REPLICATION RESEARCH

Replication Research (R2), 2026, vol 2
10.17879/replicationresearch-2026-9615
Article type: *Replication*
Discipline: *Psychology*
Published under a CC-BY 4.0 license

Are Redundant Labels Helpful during Category Learning? Replication of Lupyan, Rakison, and McClelland (2007)

Caroline Kaicher[1], Ellen M. Markman[1], Laura Gwilliams[1,2,3]

[1]Department of Psychology, Stanford University, Stanford, CA, USA
[2]Wu Tsai Neurosciences Institute, Stanford, CA, USA
[3]Stanford Data Science, Stanford University, Stanford, CA, USA



ABSTRACT. Labels are often advantageous for category learning. However, this may not be true for children as they learn subordinate, fine-grained object categories. In preparation for future research on subordinate category learning in children, we aimed to refine a procedure investigating the label advantage during subordinate category learning in adults. Thus, we conducted an online replication of Experiment 2 from Lupyan, Rakison, and McClelland (2007), which compares category learning under an auditory or written label condition versus a no-label and nonlinguistic cue condition in adults. We failed to replicate the key result, that those who received labels learned the categories faster.

KEYWORDS. category learning; language; object categorization

LAY SUMMARY. This study repeated an earlier experiment testing whether words help people learn to distinguish between similar categories of objects (for example, two types of aliens being labeled as "leebish" or "grecious"). Unlike in the original study, the researchers found that such labels did not help adults learn the categories faster, suggesting that their benefit for learning fine differences between objects may be limited.

CITATION. Kaicher, C., Markman, E. M., & Gwilliams, L. (2026). Are Redundant Labels Helpful during Category Learning? Replication of Lupyan, Rakison, and McClelland (2007). *Replication Research*, *2*. https://doi.org/10.17879/replicationresearch-2026-9615

Background

Objects can be organized into categories, on the basis of shared traits and characteristics (Murphy, 2010). These categories allow humans to carve up the world around them into meaningful classes that are necessary for concept formation and knowledge generalization. Language is an important source of category information, because linguistic labels (i.e., words) act as category markers, informing interlocutors of the category a referent does, and does not, belong to. For example, an object could be an exemplar of the dog category, and therefore implicitly not of the wolf, cat, or bowl category. In turn, labeling objects allows humans to infer properties of the object based on the category membership communicated in the label. For example, once an object is identified as a *dog*, one can infer that it likely barks, has fur, and walks on four legs.

Labels can also be useful during the process of learning object categories. It is possible to learn an object category without having a label for it, for instance, by coming to treat objects as the same kind based on having several features in common (Mervis, 1987). However, labels provide multiple advantages. They provide a simple way to represent perceptually complex distinctions, and when used in a communicative setting, they serve the pragmatic function of signaling that a certain category is relevant or important. It may come as a surprise, then, that labels do not assist humans in all category learning scenarios. Research suggests that the benefit of having labels during category learning differs across different types of categorization and age groups (e.g., Deng & Sloutsky, 2012; Waxman, 1990).

One interesting case where the role of labels is unclear is in the learning of subtle, fine-grained categories in children and adults. These subtle categories are often referred to as subordinate categories that exist within higher-level categories. For example, crow and raven are subordinate categories within the higher-level category bird. Subordinate categories are particularly difficult to learn because subordinate category members share many of the same properties as other members of the higher-level category of which they are a part (Rosch et al., 1976). This makes subordinate categories difficult to contrast with each other, and easily confused (Mervis & Crisafi, 1982). To extend the previous example, it might be difficult for a bird-watching novice to distinguish between a crow and a raven. However, learning such fine-grained categories can result in a high level of precision and even expertise in object knowledge (Tanaka et al., 2005); a bird-watching expert would have no problem distinguishing between a crow and a raven.

One might think that labels would be useful while learning fine-grained categories because words can signal that two objects belong to different subordinate categories, despite being perceptually similar. There is evidence for this position in adults; for example, Boutonnet et al. (2013) found that event-related potentials (ERPs) demonstrate distinct neural responses for perceptually similar objects that are assigned different labels. However, in children, labels do not provide a consistent behavioral advantage when learning fine-grained categories. For instance, Waxman (1990) found that novel nouns made subordinate categorization more difficult for 3– to 4 year-old children. Meanwhile, other studies have shown that 3-year-olds are not hindered by novel nouns when they are given specific contrasting information about the subordinate categories, such as information about

function, location, or perceptual attributes, or when the novel category labels are anchored to familiar higher-level labels (Mervis et al., 1994; Waxman, 1990; Waxman et al., 1991).

