COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13445

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13445

The Effects of Linear Order in Category Learning: Some Replications of Ramscar et al. (2010) and Their Implications for Replicating Training Studies

Eva Viviani,[a,b] Michael Ramscar,[c] Elizabeth Wonnacott[a]

[a]*Department of Education, University of Oxford*
[b]*Social Science and Humanities section, Netherlands eScience Center, Amsterdam*
[c]*Department of Psychology, University of Tübingen*

Received 19 December 2022; received in revised form 20 March 2024; accepted 29 March 2024

Abstract

Ramscar, Yarlett, Dye, Denny, and Thorpe (2010) showed how, consistent with the predictions of error-driven learning models, the order in which stimuli are presented in training can affect category learning. Specifically, learners exposed to artificial language input where objects preceded their labels learned the discriminating features of categories better than learners exposed to input where labels preceded objects. We sought to replicate this finding in two online experiments employing the same tests used originally: A four pictures test (match a label to one of four pictures) and a four labels test (match a picture to one of four labels). In our study, only findings from the four pictures test were consistent with the original result. Additionally, the effect sizes observed were smaller, and participants overgeneralized high-frequency category labels more than in the original study. We suggest that although Ramscar, Yarlett, Dye, Denny, and Thorpe (2010) feature-label order predictions were derived from error-driven learning, they failed to consider that this mechanism also predicts that performance in any training paradigm must inevitably be influenced by participant prior experience. We consider our findings in light of these factors, and discuss implications for the generalizability and replication of training studies.

Keywords: Language learning; Discrimination; Replication; Categorization

Correspondence should be sent to Elizabeth Wonnacott, Department of Education, University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY, UK. E-mail: elizabeth.wonnacott@education.ox.ac.uk

1. Introduction

Natural languages provide shared codes that map meaning onto form in a manner that enables their users to communicate about the world. This raises the question of how the correct "scope" of usage for the linguistic forms these codes comprise are learned. How, for example, do learners identify the critical features differentiating the referents of the labels "dog" and "cat" (as shared with other users of the code)? Ramscar et al. (2010) proposed that if such learning is underpinned by error-driven processes, it must be inherently discriminative. Error-driven learning involves acquiring and updating the predictive value (both positive and negative) of associations between cues and outcomes, which are typically strengthened when cues and outcomes *co-occur*, and weakened/inhibited by cue *background rates* (how often a given cue occurs in the absence of a given outcome) and *blocking* (the prior predictability of an outcome in a context in which it co-occurs with a cue). These factors result in *cue-competition*, which enables learners to unlearn unreliable cues to the benefit of reliable cues, and reduce their uncertainty about the features associated with a given label. Accordingly, if the world provides multiple potential cues to the label "dog," and some of these are predictive (e.g., "barks"), while others are spurious or unreliable ("has fur"), then mastering label use will require learners to discriminate (and dissociate) any unreliable cues from the cues (i.e., features of the environment) that are reliably predictive of (and hence should be associated with) a label.

A consequence of this analysis—and the key focus in Ramscar et al. (2010)—is that for effective unlearning to occur, the structure of events must support discriminative learning. A second consequence is that the effects of this support can be hugely affected by an individual's prior experience, since this influences background rates and blocking. This point was *not* considered in Ramscar et al. (2010), but it is critical when it comes to predicting the *outcome* of learning from a specific training regime, and is an issue we will return to later. In particular, Ramscar et al. (2010) suggest that because of the different functions and properties of labels versus referents, the *temporal order* in which stimuli are encountered can directly affect how any associative relationships between them are learned. Specifically, if referents contain many features, some of which are poor cues to label usage and some good, then in order for cue-competition to discriminate the appropriate from the inappropriate features, it is necessary they be encountered *before* rather than after labels. If learners encounter numerous varied exemplars of "dogs" and "cats" followed by their labels, this will allow cue-competition to strengthen the informative cues for the label "dog" (e.g., the feature "barks"), while uninformative cues ("has fur"), which result in prediction errors, will become disassociated. By contrast, when the order of presentation is reversed, such that each label serves as a *singular* predictive cue to each referent, cue-competition over the features of the referent *cannot* occur. While learning will still occur in the condition, the associative values learned will simply reflect the *frequency* with which each feature of a referent follows each label. Critically, the cue-competition that leads to the isolation of the discriminative (i.e., predictive) features of the referents will not be supported.

Ramscar et al. (2010) provide empirical evidence for this effect of linear order—which they refer to as "Feature-Label-Order" (FLO) effect —in two experiments with human learners. We focus on one of these, an artificial language learning experiment with adult participants.

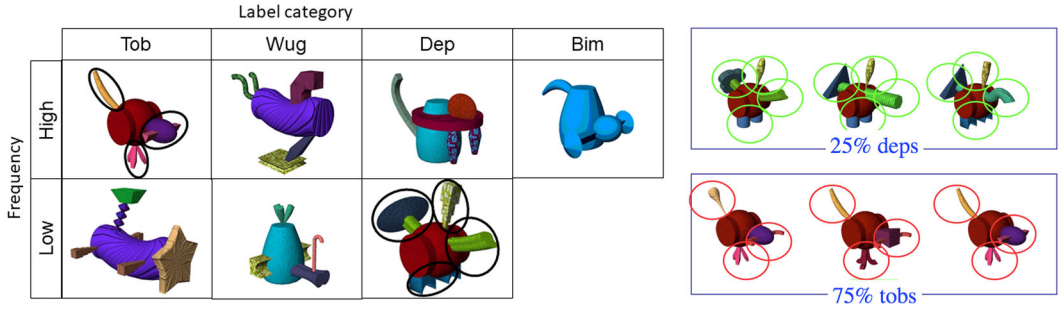

Fig. 1. Examples and stimuli occurrence rates from Ramscar et al. (2010). Left panel shows Fribble stimuli from Ramscar et al. (2010) and the current study. The label *bim* was used for a control group of uniformly blue examples. Experimental categories (*Tob*, *Wug*, and *Dep*). Each category has a high-frequency (75%) and low-frequency (25%) subset that each contain "high-saliency" feature (*body type*) that is shared with a subset of another category. The sets of discriminating features are circled in the low-frequency *dep* and high-frequency *tob* exemplars and right panel shows further exemplars from these two subcategories.

Participants (N=32) learned novel labels for categories of objects ("Fribbles," www.tarrlab.org) under either a feature-label (FL) learning condition—where they first saw a novel object (a Fribble) and then read a sentence presenting the category label ("That was a…" [wug/dep]); or in a label-feature (LF) learning condition—where they first read the category label ("This is a…" [wug/dep]) and then saw the novel object.

Fribble-stimuli were divided into the categories shown in Fig. 1. There was one easy category—all Fribbles labeled "bim" (control Fribbles) were bright blue (i.e., a single feature mapped to a single label); the critical predictions applied to the three experimental categories. For these, each category was divided into two subsets with the most salient feature—body-type (i.e., body shape and color)—systematically distributed such that it was not a defining feature for categorization. Instead, each label was predicted by other, more subtle discriminative features (circled for low-frequency deps and high-frequency tobs in Fig. 1). In addition, the frequency of training items was manipulated such that 75% of the exemplars of one category and 25% of the exemplars of another category shared the same body shape. Accordingly, to successfully learn the subcategories, participants had to learn to ignore (i.e., unlearn) the uninformative but salient body-type feature in favor of the discriminative features. This is particularly challenging for low-frequency items: For example, low-frequency "deps" share the same body-type (i.e., circular+brown) as high-frequency "tobs," such that the label "tob" and this body-type co-occur frequently: Were participants to focus on the frequency of this association, they would overgeneralize and identify low-frequency "deps" as a "tobs." This leads to the prediction that learners in the LF condition will primarily be influenced by the frequency of the co-occurrences between labels and features and will thus show poor learning of the discriminative features compared to the FL condition.

Ramscar and colleagues tested this using two four alternative force choice (4AFC) tests (administered between-participants) where participants either matched an unseen Fribble to the labels, or matched a label to one of four new Fribbles from each category. Since no

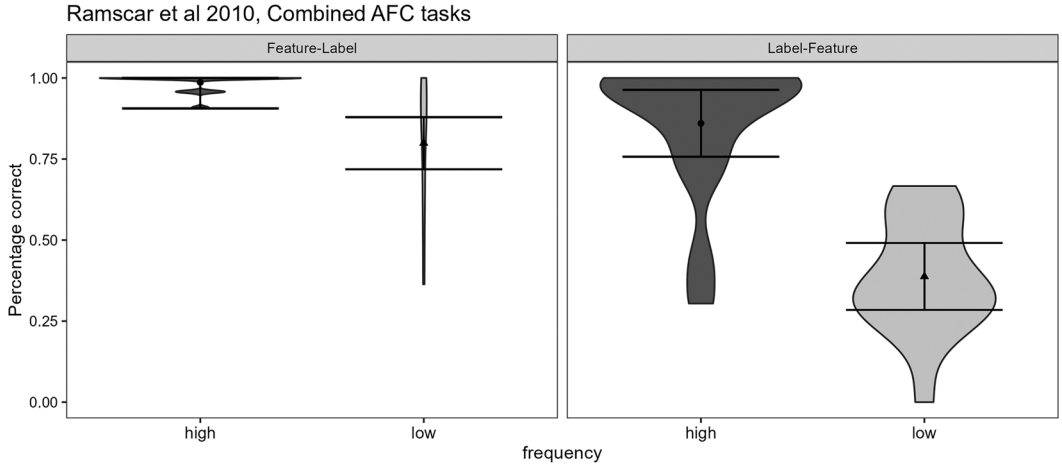

Fig. 2. Data from Ramscar et al. (2010). Violin plots show correct responses, split by frequency and learning-condition. Point shows mean and error bars 95% confidence intervals.

differences between test-sets were found, the results were pooled for analysis. Fig. 2 shows the data (replotted in the format that will be used in the current paper). It can be seen that the FL condition outperforms the LF condition, and this was shown to be consistent with a computational simulation (implemented using the Rescorla–Wagner learning rule [Rescorla & Wagner, 1972]) trained on the same input structure as the human learners. (Participants in both groups were at ceiling with the control fribbels.)

