-frequency) and *match-type* (correct, mismatch).

## 

## Randomization

Participants will be randomly allocated to one of the two learning conditions (feature-label or label-feature) and to one of the two 2AFC tasks. The learning sequence presented is randomized for each participant and presented in two blocks. No other cover task is performed.

# Analysis Plan

## Statistical models

For the Bayesian analyses we will be computing Bayes factors using the method advocated by Dienes (2008). For each hypothesis we test, we need a mean difference and a SE, as well as a predicted difference. We will get the mean difference and SE by running a logistic (for the two 2AFC tasks)/linear (for the contingency judgment task) mixed effect model and reading off the estimate and SE for the relevant coefficient. Note that for the 2AFC tasks we are therefore working in log odds space which meet assumptions of normality required by this method of computing BFs. H1 is modelled by using estimate of the mean for theory (i.e. our mean predicted differences as noted under hypotheses above) as the SD of a half normal (since all predictions are one tailed) distribution.

Logistics Mixed Effects Models:

GLMEs will be used for the 2AFC tasks. The choice of method is due to the nature of the dependent variable of correct (1) and incorrect (0) response. Frequency by participant random effects are included in the mixed models since frequency condition is within-participant.

Thus *frequency* and *learning* will be specified as fixed effects, with a random intercept by participants and a random by participants slope for frequency. The original model will be run in two ways: first with frequency and learning both given a centered coding as follows:

Version 1: glmer(acc ~ 1+ *frequency(Centered)* \**learningCondition(Centered)*+ (*frequency* |subject).

The output for this model provides co-coefficient statistics for (1) main effect of learning-condition (2) frequency and (3) their interaction. In addition, to test our key hypothesis, we also need co-efficient statistics for the simple effect of learning-condition for low-frequency items. To get this, we will rerun the model with the main effect of learningCondition and the interaction removed, and replaced with two separate slopes for learning-condition at each level of frequency – this will be achieved using the following syntax where *frequency\_Dummy* is the fixed effect of frequency using a dummy coding:

Version 2: glmer(acc ~ 1+ *frequency(Centered)* + *frequency(Dummy) : learningCondition(Centered)* + (*frequency* |subject).

Note that this has the same degrees of freedom as the original model but will output statistics for the simple effect of learning-condition for low frequency (as well as simple-effect of learning for high-frequency) which we may look at in exploratory analyses-and the main effect of frequency condition which will be identical to that obtained from the first version of the model.

Note: in the case of non-convergence of the glmer, we will first remove the correlations between the the intercept and frequency slope using the syntax (frequency||subject)., if this still doesn’t converge we will remove the slope for frequency i.e. (1|subject)).

The coefficients which will be extracted from each version of the model, and how they relate to the experiment, are shown in table 1 below.

Table 1

|  |  |  |  |
| --- | --- | --- | --- |
| Hypotheses  (as outlined in the  Hypotheses section) | Summary of the data from GLME with the following fixed effects: | Effect of interest (beta and SE for the relevant coefficient will be extracted as our model of the data). | Predicted estimates (run from the Ramscar 2010 pilot) |
| D | frequency(Centered) \* learning(Centered) | main effect of frequency | 1.7025 |
| C | main effect of learning | 0.6626 |
| B | interaction frequency by learning | 1.0265 |
| A.2 | frequency\_Dummy : learning(Centered) | simple effect of learning for high frequency | 0.1493 |
| A.1 | frequency\_Dummy : learning(Centered) | Simple effect of learning for low frequency | 1.1758 |

Linear Mixed Effect models:

An LMM will be used for contingency judgment task. The choice of method is due to the continuous nature of the dependent variable. Fixed effects are learning-condition, match-type, and frequency condition. Random by participant effects are included for frequency, match type and frequency by match type. This is because these are within participants variables. To test the hypotheses laid out above we are interested in the effect of learning-condition in each of the combinations of match-type by frequency condition. To get at this directly, rather than inspecting the model for a three way interaction, we will fit separate slopes for the effect of frequency for each combination of match-type by frequency, this will be achieved with the following syntax:

lmer(response~ 1

+ *frequency(Dummy)* : *matchType(Dummy) : learningCondition(Centered)*

+ *frequency(Centered)* \* *matchType(Centered)*

+ (*frequency* \* *matchType*|subject).