In adults, the seminal study Lupyan, Rakison, and McClelland (2007) showed that labels help adults learn fine-grained object categories faster than when they have no labels, or have other nonlinguistic cues. The authors utilized "alien" stimuli that participants learned to categorize by the shape of the alien's head and base. While not explicitly stated by the authors, these can be considered subordinate categories given that they exist within the larger category *alien* and have quite subtle distinctions. This study is particularly compelling because the labels were redundant, in other words they did not provide additional information to the participants about the category distinctions. Specifically, the authors told participants that one category of aliens should be approached and the other should be avoided, making the category labels redundant and constituting extra information to be learned. The study included 2 experiments, with the second experiment replicating and expanding upon the first experiment. In Experiment 2 of the study, the authors compared four conditions of the category learning task, auditory label, written label, location, and no label, while in Experiment 1 only written label and nolabel conditions were used. In the auditory label and written label conditions, participants were given the nonsense words *leebish* and *grecious* as labels for the two alien categories. In the location condition, participants associated the two alien categories as living in different locations on the planet, acting as a nonlinguistic cue to contrast the label conditions with. Lastly, in the no-label condition, participants did not receive any labels or get any other cues to associate with the categories. The authors found that there was a label advantage for the participants, with those in the two label conditions learning the categories faster than those in the location and no-label conditions. This suggests that for adults, there is something "special" about having a label to associate category exemplars with as they are learning fine-grained categories.

Replicating Experiment 2 of Lupyan et al. (2007) would help determine how robust the label advantage is during subordinate category learning in adults. Understanding whether the effect is reliably present in adults would clarify what to expect when trying to track its developmental trajectory, and further characterize the relationship between words and category learning. Additionally, refining a procedure to investigate the label advantage during subordinate category learning in adults is an important first step before investigating the phenomenon in children. A pilot study conducted by the present authors attempted a conceptual replication of Lupyan et al. (2007) by comparing performance of participants with labels versus no labels on a category learning task inspired by Lupyan et al. (2007). Twenty-one participants completed the task, 11 in the label condition and 10 in the no-label condition. Based on the prior study, we expected to observe higher category learning performance for the label condition as compared to the no-label condition. Instead, we found that the no-label group trended higher in category learning performance compared to the label group.

This result left open the question of whether we failed to replicate the label benefit effect because of various task differences and the relatively small sample size, or whether the label benefit is itself not a robust effect. Therefore, we wanted to conduct a close replication that reduced the number of differences between the original and replication and matched the sample size of the original. We conducted the replication online rather than in

person. The results of this replication will help address the question of how robust the benefit of labels is during fine-grained object category learning for adults. If participants learn the categories faster in the auditory label and written label conditions compared to the location and no-label conditions, this would be evidence that the original effect survives differences in task implementation and participant population. Then, future work could further investigate the phenomenon to better understand the conditions under which there is a benefit of having a label versus not having a label during fine-grained category learning.

Replicated Study and Hypothesis

We aimed to replicate Experiment 2 from Lupyan et al. (2007), which reports that participants learn object categories faster when they are given labels compared to when they have no labels, or other nonlinguistic cues. Experiment 2 was chosen because it captures the full scope of the claims made in the original paper. Specifically, Experiment 2 compares labels to both having no label and having other types of nonlinguistic cues, while Experiment 1 only includes a label versus no-label comparison.