Data were analyzed in a (2 × 2) ANOVA where the difference between conditions was seen in a main effect of learning condition as well as an interaction between learning condition and frequency, in the direction of a larger frequency effect in the LF condition, and a larger FL benefit for low-frequency items than for high-frequency items. We have since reanalyzed these original data[1] using logistic mixed effect models similar to those used in subsequent replications discussed below (and the current study) which better reflect the underlying binary nature of the data (Jaeger, 2008). These continue to reveal a strong main effect of learning condition ($\beta = 2.34$, $SE = 0.46$, $p < .001$, odds ratio = 10.4), and frequency ($\beta = 3.07$, $SE = 0.43$, $p < .001$, odds ratio = 21.5), but find no evidence of the interaction ($\beta = 0.17$, $SE = 0.84$, $p = .83$, odds ratio = 1.187).

In sum, Ramscar et al. found a difference in performance between FL and LF trained learners, demonstrating that the order in which information is presented affects the learning of appropriately constrained generalizations. Theoretically, this provides support for a central tenet of error-driven learning, namely, that the brain tries to predict incoming input and *adjusts* these predictions based on error, such that the availability of useful error is a critical aspect of learning. There are also potential educational implications: If manipulating the sequential order of information can benefit learners, this offers the possible of developing educational programs specifically designed to elicit discriminative learning (see Ramscar, Dye, Popick, & O'Donnell-McCarthy, 2011, for an application to number learning).

1.1. Replications of the FLO effect

Ramscar et al. (2010) has made a significant impact in Cognitive Science. At the time of writing, it has received over 360 citations on Google Scholar, averaging to ∼3 citations per month over the past decade. The effect of linear order in learning has also been the subject of empirical research in Education (e.g., Eitel & Scheiter, 2015). However, as for most influential studies, no direct replication has previously been published.

The most direct replication to date is an unpublished study by Ramscar and McClure (2011) which used the same learning materials with a new group of participants ($N = 3$) also recruited and tested at Stanford University. The study had methodological differences, such as being part of an FMRI experiment, and participants completing both AFC tests with some counterbalancing of test order. Notably, unlike in the original study, not all participants scored perfectly on control items. Accordingly, following an unused exclusion criterion from the original study, participants who scored less than 80% on these items were excluded from the analyses ($N = 7$, with an additional participant excluded due to having greater than 10% missing data, leaving $N = 30$). The data are shown in Fig. 3, and we have again (re)analyzed the data using logistic mixed-effect models.[2] Analyses confirmed that there was again a strong main effect of frequency ($\beta = 1.970$, $SE = 0.28$, $p < .001$, odds ratio $= 7.17$). In contrast to the original study, the main effect of learning condition was not significant ($\beta = 0.57$, $SE = 0.4$, $p < .15$, odds ratio $= 1.77$); however, there was a marginally significant interaction between frequency and learning condition ($\beta = 1$, $SE = 0.54$, $p < .06$, odds ratio $= 3$) which broken down to show no significant effect of learning condition for high-frequency items ($\beta = 0.075$, SE $= 0.519$, $p = .871$) but a significant effect of learning condition—in the theoretically predicted direction of an FL benefit—for low-frequency items ($\beta = 1.07$, $SE = 0.52$, $p < .04$, odds ratio $= 2.92$).

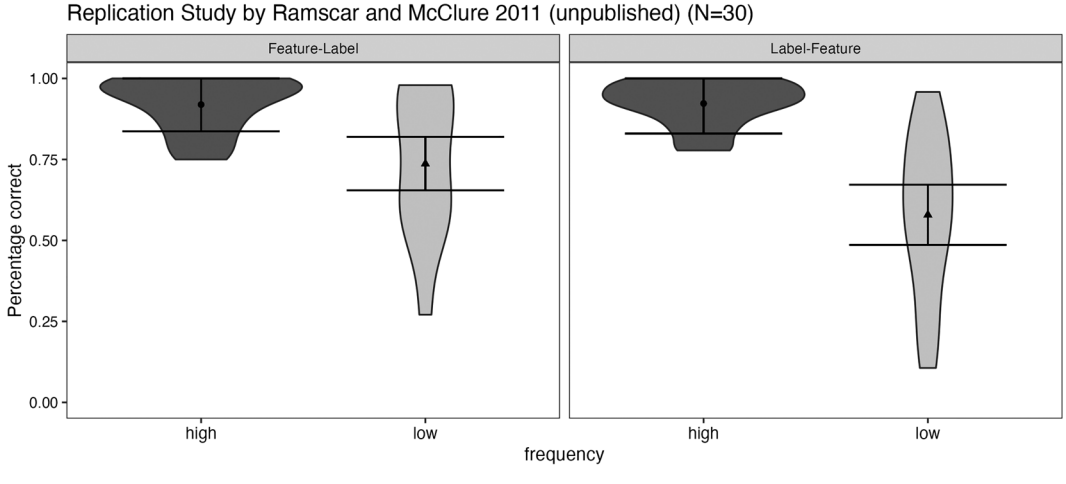

Fig. 3. Data ($N = 30$) replicating the FLO effect (Ramscar & McClure, 2011). Violin plots show correct response rates, split by frequency and learning-condition. Point shows mean and error bars 95% confidence intervals.

In addition to this more direct replication, numerous corollary experiments published over the years have supported the FLO hypothesis (Apfelbaum & McMurray, 2017; Dye, Ramscar, & Suh, 2011; Hupp, Sloutsky, & Culicover, 2009; Hoppe, van Rij, Hendriks, & Ramscar, 2020; Nixon, 2020; Ramscar, Thorpe, & Denny, 2007; Ramscar et al., 2011; Ramscar, Dye, Gustafson, & Klein, 2013; Ramscar, 2013; St. Clair, Monaghan, & Ramscar, 2009; Vujović, Ramscar, & Wonnacott, 2021). We consider the three most recent of these in detail.

Hoppe et al. (2020) studied the effect of the different linear ordering in prefixing versus suffixing on learning of the relationships between nouns and affixes in an artificial language. Under a discriminative learning account, suffixing benefits learning of abstract common dimensions from the stem nouns, that is, a benefit of the linear ordering where features (of the nouns) proceed outcomes (affixes). Their online experiment found a suffixing benefit (a FLO effect), but interestingly only for categories distinguished by multiple overlapping features; when cues were nonambiguous, both orders yield similar learning outcomes. This suggests that suffixing (i.e., as shown in Hoppe et al., 2020) offers a discriminative learning advantage specifically in the circumstances where cue-competition is necessary to disambiguate cues in the environment.

Nixon (2020) found a FLO effect in an (online) study that closely matched the design of the original experiment in Ramscar et al. (2010) except that the stimuli features over which generalization must occur were acoustic rather than visual. Specifically, English speakers learned to associate novel syllables (analogous to the Fribbles) with geometrical shapes (analogous to the labels), with the key discriminative feature of the syllables being lexical tone, which is not used discriminatively in English (and thus should not be salient). The design followed Ramscar et al. (2010) in having low-frequency items where there was a conflicting high-frequency association between a salient feature (here the base syllable) and an alternative meaning. Strong learning for low-frequency items thus involved down-weighting the high-frequency uninformative cue in favor of the discriminating tone cue. In the FL condition, audio syllables (multiple cues) were presented before geometric shapes (a single outcome), and in the LF, the order was reversed (geometric shapes before audio syllables). The key result was an interaction between frequency and learning condition and a simple effect of stronger performance in FL than in LF specifically for low-frequency items (the equivalent statistic for high frequency was not reported, but the means were in the reverse direction, though the items were near ceiling). This benefit of FL for low-frequency items is consistent with the key predictions of the discriminative account and confirms that the FLO effect is *not* about the ordering of visual versus linguistic stimuli per se, but rather about the arrangement of *information* to ensure that cue-competition over relevant features is possible.

Finally, online experiments by Vujović et al. (2021) looked at the same advantage of *suffixing* over *prefixing* as Hoppe et al. (2020), but using Fribble stimuli and the same basic class structure as Ramscar et al. (2010). Four hundred participants in two experiments were exposed to novel spoken words (in contrast to the written stimuli in Ramscar et al., 2010) comprising a CVC stem and a CV affix, which appeared either before (prefix) or after (suffix) the stem. The stems were accompanied by visual stimuli (a subset of the Fribbles used in Ramscar et al., 2010) and semantic and phonological properties determined the affix use. The design mirrored the key feature of the original in that correct classification–and thus

affix use—depended on unlearning the salient body shape cue in favor of discriminating cues, and this was particularly challenging for low-frequency items due to the frequent association of the body-type with an alternative label. Two experiments were reported that differed only in the number of individual nouns in the experiment (16 in the first, 8 in the second). Participants underwent various tests, including an AFC generalization test where they selected the best match for a novel Fribble from two audio heard stem + affix combinations. Analyses indicated an interaction between frequency and learning condition in both experiments, reflecting both stronger frequency effects with prefixes and a stronger suffix benefit for low-frequency items. In contrast to the original study, there was no evidence of an FL benefit for high-frequency items—as in Nixon, the means were in the reverse direction.

Vujović et al. (2021) also included computational simulations building on those in the original paper, which suggested that whether error-driven learning lead to an FL specific benefit for high-frequency items (and thus an overall "main effect" of FL) is dependent on exactly how the AFC test is simulated—that is, on how the process which evaluates the weights of associations with targets and with foils is implemented. In other words, seeing an FL benefit for high-frequency items is not predicted by the error-driven learning algorithm per se. This contrasts with low-frequency items, which show an FL benefit in the model regardless of decision rule implementation, due to the specific need to "unlearn" the conflicting salient cue for these items (and the model also consistently shows a stronger frequency effect for in the LF condition regardless of implementation). Interestingly, even for low-frequency items, Vujovic's participants showed a smaller benefit for the FL (i.e., suffixing) condition than in the original items, due to much greater overgeneralization based on body-type. Their modeling suggested that this was consistent with learners being in an earlier stage of learning relative to those original study, since the model also exhibited overgeneralization with suffixing stimuli in the early stages of learning. They further suggested this could be due to their more complex learning paradigm, which asked learners to learn noun names and affix categorization simultaneously. Consistent with the idea that the overall complexity of the input might modulate the FLO effect, a later version of the experiment (reported as experiment 6 in Vujović, 2020) using a more complex artificial language with more lexical items, did *not* show the FLO effect—with strong overgeneralization seen in both conditions. This suggests that the ability of testing to reveal learning (and unlearning) is far more dependent on participants meeting a certain threshold of discrimination more than Ramscar et al. (2010) considered.