Note that the output of this model will give statistics for coefficients (1) main effect of match-type (2) main effect of frequency (3) match-type by frequency interaction and simple effect of learningCondition for: (4) high-frequency, match (5) high-frequency, mismatch (6) low-frequency, match (7) low-frequency mismatch *type[[3]](#footnote-4).* We need (4-7 to test our four key hypotheses, and 1 and 3 to test our secondary hypotheses).

In the case of non-convergence, we will first remove the correlations between slopes, then the slope for the interaction, then the slope for frequency, then for match type i.e. (*frequency* \* *matchType* |subject) (*frequency* + *matchType* ||subject) (*matchType* ||subject).

These coefficients provide the data used to test each hypothesis as shown in table 2 below.

Table 2

|  |  |  |  |
| --- | --- | --- | --- |
| Hypotheses  (as outlined in the  Hypotheses section) | Summary of the data from: | Effect of interest (beta and SE for the relevant coefficient will be extracted as our model of the data). | Predicted estimates (run from our own pilot) |
| D | LME model with frequency(Dummy) : type(Dummy) :learning(Centered) + frequency(Centered) \* type(Centered) | main effect of type | 53.78 |
| E | Interaction frequency by type | 83.618 |
| C1 | Simple effect of learning for frequency high – match | 16.001 |
| C2 | Simple effect of learning for frequency low – match | -15.05 |
| C3 | Simple effect of learning for frequency high – mismatch-type1 | 15.477 |
| C4 | Simple effect of learning for frequency low – mismatch-type1 | 17.270 |

## Follow-up analyses

None.

## Inference criteria

We will base our inferences on Bayes Factors analysis, though we will also report *p* values due to their greater familiarity to the readership. We will continue to work in log odds space (as for Frequentist) to meet assumptions of normality, using estimates and standard errors which come from the logistic mixed effects models. Following Dienes (2008), we test one sided predictions, modelling H1 as a half normal distribution with the SD set to the predicted effect size as laid out in the Hypothesis section. In each case, we take our model of the data to be the beta (i.e. mean difference) and SE for the coefficient from the relevant mixed effect model as laid out in the hypothesis section. We will say we have substantial evidence for H1 if BF is larger than 3 and for H0 if BF is less than .33.

## Data exclusion

Participants who score less or equal to 80% in the control condition will be excluded from the analysis.

## Missing data

The responses in the experiment are timed out and so there may be some missing data. If more than 10% of the responses in either test tasks from a participant are missing, that participant will be excluded from the analyses.

## Exploratory analysis

We will explore correlations between the generalization tasks and the contingency judgment task. We will also look for differences between the 2FAC tasks.

# 

# References

Dienes, Zoltan. 2008. *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Macmillan International Higher Education.

Nixon, Jessie S. 2020. “Of Mice and Men: Speech Sound Acquisition as Discriminative Learning from Prediction Error, Not Just Statistical Tracking.” *Cognition* 197: 104081.

Ramscar, Michael, Daniel Yarlett, Melody Dye, Katie Denny, and Kirsten Thorpe. 2010. “The Effects of Feature-Label-Order and Their Implications for Symbolic Learning.” *Cognitive Science* 34 (6): 909–57.

Rescorla, Robert A, Allan R Wagner, and others. 1972. “A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement.” *Classical Conditioning II: Current Research and Theory* 2: 64–99.

Vujović, Maša, Michael Ramscar, and Elizabeth Wonnacott. 2019. “Vujović, Ramscar & Wonnacott (Pre-Print). Language Learning as Uncertainty Reduction: The Role of Prediction Error in Linguistic Generalization and Item Learning.”

1. The reason we use this data rather than from the published study is that we have it available in a format where it is possible to run the necessary logistic mixed effect models to obtain our predicted effects [↑](#footnote-ref-2)
2. In Ramscar et al., (2010) there was no difference between tasks. In the pilot replication, although the key effects were present when collapsing across tasks there *was* a difference between tasks, although the labelling of the conditions is such that we are not clear which task showed the effect. [↑](#footnote-ref-3)
3. It thus has the same degrees of freedom – and is equivalent to - a model with the three main effects, three two way interactions + three way interaction. [↑](#footnote-ref-4)