Methods

Participants

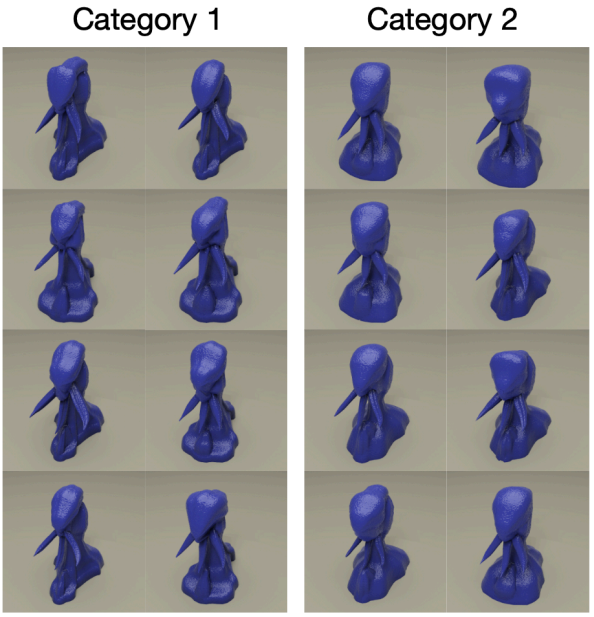
76 participants completed the study, ages 22–78 (M = 44, SD = 13). One participant reported an age of 7, which is not possible on Prolific as the platform requires participants to be at least 18 years old. Therefore, this participant's age was recoded to N/A during data processing. Nineteen participants were randomly assigned to each of the four conditions. Of those that reported their gender, 35 were female and 38 were male. Three participants did not report their gender. Participants were recruited online through Prolific (2026; https://www.prolific.com) and compensated $8/hr. Participants were required to be fluent in English, to ensure they would understand the instructions of the task. The sample was also limited to participants from the United States, since the original sample were people living in the U.S.

Materials and Procedure

To conduct this experiment, we recreated the original category learning task using PsychoPy (Peirce et al., 2019) and hosted it online using Pavlovia (https://pavlovia.org/). The 16 alien images used in the replication were the same as the original, as they are a subset of the publicly available Yu's Un-Facelike Objects (YUFO) stimulus set originally used in Gauthier et al. (2003; Figure 1).

Figure 1

Subset of YUFO Stimulus Set Used



Note. Courtesy of Michael J. Tarr, Carnegie Mellon University, http://www.tarrlab.org/

The category learning task was designed as close as possible to the original. In the task, participants were tasked with learning two groups of aliens. They were asked to imagine that they were an explorer on another planet, learning which group to approach and which to avoid. The task had four conditions: auditory label, written label, location (nonlinguistic cue), and no-label. For those in the two label conditions (auditory and written), the instructions included the statement, “Previous visitors to the planet have found it useful to name the two kinds of aliens, and you should pay careful attention to the labels” (Lupyan et al., 2007, p. 1079). For those in the location condition, the instructions stated, “One group of aliens live on one side of the planet, and the other lives on the other side … The alien will move up or down to signal where it lives.” The no-label condition did not have any unique instructions.

On each trial, one of the aliens was presented in the center of the screen, and then an astronaut character appeared on one side of the alien (left, right, above, or below) after 500 ms. In the instructions, participants were told that if they believed the alien should be approached, they should press the arrow key on their keyboard that would result in the astronaut moving toward the alien. Likewise, if they believed the alien should be avoided, the participant should press the arrow key on their keyboard that would result in the astronaut moving away from the alien. After the participant made their response, the astronaut moved in their chosen direction. Then, 200 ms after the astronaut stopped moving, the participant received auditory feedback of whether they were correct (bell sound) or incorrect (buzz sound).

After getting feedback, the trial differed depending on which condition the participant was assigned to. In the auditory label condition, participants heard the alien’s category label 300 ms after getting feedback. The labels used were the same as the original study: the nonsense words *leebish* and *grecious*. A text-to-speech converter was used to generate the auditory labels in a female voice. In the written label condition, a written label was used,

which appeared to the right of the alien 300 ms after feedback. In the location condition, the alien moved up or down to signal where it "lives" 300 ms after feedback. Lastly, in the no-label condition, no additional cue was given to participants after feedback. However, the alien remained on the screen for the same duration as the other conditions to equate exposure to the stimuli. The total time the stimuli remained on the screen after participants got feedback on each trial was 1,800 ms for the auditory label, written label, and no-label conditions and 2,000 ms for the location condition. Then, the trial ended and a fixation cross appeared to signal the next trial would be starting. The timing of stimulus presentation throughout the task was the same as the original study. The pairing of the labels or locations with the approach versus avoid categories, and with the set of images used, were counterbalanced across subjects.