Overall, these later studies offer support for Ramscar et al. (2010)'s analysis, though they suggest that whether a "features first" benefit is seen may depend on a variety of factors such as cue ambiguity, whether there is competition between discriminating and more frequent cues, the effect of overall task complexity, and learners' experience/state of learning. It is also important to note that where a FLO effect has been seen, effect sizes have been smaller (see Fig. 11 in the General Discussion[3]).

1.2. Overview of the experiments in the current paper

Taken together, the experiments described above provide corroborating evidence in support of the FLO effect, though even in the closest replications, the effect sizes were smaller and

data generally noisier than in the original study. There is also some question as to whether the FL benefit in the original paradigm should be expected to hold across high- and low-frequency items, or just for low-frequency items. Given the theoretical and potential educational implications of the original finding, and bearing in mind the increased emphasis on the need for replication in Psychology (Open Science Collaboration, 2015), we conducted two large scale (near) direct replications of the original study. In contrast to the original, but consistent with the three more recent studies described above, we ran these studies online. Some researchers have raised concerns regarding the validity of online versus in-lab experiments (e.g., Finley & Penningroth, 2015), although these primarily concern the collection of behavior data such as reaction times. On the other hand, online recruitment methods have grown in popularity due to the possibility of recruiting larger samples in short periods, thus addressing concerns about the use of low-powered samples in Psychology experiments (e.g., Morey & Lakens, 2016), and due to the necessity in the context of the COVID-19 pandemic. It is thus important to establish the replicability of various experimental paradigms in online contexts. In what follows, we report these replications of Ramscar et al. (2010)'s experiment 1, which were conducted exclusively online and were preregistered on OSF.

2. Experiment 1

2.1. Overview and predictions

This experiment used the same stimuli and training and test structure as the original experiment. One extra test was added at the end of the experiment —a contingency test (where they had to judge how well a given Fribble matched a label; Fig. 4). For clarity and conciseness of presentation, the methods and results for this test (which are somewhat mixed) are provided in the online Appendices rather than the main text (see online Appendix: Contingency and additional subset analysis).

Analysis plans were preregistered. Importantly, we preregistered that in the light of findings suggesting an effect of test-type in this paradigm (Vujović, 2020), we would examine the results of the two AFC tests separately.[4]

For each test, we preregistered the following predictions: We expected to find a main effect of frequency, that is, overall better performance with high-frequency items. This is essentially a positive control (i.e., if it was not observed, it would indicate an issue with the paradigm). On the basis of Ramscar et al. (2010), we also looked for a main effect of frequency in the direction of overall higher performance in FL than in LF. However, since this was not observed in the other replications described above (or originally predicted), we preregistered it as a secondary prediction.

The most critical are our primary predictions: (i) An interaction between frequency and learning condition, reflecting a larger frequency effect in the LF condition and (ii) A simple effect of learning condition for low-frequency items, reflecting stronger learning in the FL condition. We pre-registered that observing either of these would support the FLO analysis.

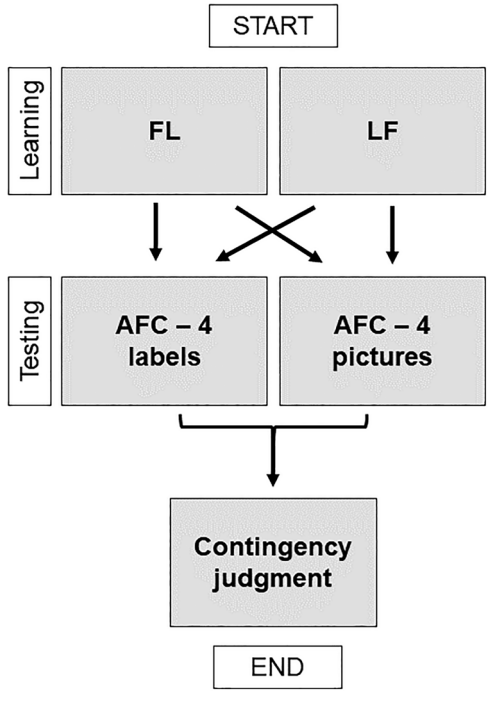

Fig. 4. Design of the experiment.

2.2. Methods[5]

2.2.1. Participants

Two hundred and ninety-nine participants were recruited via Prolific. Each confirmed they had native or native-like English proficiency and were between 18 and 35 years old. They were randomly assigned to the LF and FL conditions, and to the Four Pictures + Contingency or Four Labels + Contingency test (see Fig. 4). Data from 101 participants who scored below 80% on control items and 22 with 10% or more missing trials on either test were excluded following preregistered criteria, leaving 75 in the four pictures, and 93 in the four labels task. Data collection used a preregistered optional stopping procedure, starting with 50 participants and adding in batches of 20 up to a preset maximum (see Sample size planning, and also Data loss and deviations from preregistration), with Bayes Factors checked to see if we had substantial evidence for *either* H1 *or* H0 for each of our hypotheses (in which case data collection stopped) using the process explained below. [6]

2.2.2. Stimuli

These were identical to Ramscar et al. (2010): 75 "Fribble" pictures (www.tarrlab.org) were divided into four categories, including a "control" Fribble (always blue) and the three experimental categories (comprising six subcategories) described above (Fig. 1). The control category comprised 15 exemplars, each high-frequency subcategory 15 exemplars, and each

low-frequency subcategory 5 exemplars. Labels were presented as text. The experiment was redeveloped with the JsPsych library (De Leeuw, 2015) and hosted on the Gorilla platform.

2.2.3. Procedure

Participants were told that they would "learn an alien language." They saw pictures of "aliens" and heard "how they are referred to in the alien language." The training and testing structure employed is depicted in Fig. 4.

2.2.3.1. Learning: A total of 90 high-frequency, 30 low-frequency, and 30 control trials were presented in two identical blocks separated by a short break. Each trial presented a Fribble and its label, with the order and timing of Fribble/label presentations varying between conditions, as illustrated in Fig. 5.[7]

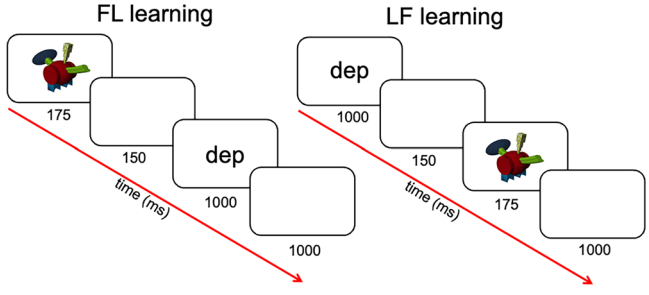

Fig. 5. Temporal structure of the learning trials.

2.2.3.2. Four pictures AFC task: Test-trials presented a label and four previously unseen Fribbles, and participants had to pick the Fribble matching the label (Fig. 6). Responses made

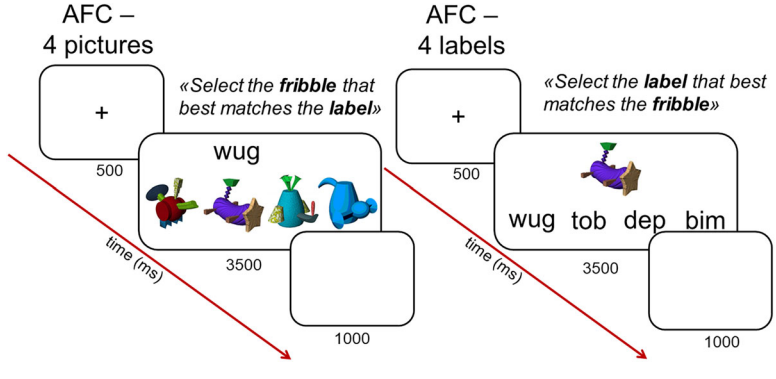

Fig. 6. Schematic representation of a single trial in both AFC tasks.

after 3500 ms were coded as timed-out. There were 8 control tests (control fribble as target, and foils from both high- and low-frequency categories), 24 high-frequency tests (a high-frequency target + two foils from the other two high-frequency categories + one control

fribble), and 24 low-frequency trials (a low-frequency target + two foils from the two other low-frequency categories + a control fribble foil). Note that these last tests can be expected to be particularly difficult because one of the foils will have a body-type that has frequently been associated with the target label (Fig. 1).

2.2.3.3. Four labels AFC task: In each trial, participants saw a picture of a Fribble—either a control Fribble (8 trials), a high-frequency Fribble (24 trials), or a low-frequency fribble (24 trials)—above four labels: dep, tob, wug, and bim.

2.2.4. Analyses

We conduct equivalent analyses to the ANOVAs/*t*-tests in Ramscar et al. (2010) but use logistic/linear mixed-effect models (using the package `lme4`; Bates, Mächler, Bolker, & Walker, 2015) with participant as the random effect and fixed effects for learning condition, frequency, and their interactions. We run different versions of the models with different codings to capture main versus simple effects, with the fixed effect coefficients providing statistics for our hypotheses. Instead of interpreting frequentist *z*-values and *p*-values, we use the relevant beta and SE values from the fixed effects to calculate Bayes Factors, which provide a way of performing the Bayesian equivalent of significance testing, but have the advantage of providing information that a *p*-value cannot: A "null" result (i.e., $p>.05$) does not tell us whether we have evidence for the null or no evidence for any conclusion at all (or even evidence against the null). This information is important, particularly in a replication study where in the case of null results we wish to know whether we do indeed have evidence against the effects which were originally found.

Bayes Factors were computed following the approach advocated by Dienes (2008, 2014, Dienes, & Wonnacott, 2021) using Diene's calculator (Baguley & Kaye, 2010). The calculation requires three numbers: (i) the estimate (beta), (ii) SE from the relevant coefficient in the mixed-effect model, and (iii) a rough estimate of the predicted difference (i.e., predicted size of the beta) for the hypothesis. We based these on beta values extracted from the coefficients of equivalent models run over the data from the earlier replication by Ramscar & McClure (2011). In the calculator, the predicted value is used as a parameter (or the *scale factor*) in a model representing the plausibility of different effect sizes if H1 is true. The calculator tests whether the data summary is more likely under this model of H1 than under a model representing the null (i.e., only plausible effect is 0). The result is a ratio representing the relative strength of evidence for H1 versus the null—this is the Bayes Factor. Values above 1 indicate more evidence for H1 and below 1 more evidence for H0.[8] Bayes Factors are interpreted continuously; however, for hypothesis testing, we can also use discrete evidential categories. We use: $BF > 3$ indicates substantial/moderate evidence for H1 and a $BF < 1/3$ indicates moderate/substantial evidence for H0, otherwise the evidence is ambiguous (i.e., the data are insensitive to test the hypothesis). Note that $BF > 3$ is approximately as conservative as $p<.05$, though alignment is not guaranteed.

Since Bayes Factors are sensitive to the choice of values for the predicted effects, and since there is some subjectivity in this, we also calculated "robustness regions" for each BF (indicated as *Robustness Region* = [x:y]). These show the range of predicted values we could

have used as the parameter (scale factor) for the model of H1 and still have drawn the same conclusion based on the cutoffs of BF $>$ 3 or BF $<$ 1/3. That is, x and y represent how low/high a value we could have used and still obtained a BF which was greater than 3 (if the BF is $>$ 3) lower than 1/3 (if the BF is lower than 1/3) or between 1/3 and a third (if the BF is between 1/3 and 3). Further details can be found in the Details of statistical analyses. Note we chose (and preregistered) to test one-tailed hypotheses (since the predictions are clearly directional) and throughout the reporting beta estimates are consistently reported as positive when they are in the predicted direction and negative when they are not. We also report the (more familiar) *p*-values, without interpreting them.

2.3. Results

Control trials were not included in analyses (average accuracy on these was 99% after participants' exclusion).

2.3.1. Four picture AFC task

Data are plotted in Fig. 7. Note that the very low performance with low-frequency items is due to the fact that participants overgeneralized in 66% of trials (i.e., they erroneously matched based on body shape).

There is strong evidence for higher overall performance for high than low frequency (β = 4.46, SE = 0.35, tails = 1, p = <.001, Predicted Effect = 1.97, BF = 2.2×10^{34}, Robustness Region = [0.03:>10]). Evidence for overall higher performance in FL than in LF tends toward the null but is ambiguous (β = 0.13, SE = 0.32, tails = 1, p = .34, Predicted Effect = 0.58, BF = 0.67, Robustness Region = [0:1.3]). The evidence for each of our primary predictions is ambiguous: Interaction between frequency and learning condition (β = 0.98, SE = 0.66, tails = 1, p = .07, Predicted Effect = 1, BF = 2.1, Robustness Region = [0:>10])

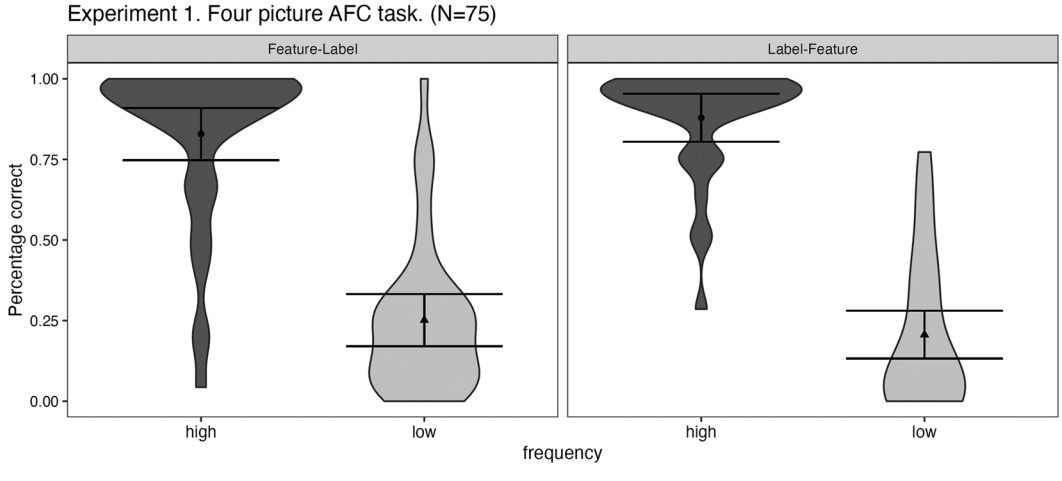

Fig. 7. Violin plots showing correct response rates split by frequency and learning-condition. Point shows mean and error bars 95% confidence intervals.

evidence for greater performance in FL for low-frequency items (simple effect) $\beta = 0.62$, $SE = 0.42$, tails = 1, $p = .07$, Predicted Effect = 1.07, BF = 1.74, Robustness Region = [0 : 6.89]). In sum, there is a strong overall effect of frequency, and the evidence that this effect is smaller in the FL condition tends in the direction of supporting the hypothesis (BF > 1) but is in the ambiguous range. The evidence also tends toward supporting the hypothesis of more accurate performance with low-frequency items in FL than in LF, but again is ambiguous.

2.3.2. *Four labels AFC task*

Data are in Fig. 8. Again, the very low performance with low-frequency items is due to participants overgeneralizing based on body shape (64% of low-frequency trials).

There is strong evidence for higher overall performance for high than low frequency ($\beta = 4.34$, $SE = 0.39$, tails = 1, $p = <.001$, Predicted Effect = 1.97, BF = 8.31×10^{25}, Robustness Region = [0.04:>10]). There is substantial evidence for the null hypothesis that overall performance is higher in FL than in LF ($\beta = –0.36$, $SE = 0.38$, tails = 1, $p = .825$, Predicted Effect = 0.58, BF = 0.33, Robustness Region = [0.56:inf]). Critically, there is also evidence for the null for our primary predictions: Interaction: $\beta = –0.7$, $SE = 0.66$, tails = 1, $p = .855$, Predicted Effect = 1, BF = 0.31, Robustness Region = [0.92:inf]); Simple Effect: $\beta = –.71$, $SE = 0.39$, tails = 1, $p = .965$, Predicted Effect = 1.07, BF = 0.13, Robustness Region = [0.37:inf]).

In sum, there is a strong overall effect of frequency and there is evidence against the hypothesis that this effect is smaller in the FL condition. There is also evidence against FL participants being more accurate with low-frequency items than those in LF participants.

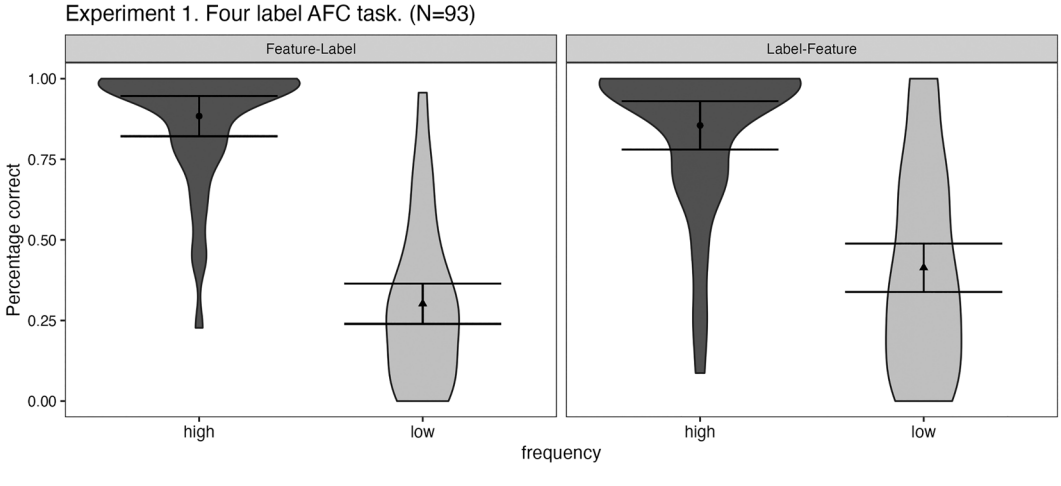

Fig. 8. Violin plots showing correct response rates split by frequency and learning-condition. Point shows mean and error bars 95% confidence intervals.

2.4. Discussion

Experiment 1 revealed strong frequency effects and high performance on high-frequency items in both the LF and FL conditions. Critically, we did not see evidence for either of the key hypotheses or the secondary hypothesis. However, the pattern of evidence differed for the two tests. In the four pictures test, although evidence did not meet our criteria for substantial, there was more evidence for H1 than for the null for both the interaction and the simple effect. In contrast, in the four labels test, we found evidence for the null for both effects. Thus, overall, this experiment does not confirm the key findings of Ramscar et al. (2010).

In considering why we did not find the predicted effects, it is important to note the extremely low performance for the items from the low-frequency subsets in both conditions and in both tests. Looking at the AFC tests, the majority of participants have scores well below 33% (which could be viewed as "chance" given that participants so rarely chose the control Fribble), which was not the case in either the original study or the unpublished replication study by Ramscar & McClure (2011)(see Fig. 3). The poor performance in this experiment was mostly because participants overgeneralized and picked the Fribbles based on body shape, indicating that they did not learn to discriminate low- from high-frequency items (there was also a significant data loss, with about one-third of participants not meeting our criteria). Given the floor effects observed in the low-frequency items, it is difficult to determine whether the differences in learning predicted for the FL and LF conditions are supported or not.

To account for the reduced performance, we considered whether there might be some unintended differences in the learning phase compared to the original lab experiments. One potential difference is that originally, the stimuli had a consistent size across participants, filling a large monitor screen. By contrast, stimulus size in this experiment depended on the size of the monitor of the participants' computer, introducing variability. The possibility that smaller stimuli did not allow participants to learn the discriminative features of the objects prompted Experiment 2: A follow-up replication with a monitor calibration phase.

3. Experiment 2

3.1. Methods

3.1.1. Participants

Two hundred and fifty-five participants were recruited using the same criteria as before. However, this time participants were recruited both through Prolificand through the University of Tübingen. Prolific participants were compensated as in Experiment 1, while the university participants participated for credit. The stopping procedure described in the Participants section was employed again.

3.1.2. Participant exclusion

Of the participants, 131 completed the Four Pictures task followed by the contingency task, while 124 completed the Four Labels task followed by the contingency task. Again, as preregistered, participants scoring below 80% on control items in the 4AFC tasks were

excluded (35 participants from Four Pictures and 20 from the Four Labels). Exclusions for missing data (over 10% in any task) resulted in a further 13 from the Pictures and 6 from Four Labels being excluded, leaving 83 participants in the Four pictures and 98 in the Four Labels task.