Participants completed nine blocks of 16 trials of the category learning trials, which is the same length as the original study. Just as in the original study, verification trials were also included throughout the task. In the verification trials for the two label conditions, an alien appeared in the center of screen with the question: "Is this one leebish/grecious? Left arrow key = yes, right arrow key = no." In the location condition, the alien instead moved up or down and the question "Is this correct? Left arrow key = yes, right arrow key = no" was presented. No verification trials were given in the no-label condition, as their purpose was to measure how well participants were learning the association between the alien stimuli and cues (labels or locations).

Just as in the original study, participants did not get feedback on the verification trials. One verification trial was presented at the end of each block of 16 training trials. This resulted in nine verification trials throughout the task. This differs from the original study, which had 14 verification trials that were presented "after a random 10% of training trials" (Lupyan et al., 2007, p. 1080). This difference is due to limitations of the experimental presentation software PsychoPy; the setup of loops for trial blocks makes it difficult to insert a new trial type within a block without it being repeated every iteration of the loop. One other difference in the verification trials was that in the original location condition, participants could repeat the motion before making their response. In the replication, we decided to not include this option because after the alien moves, it remains in the location where it stopped moving, so it is clear to participants which direction the alien moved even if they missed the initial movement (because the alien always starts in the center of the screen).

A black background and white text were used in accordance with the original task. One aspect of the task that could not be equated with the original study is the size and visual angle subtended by the screen. In the original study, a 17-in. (43-cm) computer screen was used that subtended 8° of visual angle. However, since we ran our replication online, participants could be using a range of laptop or desktop screen sizes. One other small difference is that in the original study, participant responses were collected using a gamepad controller. In the replication, participants used their computer keyboards. Other minor differences include the buzz and bell sounds used for feedback, the recordings for the auditory labels, and the astronaut image used. Summaries of some key differences and replication closeness between the original and present study are provided in Table 1 and Table 2, respectively.

Analysis

The analysis plan is preregistered here: https://osf.io/rktc6/overview. All data analysis was conducted using R (Version 4.5.2; R Core Team, 2025), implemented in RStudio (Posit Team, 2025), and the lme4 (*v*2.0.1; Bates et al., 2015), lmerTest (*v*3.2.1; Kuznetsova et al., 2017), emmeans (*v*2.0.2; Lenth & Piaskowski, 2019), and effectsize packages (*v*1.0.2; Ben-Shachar et al., 2020). Data were cleaned and tidied such that trials were excluded if their response time is more than 3 minutes. We also checked for our preregistered exclusion criterion of pressing the same arrow key for more than 90% of trials.

The key analysis of interest is an ANOVA with condition and block as factors. This is used to test the hypothesis that those who received category labels (auditory and written label conditions) learned the categories faster than those who did not receive labels (no-label condition) or got nonlinguistic cues (location condition). A significant Block × Condition interaction and main effect of condition would suggest that there is a difference in category learning efficiency and accuracy across the conditions. Additionally, planned comparisons of the conditions were conducted to see whether the two label conditions differ from one another and whether the no-label and location conditions differ from one another. The original authors also conducted two more Condition × Block ANOVAs: 1) with the pooled data from the label conditions and pooled data from the no-label and location conditions, and 2) with just the written label and location conditions. We also conducted these analyses.

Other analyses used participants' performance on the verification trials to see if their verification accuracy correlates with training accuracy, and whether an ANOVA with Condition as a between-subjects factor is significant. If the ANOVA was significant, we planned to follow up with pairwise comparisons of the conditions. These verification trial analyses were only conducted for the auditory label, written label, and location conditions, as the no-label condition does not have verification trials.

Original and Replication Study Features

Table 1

Differences Between the Original and the Replication Study

Study Feature	Original Study	Replication Study	Justification and Possible Effect on Interpretability
Data collection time	2007	12/2025	Does not apply.
Administration	In the lab	Online	Facilitate participant recruitment and save time spent on data collection. May cause less reliable participant engagement.
Type of sample	University undergraduates	Prolific participants, education unknown	If the effect is robust, it should extend to online participants and those with more variable education backgrounds.
Exclusion criteria	Not reported	Pressing same key for more than 90% of trials	Want to ensure participants were engaging with the task.
Compensation	Course credit	Monetary	Necessary because sample was recruited on Prolific.