3.1.3. Stimuli and procedure

Experiment 2 was the same as Experiment 1 but added a resizing procedure to standardize stimuli size across monitors, using the `JsPsych -resize- plugin` (De Leeuw, 2015). Participants matched the on-screen size of a container to a credit card using a slider, allowing us to set a consistent Fribble size (160x160mm) regardless of monitor.

3.2. Results

3.2.1. Four picture AFC task

We plot the accuracy by frequency and condition in Fig. 9. Again, the very low performance with low-frequency items is due to participants overgeneralizing based on body shape—here in 69% of low-frequency trials.

There is strong evidence for higher overall performance for high than low frequency ($\beta = 4.89$, $SE = 0.37$, tails $= 1$, $p = <.001$, Predicted Effect $= 1.97$, BF $= 4.88 \times 10^{35}$, Robustness Region $= [0.03{:}{>}10]$). Evidence for overall higher performance in FL than in LF tends toward the null but is ambiguous ($\beta = 0.004$, $SE = 0.28$, tails $= 1$, $p = .505$, Predicted Effect $= 0.58$, BF $= 0.44$, Robustness Region $= [0{:}0.79]$). Turning to our primary predictions, critically, there is evidence for the interaction that crosses substantial criteria ($\beta = 1.35$, $SE = 0.71$, tails $= 1$, $p = .03$, Predicted Effect $= 1$, BF $= 3.58$, Robustness Region $= [0.64{:}2.26]$). The evidence for the simple effect of an FL benefit for low-frequency items was ambiguous $\beta = 0.68$, $SE = 0.42$, tails $= 1$, $p = .05$, Predicted Effect $= 1.07$, BF $= 2.14$, Robustness

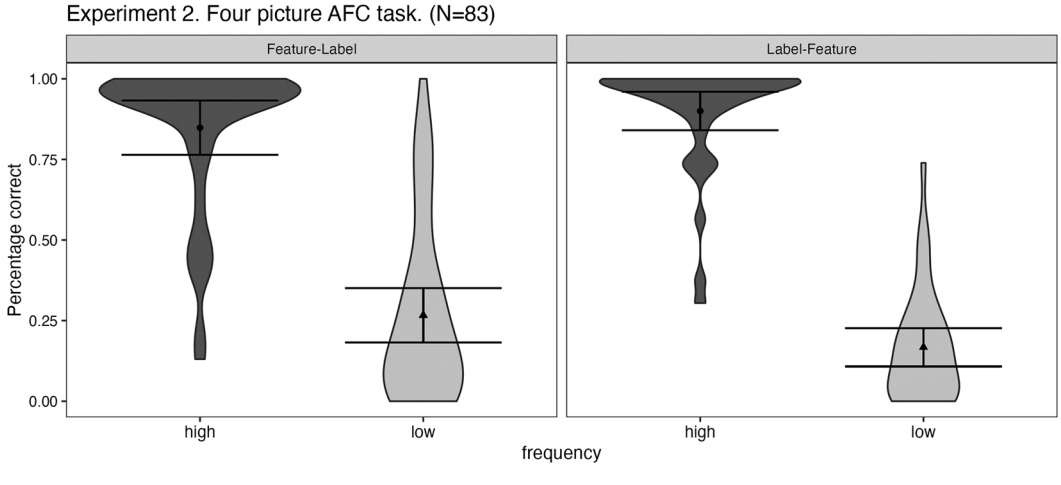

Fig. 9. Violin plots showing correct response rates split by frequency and learning-condition. Point shows mean and error bars 95% confidence intervals.

Region = [0:8.76]). In sum, there is a strong overall effect of frequency; however, in line with our prediction, there is evidence that this effect is smaller in the FL than in the LF condition. The evidence also tends in the direction of supporting the hypothesis of more accurate performance with low-frequency items in FL than in LF, but it is in the ambiguous range.

3.2.2. Four labels AFC task

We plot the accuracy by frequency and condition in Fig. 10. Again, the very low perfor-

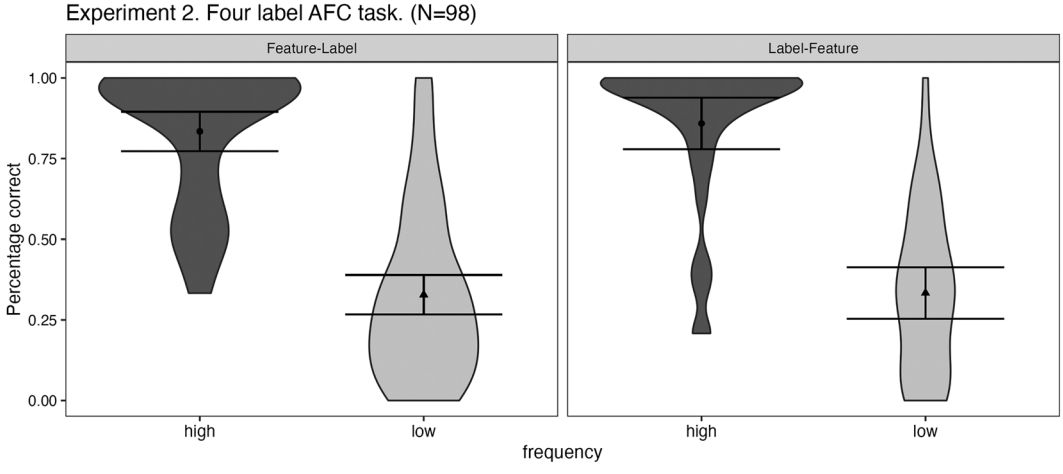

Fig. 10. Violin plots showing correct response rates split by frequency and learning-condition. Point shows mean and error bars 95% confidence intervals.

mance with low-frequency items is due to the fact that participants are overgeneralizing based on body shape (52% of low-frequency trials). There is strong evidence for higher overall performance for high than low frequency ($\beta = 3.86$, $SE = 0.31$, tails = 1, $p = <.001$, Predicted Effect = 1.97, BF = 4.37×10^{32}, Robustness Region = [0.03:>10]). There is substantial evidence for the null hypothesis that there is higher performance in FL than in LF ($\beta = -0.19$, $SE = 0.31$, tails = 1, $p = .735$, Predicted Effect = .58, BF = 0.32, Robustness Region = [0.56:inf]). For the primary predictions, the evidence for the interaction is ambiguous ($\beta = 0.39$, $SE = 0.59$, tails = 1, $p = .25$, Predicted Effect = 1, BF = 0.86, Robustness Region = [0:3.21]) and there is evidence for the null for the simple effect ($\beta = 0.003$, $SE = 0.34$, tails = 1, $p = .495$, Predicted Effect = 1.07, BF = 0.31, Robustness Region = [0.97:Inf]).

In sum, there is a strong overall effect of frequency. The evidence that this effect is larger in the LF condition than in the FL tends toward the null, but is ambiguous. There is evidence against the hypothesis that FL participants are more accurate with low-frequency items than LF participants.

3.3. Discussion

As in Experiment 1, performance was generally lower than in the original study, with most participants scoring at or below "chance" (33%) on low-frequency items due to extensive

overgeneralization. The pattern again differed between the two tests and in each case the direction of the evidence matched Experiment 1. However, in terms of our preregistered evidence thresholds, in the four pictures task, the evidence for H1 for the interaction between frequency and learning-condition reaches substantial criteria (BF $>$ 3), indicating that this test taps into the effects of stimuli ordering in training, as predicted. However, the evidence for the simple effect, although in direction of H1, was ambiguous. In the four labels test, evidence favored H0 for both key hypotheses (interaction and simple effect), with evidence for the null reaching the substantial criterion (BF $<$ 1/3) for the simple effect. Again, we did not see evidence for a main effect in either AFC test, aligning with the preregistered expectations and previous replications (we consider this further in the General Discussion).

Overall, the evidence pattern suggests quantitative rather than qualitative differences between Experiments 1 and 2. However, in terms of our preregistered inference criteria, only in Experiment 2 did we find evidence for H1 for one of the tests. Accordingly, on the basis of the analyses conducted so far, it is not possible to ascertain whether this reflects a genuine difference between the experiments. To determine this, we performed additional exploratory (nonpreregistered) analyses to see whether the evidence differs.

4. Comparing and combining data sets: Further exploratory analyses

There were two differences between the experiments: (1) standardizing stimulus size in the current experiment, and (2) different recruitment methods: Solely Prolific recruited in Experiment 1, versus a mix of Prolific and university students in Experiment 2. Although (2) was not intentional, the original experiments were done with university students, and as we noted in the introduction, the same learning mechanisms that predict the FLO effect also predict that populations with different levels of experience can be expected to perform differently in a learning experiment (albeit this fact is not usually acknowledged when training studies are reported). Given this, we ran analyses (1) over data from the Prolific-recruited students looking for a modulating effect of *stimuli-consistency-type* (i.e., inconsistent size vs. consistent size) and (2) on Experiment 2 data looking for a modulating effect of *recruitment-type* (Prolific recruited vs. university recruited).

We used the same approach as our previous preregistered analyses, except that we included the fixed effects of *stimuli-consistency-type/recruitment-type* and their interactions in the models. We were interested in evidence for interactions between each of these factors and the primary tests of the FLO effect—that is, to see if there was a three-way interaction of *stimuli-consistency-type/recruitment-type* by *learning condition* by *frequency* and if there was an interaction of *stimuli-consistency-type/recruitment-type* by *learning condition* specifically for low-frequency items. While we again used Bayes Factors, these were two-tailed as we did not start with predictions in one direction (see Details of statistical analyses). For both test tasks, in every case, the evidence for an interaction was around 1, that is, highly ambiguous (BFs between 0.81 and 1.12).[9]

Since these analyses find no evidence that the changes we made between the experiments modulated the FLO effect, we reran the analyses for the four pictures and four label tasks

using our preregistered methods but using pooled data sets from Experiments 1 and 2. For the four picture test, this exploratory analysis showed substantial evidence for H1 for both the interaction between *frequency* and *learning condition* ($\beta = 1.18$, $SE = 0.49$, tails = 1, $p = .005$, Predicted Effect = 1, BF = 9.18, Robustness Region = [0.27: 5.74]) and the simple effect: ($\beta = 0.66$, $SE = 0.3$, tails = 1, $p = .007$, Predicted Effect = 1.07, BF = 5.11, Robustness Region = [0.2: 2.08]). For the four labels test, there was evidence for the null for both the interaction between *frequency* and *learning condition* ($\beta = –0.12$, $SE = 0.44$, tails = 1, $p = .695$, Predicted Effect = 1, BF = 0.33, Robustness Region = [1:inf]) and the simple effect ($\beta = –0.35$, $SE = 0.26$, tails = 1, $p = .542$, Predicted Effect = 1.07, BF = 0.11, Robustness Region = [0.31:inf]).