Replication Closeness

Table 2

Replication Closeness According to Features Proposed by Brandt et al. (2014)

Study Feature	Closeness (type: exact, close, different, or does not apply)	Description
Instructions	Close	Complete instructions not reported in original, though anything that was reported was used in replication.
Measures	Exact	Both using trial accuracy
Stimuli	Exact	YUFO stimulus set used for both
Procedure	Close	Procedure followed as close as possible to the original, implemented in PsychoPy
Location	Different	Lab (original) vs online (replication)
Remuneration	Different	Course credit (original) vs monetary (replication)
Participant populations	Different	Undergraduate students aged 18-24 (original) vs Prolific participants ages 22-78 (replication)
Participant exclusions / exclusion criteria	Different	Exclusion criteria not reported in the original study
Language	Exact	English used for both
Nationality of the participants	Exact	US residents used for both

Involvement of Original Authors

We reached out to Gary Lupyan, the first author of the original paper, before data collection and he sent us the stimuli used in the original paper. We also contacted him with an initial draft of the report, at which point he made analysis suggestions that we included in our Further Results section.

Sample Size Justification

Our target sample size was 76 participants, 19 in each condition. The original study had 75 participants, and we adjusted it to 76 to have the same number of participants in each group.

The original authors reported a partial eta-squared effect size of .07 for their key analysis of interest, the Condition × Block interaction. This translates to a partial Cohen's *f* of 0.27, indicating a medium effect size. Thus, a power analysis to detect an interaction with a medium effect size in a mixed ANOVA was conducted using G*Power (*v*3.1.9.6; Faul et al., 2009). The power analysis protocol can be found here: https://osf.io/xjfpu/files/5adup. The power analysis indicated that a sample of 20 participants was needed to achieve 80% power, 24 participants were needed to achieve 90% power, and 28 participants were needed to achieve 95% power. However, we decided to match the original study's sample size in order to replicate it as closely as possible. We also thought this would allow for the detection of smaller effects, as we anticipated may be the case since we were running the study online rather than in the lab.

Preregistration

The original study was not preregistered. The replication was preregistered on the Open Science Framework (OSF) prior to data collection, using the OSF Standard Pre-Data Collection Registration template (https://osf.io/rktc6).

Results

The raw and processed data are available online (https://doi.org/10.17605/OSF.IO/XJFPU). Instructions to reproduce results are also available (https://osf.io/xjfpu/files/jk3my). No participants met the exclusion criterion of pressing the same key for more than 90% of trials. No trials were excluded based on a response time of more than 3 minutes.