4.1. Summary of findings from these analyses

Our exploratory analysis of the combined data from the experiments provides substantial evidence *for* both of the primary predictions in the four picture task, and *against* (relative evidence for the null) both primary predictions in the four labels task. In other words, we see patterns of evidence which are qualitatively in line with each of the experiments individually, but the weight of evidence in this larger sample is now sufficient to cross our "substantial" threshold in all cases. Although not preregistered, this can be interpreted bearing in mind that Bayes Factors remain a valid measure of evidence in a combined sample, with larger samples expected to yield stronger evidence. We also point out that the sample sizes in our preregistered experiments ended up being somewhat smaller than originally planned due to a data collection error (see Data loss and deviations from preregistration).[10]

These analyses revealed no significant impact from changes in stimulus size or recruitment methods on the 2AFC tests between Experiments 1 and 2. Since these results are ambiguous (i.e., we do not have substantial evidence for either H1 *or* the null), we cannot draw strong conclusions. We also note that in the contingency data (see Contingency and additional subset analysis), we *did* find some evidence for an interaction with recruitment type (in the direction of stronger evidence for the simple effect for low-frequency items in the university-recruited than in the Prolific-recruited individuals). However, for the AFC tests, the key conclusion must be that our paradigm and sample size are not sufficiently sensitive to test for these differences between experiments.

5. General discussion

Ramscar et al. (2010) presented an analysis and simulations that predicted stimuli sequencing effects in error-driven learning. In the current work, we set out to look for evidence of these key effects in two preregistered replications conducted online, which differed from one another in terms of stimuli size and recruitment methods. In both experiments, we saw different patterns of results for the two AFC tests. In our preregistered analyses: For the four pictures task, the evidence tended toward H1 for both the interaction and the simple effect in both experiments, crossing the substantial criteria (BF > 3) for the interaction in Experi-

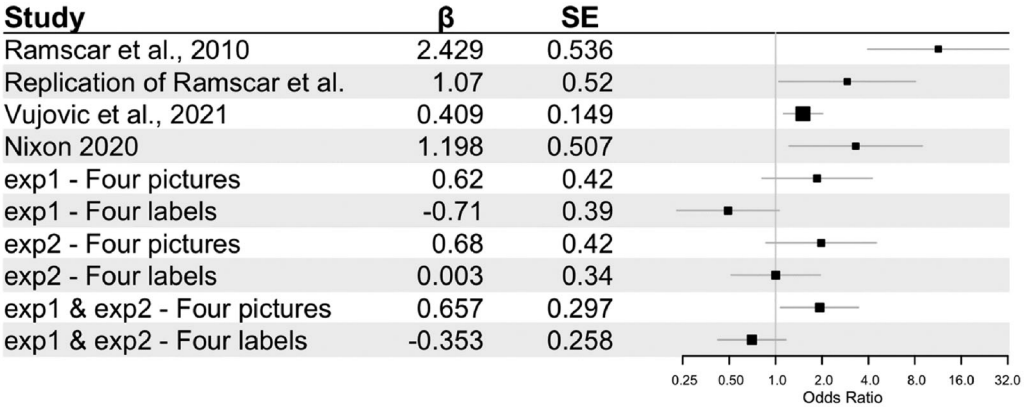

Fig. 11. Effect sizes for the simple effect—that is, the benefit of the FL over the LF condition specifically for low-frequency items; beta and SE are log odds; right shows forest plot as odds ratio.

ment 2. In the four labels task, in both experiments, evidence tended toward the null for both the interaction and the simple effect, crossing the substantial criteria (BF $<$ 1/3) for both in Experiment 1, and for the simple effect in Experiment 2. In exploratory analyses over the combined data, there was substantial evidence for both the interaction and the simple effect in the four pictures task, and evidence for the null for both effects in the four labels task.

The evidence from the four picture task is thus (broadly) in line with that from the original study–and, critically, with the predictions of computational models implementing error-driven learning—albeit with moderate evidence levels. This adds to the existing body of evidence supporting the theory (e.g., Nixon, 2020; Vujović et al., 2021). However, the null results in the four labels test suggest that observations of the FLO effect can be affected by the test task. We discuss this point below in the section Test-task effects.

Focusing on the four pictures task, while the results are overall consistent with predictions, there are important differences compared to the original experiment. One key observation is that learning of the low-frequency items in the FL condition is much lower than in the original study, and thus the benefit over the LF condition is much smaller. This can be clearly seen in the comparison of effect sizes in Fig. 11: The effect is much smaller than in the original, though roughly consistent with the more recent studies by Vujovic and Nixon—which were also conducted online.

Not-withstanding that effects are generally smaller in replications, it is worth considering why learning of low-frequency items is not stronger in FL in our study. We saw that this poor performance reflects overgeneralization based on body-type: Almost 70% of the time, our participants ignored the discriminating features and erroneously picked the foil Fribble whose body-type was frequently associated with the label. Vujović et al. (2021) showed that a similar pattern of overgeneralization also occurs during the early stages of training in computational simulations. Critically, in these simulations, overgeneralization in FL decreases as exposure to the input increases, due to increased opportunity to use prediction error to unlearn the erroneous cue. This suggests that (at test) our online participants may have reached an average

point in learning than is different to those in the original study. We return to this point in the section Implications for conducting online training experiments.

Another difference is that the original study found an FL benefit for both low- and high-frequency items (reflected in a main effect), but in the current study that was only seen for low-frequency items (as originally predicted by Ramscar et al., 2010). Indeed—though we did not analyze high-frequency items in isolation—the data trends in the direction of an LF benefit consistent with patterns seen in Nixon (2020) and Vujović et al. (2021) (this is why a main effect was preregistered only as a secondary hypothesis). Further, Vujović et al. (2021) pointed out that stronger LF learning for high-frequency items is consistent with the model if participants base their choice on the picture whose features are (in sum) most positively associated with the label, assuming positive associations are mapped to raw associative weights in the model (since though negative weights for the incorrect label are greater in FL, *positive* weights for the correct label are greater in the LF condition). In contrast, the behavior of the original participants in showing an FL benefit with these items is more consistent with the use of a "choice" rule in the AFC task, in which in determining the probability that a label matches a particular set of features also takes into account the match with the other label (though in interpreting the findings with high-frequency items in the original, it should also be noted that responses are close to ceiling). For a fuller discussion of these issues, see Vujović et al. (2021).

5.1. Implications for conducting online training experiments

This paper adds to the growing body of research conducted online rather than in laboratory. We suggested above that our results (at least with the four pictures tasks) are consistent with learners being at an earlier average point in learning as compared to those in the original study (and, though to a lesser extent, in the unpublished lab-based replication by Ramscar & McClure, 2011). Vujović et al. (2021) suggested something similar for their online study.

Could the move to an online platform have led to this difference? Recent discussions have highlighted numerous factors affecting online versus lab experiments (Gagné & Franzen, 2021; Rodd, 2019). One difference is that every online participant uses their own hardware. In this study, we explored the impact of monitor variability: Experiment 1 did not take into account the fact that monitor size could affect stimuli size, which was corrected in Experiment 2. Our subsequent analyses did *not* find substantial evidence that this factor modulated the effects of interest (although we also did not find substantial evidence for the null, meaning that we cannot draw very strong conclusions about this methodological aspect). However, we did not consider other hardware aspects, such as processing capacity and network quality, which could potentially affect stimuli presentation (including the critical timing differences between FL and LF). Future work could mitigate against this by restricting the experiment to participants with computers that meet specific hardware criteria.

Another key difference in online experiments is the lack of control over participants' environments. Participants in lab experiments may have less risk of distractions and may be more motivated and attentive if they believe they are being observed. We mitigate against this to some extent via our exclusion criteria—they must have least paid sufficient attention to learn

the easy control category and be sufficiently attentive during the test not to have too many "timed-out" trials. However, we cannot rule out that the different environments led to reduced attention to the input stimuli.

A final difference between lab-based experiments and online experiments —which we would argue is particularly relevant for *learning* paradigms —is in recruitment procedures, which can change the profile of participant populations. Platforms such as Prolific provide a mechanism for wide-scale recruitment across (relatively) diverse participants; however, while recruitment criteria can be stipulated (e.g., age range, English proficiency, lack of learning disabilities, etc.), they cannot be guaranteed. Critically, as we noted at the outset, it is well established that in error-driven learning, while the *association rates* between cues and outcomes promote the learning of positive associations, what actually gets learned also depend on factors that tend to inhibit learning: The *background rates* of cues, and the prior predictability of outcomes in context (*blocking*). Not only do these factors interact to produce cue competition (which led Ramscar et al., 2010 to predict the FLO effect), they also predict that the outcome of training in any given task is itself a function of prior experience. This is best explained by considering a far simpler learning paradigm than the one tested here. In paired-associate learning (PAL), participants learn word pairs and then recall one word—the target, when given another—the cue. Ramscar, Hendrix, Love, and Baayen (2013); Ramscar, Hendrix, Shaoul, Milin, & Baayen (2014); Ramscar, Sun, Hendrix, and Baayen (2017) have shown that age-related "declines" in adult PAL performance can be accurately *predicted* by error-driven models that estimate the learnability of word pairs as a function of learner's previous experience with the actual words in question, modulated by the three factors that promote and inhibit learning described above. These models/factors explain why adults of all ages are better at learning "easy" PAL pairs like *baby-cries* than "hard" pairs like *obey-eagle*: because the high association rate of *baby-and-cries* successfully promotes learning of this association. They can explain why harder word pairs become proportionally far more difficult to learn as experience grows: because *obey* rarely occurs with *eagle*, increased knowledge of background rates inhibits learning. Further, they successfully predicted that older age-matched L2 speakers would outperform native German speakers when asked to learn German PAL pairs (Ramscar et al., 2017): because on average native speakers have better knowledge of background rates than bilinguals (see also Qiu & Johns, 2020). All these findings highlight a fact that was overlooked by Ramscar et al. (2010): that the outcome of learning can *never* be independent of prior experience.

To understand how experience impact the Fribbles task, consider that successful learning of the low-frequency exemplars involves unlearning a highly salient distractor feature. This is in contrast to most categories of natural objects, which actually tend to share common features (Torralba & Oliva, 2003). This must inevitably cause people exposed to natural categories to increasingly learn that common features are salient to category learning, and given this, it follows that learners with more experience of the world (i.e., older learners) ought to find the fribble task harder than learners with less experience (who will have had less training on actual co-occurrences between common features and natural categories). In a similar vein, learners with more experience of taking psychological experiments will have more experience of the ways in which such experiments can contain manipulations which violate real-world

expectations. Although we did not collect age in our study, since the maximum age was 35, it seems likely that the participants are on average older than students in the original study, who were undergraduate psychology students with an average age of 19 (meanwhile, our Tübingen participants, many of whom were masters' students, were mainly linguists). These differences might begin to explain the performance of the original participants when it came to learning to discriminate the low-frequency Fribbles.

Speaking against this account is the fact that we did *not* find evidence for performance differences between our student participants and our Prolific-recruited participants in our AFC tests. On the other hand, although we did not see evidence for this difference between students and nonstudents in the present study, we also did not see evidence for *no* difference. Moreover, in the analyses of the contingency data reported in Appendix C, and Fig. S6 in particular, we did see evidence of better learning of the low-frequency discriminating features in students than in Prolific-recruited participants. Further exploratory analyses in Appendix B are consistent with stronger learning of the discriminating features for low-frequency items in the student group. The more general point is that there may be differences in experience between the current participants and those in the original experiment, and if there are, it is highly likely that this will impact their performance, which suggests that researchers should be more circumspect in assuming the generality of their findings when it comes to results from training paradigms (it is notable in this regard that the *biggest* by-decade change in PAL performance observed across the lifespan is between adults in their 20s and 30s; Ramscar et al., 2013).

5.2. Test-task effects

A surprising outcome in this study was that we did *not* find evidence for our predicted effects in the four labels task in either experiment. In fact, when combining data sets, we see evidence for the null. Assuming that participants learn as predicted—as shown by the Four Pictures task and the original study—why were the primary predictions not met in the four label task?

It is notable here that performance with low-frequency items was stronger in the four labels task, and particularly in the LF condition. A possible explanation for this is that the test itself might result in cue competition, and this could wipe out differences from training condition. Recall that in the four labels task, a single Fribble is presented simultaneously with four label choices. Thus, a natural way to approach the task would be to consider the picture against each label in turn, from left to right. For example, if the Fribble picture is a low-frequency "dep," the participant will see this alongside all four labels. If they look back and forth between picture and labels, when they reach "tob," this could itself provide an error signal, helping them to discard "tob" in favor of the correct label "dep."

Speaking against this explanation is that, in principle, something similar could occur in the four pictures task: Participants could in turn attempt to match each of the four Fribble pictures against the single label. However, since the complexity of the Fribble pictures would make it much harder to do this within the 3500 ms available, it seems that the time constraint—which was the same in both tasks—may have been more effective in blocking this process in the four

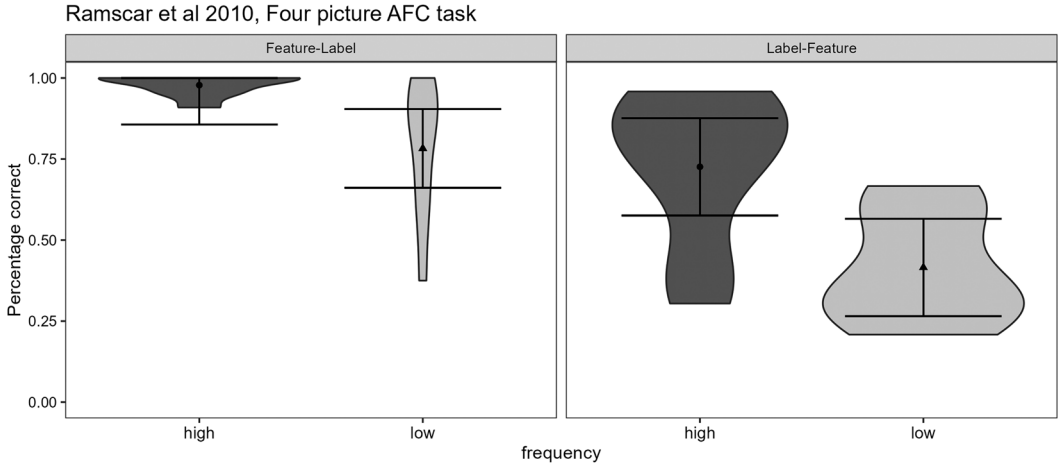

Fig. 12. Data from the Four Pictures task in Ramscar et al. (2010).

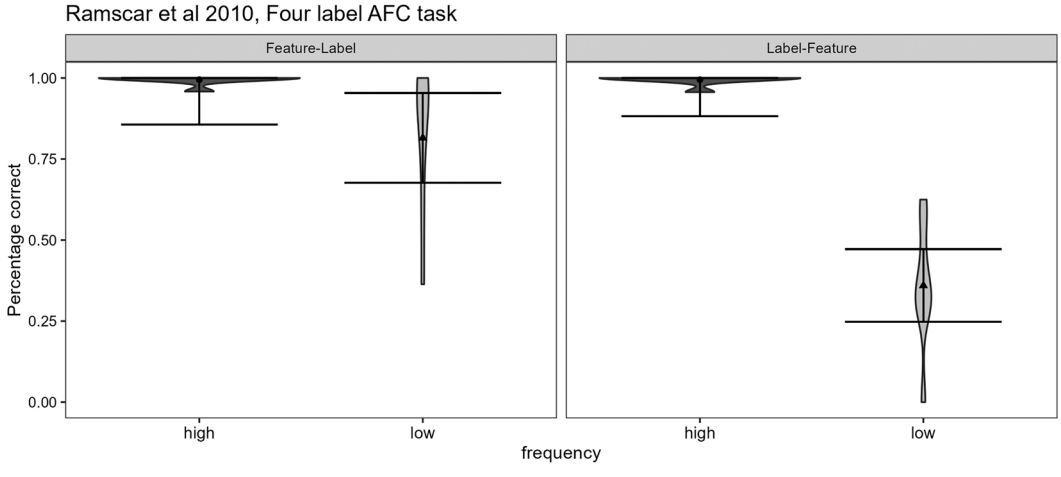

Fig. 13. Data from the Four Labels task in Ramscar et al. (2010).

pictures task. A supplementary analysis comparing the likelihood of a "timed-out" trial for the two responses showed that this was greater for the four picture task—four picture: 4.8%, four labels: 2.5%; fixed effect of test task in the logistic mixed effect model: $\beta = 0.74$, $SE = 0.15$, $z = 5.3$ $p < .001$—supporting the claim that the time limit was indeed more constraining. Thus, it seems responses in the four pictures task may better reflect the implicit knowledge actually gleaned in training.

How does this explanation fit with previous studies? Ramscar et al. (2010) reported their results across data from the combined AFC tasks; however, we have since been able to look at their data separately for the two tasks (Figs. 12 and 13), it is clear that performance for low-frequency items is stronger in FL than in LF for *both* tests. One possibility for the difference—

consistent with our discussion above of how learners in the current study might be at a different point in learning—is that learners in the original study had progressed to a stage where the FL benefit was sufficiently strong that it could not be confounded by learning during the test.

Interestingly, Vujović et al. (2021) also included different types of AFC tests and also found stronger patterns in some tests than others. However, the strongest results in that paper were in fact found in the task which, at first glance, is more similar to the current four labels test. Specifically, there was a single Fribble and two phrases to choose from. However, recall that Vujović et al. (2021) used *audio* rather than *written* stimuli, and in this task, participants had to listen to the two phrases sequentially before making their choice. Therefore, participants did *not* have the same opportunity to deliberately consider the Fribble against a set of labels in turn.

More generally, this result shows again that the choice of test task type can have unexpected consequences when trying to tap into these learning effects. Finally, we note that there are also further test-type differences when we consider the results of the additional contingency test.[11] The results of this new test were mixed, with evidence for the null for our predictions in some cases, and for H1 in others, and exploratory analyses finding tentative evidence of differences between learner groups. This is fully discussed in the relevant online Appendix (C)—here, we note that this is again consistent with a modulating role of test-type which affects our ability to tap into the effects which signature discriminative learning. The more general conclusion for training experiments is that effects gleaned from short periods of exposure must inevitably depend on the ability of a test to detect any effects of training.

5.3. Data loss and deviations from preregistration

While we view the use of preregistration as a strength of this work, it does lay bare deviations and procedural errors which must be acknowledged. We deviated from our preregistration in the size of the final sample for both Experiments 1 and 2. Our preregistration said that unless we had sufficient evidence from a smaller sample, we would continue collecting data up to a maximum of 100 participants in each of the 4AFC tasks in each *net of exclusions*. Exclusions were specified as depending both on participants' performance in control tasks and amount of missing data for their AFC test (no more than 10% timed-out trials being permissible). Unfortunately, an error in the interim analyses we conducted during data collection meant that exclusions were in fact only identified on the basis of control trials, not missing data. This was not identified until after our data collection for both experiments was completed, meaning that those participants had not been replaced. Thus, our final data sets in each experiment are less than the planned maximum. Although we could theoretical rectify this by collecting more data in each experiment, in the end we were able to gain a larger sample by combining across the experiments. Since this combined data set actually led to a clear pattern of results for our key hypotheses for each AFC test, we did not feel that additional data collection in individual hypotheses was a good use of resources at this point.

Another deviation is that we preregistered an analyses plan for Experiments 1 and 2 which included a set of specific priors for the Bayes Factors, basing these on data from Ramscar & McClure (2011). We later found a small error in our initial analysis of that experiment

which (slightly) changed the relevant values. We decided to use the new values (i.e., those based on more accurate analyses) as the priors for the later experiments, rather than sticking with the preregistered values. Critically, the differences are small and there is no place where using these rather than the preregistered values led to qualitatively different results (i.e., there was no case where this affected the direction of the evidence as for/against our hypotheses, or where we claim substantial evidence for H1 or H0 where the use of the original values would have indicated ambiguous evidence). The reader can verify this by checking that the preregistered values (see footnote 4 above) fall within the robustness regions in each case (it is also shown in our analyses script).[12]

Open practises

Data, analysis, and preregistration of all experiments are available on OSF. All analyses are carried out in R, and all analysis scripts are on OSF.

Funding

This work was supported by the Leverhulme Research Project RPG-2019-160 "Language Learning as Expectation: a Discriminative Perspective" awarded to the last author. The second author was supported in part by a grant from the Deutsche Forschungsgemeinschaft (DFG 381713393).

Competing interests

We have no competing interests.

Open Research Badges

 This article has earned Open Data and Open Materials badges. Data and materials are available at https://10.17605/OSF.IO/ZH8AJ

Notes

1 Thank you to an anonymous reviewer for suggesting this analysis.
2 We are not able to present separate plots/analyses for the two tests since labeling of the testing conditions in the data files was unclear and information about the order in which the tests were administered to each participant has been lost.
3 In this figure, we have plotted the simple effect of a benefit for FL over LF specifically for low-frequency items (i.e., the effect most clearly predicted by the computational

model which has been most consistently found; Hopp et al. is not included due to the differences in the design).

4 We also preregistered additional analyses over the tests combined, following the original study. Since we found different patterns of results for the two tests, we do not report these, though they can be viewed on OSF.

5 Methods including exclusion criteria, optional stopping criteria, and analyses were preregistered on OSF. Any deviations from this plan are noted in the text.

6 Dienes (2016) states that optional stopping does not affect Bayes factors as it does *p*-values in frequentist analyses. Frequentist methods may yield significant *p*-values under optional stopping even if the null hypothesis (H0) is true due to *p*-value fluctuations. In contrast, Bayes factors, which are symmetric, will increase if H0 is false and decrease if H0 is true, providing a valid measure of evidence regardless of the stopping rule used. This respects the "stopping rule principle" which ensures that the evidence is derived solely from the data, not the method of collection. Thus, Bayes Factors remain a valid measure of evidence regardless of data collection procedure.

7 Rapid presentation was used to prevent explicit learning and strategizing.

8 Note that our approach means that we are only using predicted effect sizes as part of the hypothesis testing process, they are *not* influencing our *estimations* of effects in the current study (which are taken from the mixed effect models which do not incorporate priors into estimation).

9 Four picture test, interaction stimuli-consistency-type with learning condition by frequency $\beta = 0.352$, $SE = 1.19$, tails = 2, $p = .768$, Predicted Effect = 0.501, BF = 0.928, Robustness Region = [0: 3.51]; four picture test, interaction stimuli-consistency-type by learning condition (for low-frequency items) $\beta = –0.643$, $SE = 0.682$, tails = 2, $p = .346$, Predicted Effect = 0.537, BF = 0.931, Robustness Region = [0:3.05]; four picture test, interaction recruitment-type with learning condition by frequency $\beta = 0.028$, $SE = 1.426$, tails = 2, $p = .984$, Predicted Effect = 0.501, BF = 0.984, Robustness Region = [0: 4.03]; Four picture test, interaction recruitment-consistency-type by learning condition (for low-frequency items) $\beta = 1.258$, $SE = 0.807$, tails = 2, $p = .119$, Predicted Effect = 0.537, BF = 1.21, Robustness Region = [0:8.02]; Four label test, interaction stimuli-consistency-type with learning condition by frequency $\beta = 0.863$, $SE = 1.281$, tails = 2, $p = .5$, Predicted Effect = 0.501, BF = 0.96, Robustness Region = [0: 4.56]; four label test, interaction stimuli-consistency-type by learning condition (for low-frequency items) $\beta = 0.065$, $SE = 0.743$, tails = 2, $p = .93$, Predicted Effect = 0.537, BF = 0.811, Robustness Region = [0:2.11]; four label test, interaction recruitment-type with learning condition by frequency $\beta = 0.321$, $SE = 1.25$, tails = 2, $p = .797$, Predicted Effect = 0.501, BF = 0.932, Robustness Region = [0:3.65]; four label test, interaction recruitment-consistency-type by learning condition (for low-frequency items) $\beta = 0.992$, $SE = 0.724$, tails = 2, $p = .171$, Predicted Effect = 0.537, BF = 1.12, Robustness Region = [0:5.41].

10 We also confirmed the modulating role of test-type in this combined sample, that is, the results are different in the Four Pictures compared with Four Labels test: There is evidence that both the frequency by learning condition interaction and the simple effect

are larger in the four pictures test than in the four labels test (interaction: $\beta = 1.391$, $SE = 0.647$, tails = 1, $p = .016$, Predicted Effect = 0.501, BF = 3.412, Robustness Region = [0.44:3.97]; simple effect: $\beta = 1.006$, $SE = 0.388$, tails = 1, $p = .005$, Predicted Effect = 0.537, BF = 10.443, Robustness Region = [0.2: 7.33).

11 We added it at the end of the current experiment and report in the online Appendix C: Contingency and additional subset analysis.

12 Note that more generally, the inclusion of robustness regions throughout the paper mitigates against a potential criticism of the Bayes Factor, which is that any choice of priors is subjective. Given the regions, a researcher who thinks a different choice of prior would be more appropriate can easily check whether this would affect our conclusions.

References

Apfelbaum, K. S., & McMurray, B. (2017). Learning during processing: Word learning doesn't wait for word recognition to finish. *Cognitive Science*, *41*, 706–747.

Arppe, A., Hendrix, P., Milin, P., Baayen, R. H., Sering, T., & Shaoul, C. (2018). ndl: Naive discriminative learning. R package version 0.2.18. https://CRAN.R-project.org/package=ndl.

Baguley, T., & Kaye, W. (2010). Review of: Understanding psychology as a science: An introduction to scientific and statistical inference, by Z. Dienes. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 695–698.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

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

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan International Higher Education.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89.

Dye, M., Ramscar, M., & Suh, E. (2011). For the price of a song: How pitch category learning comes at a cost to absolute frequency representations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Eitel, A., & Scheiter, K. (2015). Picture or text first? Explaining sequence effects when learning with pictures and text. *Educational Psychology Review*, *27*(1), 153–180.

Finley, A. J., & Penningroth, S. L. (2015). Online versus in-lab: Pros and cons of an online prospective memory experiment. *Advances in Psychology Research*, *113*, 135–162.

Gagné, N., & Franzen, L. (2021). How to run behavioural experiments online: Best practice suggestions for cognitive psychology and neuroscience.

Hoppe, D. B., van Rij, J., Hendriks, P., & Ramscar, M. (2020). Order matters! Influences of linear order on linguistic category learning. *Cognitive Science*, *44*(11), e12910.

Hupp, J. M., Sloutsky, V. M., & Culicover, P. W. (2009). Evidence for a domain-general mechanism underlying the suffixation preference in language. *Language and Cognitive Processes*, *24*(6), 876–909.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Kopp, B., & Wolff, M. (2000). Brain mechanisms of selective learning: Event-related potentials provide evidence for error-driven learning in humans. *Biological Psychology*, *51*(2–3), 223–246.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Morey, R. D., & Lakens, D. (2016). Why most of psychology is statistically unfalsifiable.

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*, 104081.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. https://doi.org/10.1126/science.aac4716

Powell, M. (2009). The bound optimization by quadratic approximation (BOBYQA) algorithm for bound constrained optimization without derivatives. Technical report, Cambridge, England.

Qiu, M., & Johns, B. T. (2020). Semantic diversity in paired-associate learning: Further evidence for the information accumulation perspective of cognitive aging. *Psychonomic Bulletin & Review*, *27*(1), 114–121.

Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, *46*(4), 377–396.

Ramscar, M., Dye, M., Gustafson, J. W., & Klein, J. (2013). Dual routes to cognitive flexibility: Learning and response-conflict resolution in the dimensional change card sort task. *Child Development*, *84*(4), 1308–1323.

Ramscar, M., Dye, M., Popick, H. M., & O'Donnell-McCarthy, F. (2011). The enigma of number: Why children find the meanings of even small number words hard to learn and how we can help them do better. *PLoS One*, *6*(7), e22501.

Ramscar, M., Hendrix, P., Love, B., & Baayen, R. H. (2013). Learning is not decline: The mental lexicon as a window into cognition across the lifespan. *Mental Lexicon*, *8*(3), 450–481.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*(1), 5–42.

Ramscar, M., & McClure, S. M. (2011). Manipulating information structure as a method of localizing information processing in the brain. In *Poster presented at the 18th Annual Meeting of the Cognitive Neuroscience Society; April 2011*.

Ramscar, M., Sun, C. C., Hendrix, P., & Baayen, H. (2017). The mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the "cost" of learning, not cognitive decline. *Psychological Science*, *28*(8), 1171–1179.

Ramscar, M., Thorpe, K., & Denny, K. (2007). Surprise in the learning of color words. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 29.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive Science*, *34*(6), 909–957.

Recorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). New York: Appleton- Century-Crofts.

Rodd, J. (2019). How to maintain data quality when you can't see your participants. *APS Observer*, *32*.

Silvey, C., Dienes, Z., & Wonnacott, E. (2021).Bayes factors for mixed-effects models. *PsyArXiv*. https://doi.org/10.31234/osf.io/m4hju

St Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, *33*(7), 1317–1329.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391.

Vujović, M. (2020). *Exploring language learning as uncertainty reduction using artificial language learning*. PhD thesis, University College London.

Vujović, M., Ramscar, M., & Wonnacott, E. (2021). Language learning as uncertainty reduction: The role of prediction error in linguistic generalization and item-learning. *Journal of Memory and Language*, *119*, 104231.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1: Four pictures task. Combined data from the two experiments split according to participant recruitment type.

Fig. S2: Summed weights from the simulation plotted (averaged over 50 runs).

Fig. S3: Example of high-frequency "match" and low-frequency "mismatch" trial for the category tob used in the Contingency Judgment task and in the modeling.

Fig. S4: Contingency task. Experiment 1. Violin plots showing average score (between –100 +100) for match and mismatch trials for high and low frequency in the two learning conditions.

Fig. S5: Contingency task. Experiment 2. Violin plots showing average score (between –100 +100) for match and mismatch trials for high and low frequency in the two learning conditions.

Table S1: Results of Bayes Factors tests for each primary prediction in each experiment.

Fig. S6: Contingency task. Combined data from the two experiments split according to participant recruitment type.

Fig. S7: Proportion of runs of the simulations (1000 runs per *N* participants) where the BF met various criteria for establishing evidence for H1/H0.