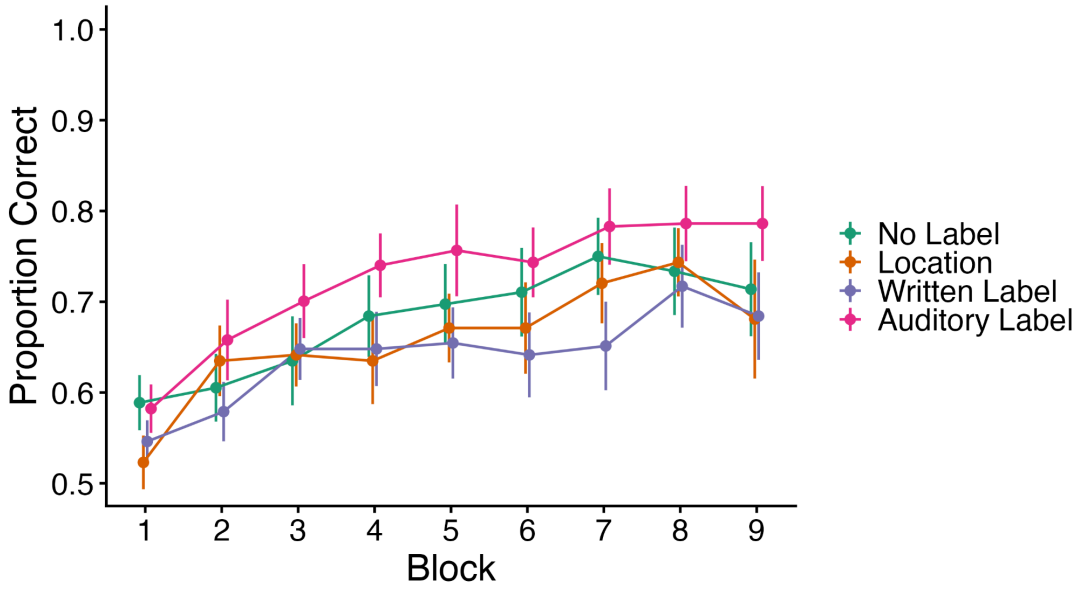
Statistical test results from the original paper, along with confidence intervals inferred from the available information, are shown in Table 3. The results from the present study are shown in Table 4. The key analysis of interest was a Condition x Block ANOVA, from which the original paper reported a significant Condition × Block interaction, $F(24, 568) = 1.66$, $p < .01$[1], $\eta^2_p = .07$, 90% CI [.00, .06], and a main effect of condition, $F(3, 71) = 3.18$, $p < .03$, $\eta^2_p = .08$, 90% CI [.01, .22]. In the replication data (Figure 2), there was not a significant Condition × Block interaction, $F(24, 648) = 0.24$, $p > .99$, $\eta^2_p = .01$[2], but there was a main effect of condition, $F(3, 648) = 6.77$, $p < .001$, $\eta^2_p = .03$, 90% CI [.01, .05]. These effects both lie within the inferred confidence intervals of the original effects. There was also a main effect of block such that participants improved their performance over time, $F(8, 648) = 7.34$, $p < .001$, $\eta^2_p = 0.08$, 90% CI [.04, .11].

[1] A statcheck report detected an error in which the recalculated *p*-value for this statistic was 0.026 (Nuijten et al., 2016). The statcheck report can be found here: https://pubpeer.com/publications/EB3FD6CF5EB7E2EA9C9A1E424CB18D

[2] The reported 90% confidence interval was degenerate at zero [0,0].

Figure 2

Category Learning Performance by Condition

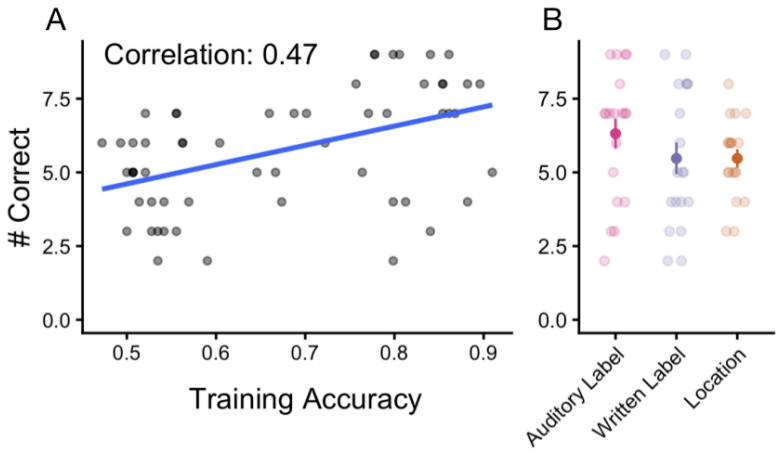


We conducted the same follow-up analyses as the original authors, as detailed in the Analysis section. First, we conducted planned comparisons of the conditions to test whether the two label conditions differ significantly from one another, and whether the no-label and location conditions differ significantly from one another. In the original paper, the authors found that neither pair differed significantly. Therefore, they followed up with a Condition × Block ANOVA with the pooled data from the label conditions and pooled data from the no-label and location conditions and found a significant interaction. In the replication, the auditory label and written label conditions were significantly different from each other, $t(648) = 4.247$, Šidák adjusted $p < .001$, Cohen's $d = 0.459$, 95% CI [0.25, 0.67], and the location and no-label conditions were not, $t(648) = -1.09$, Šidák adjusted $p = .474$, Cohen's $d = -0.118$, 95% CI [-0.33, 0.09]. In the ANOVA with the pooled data, there was not a significant Condition × Block interaction, $F(8, 666) = 0.205$, $p = .990$, $\eta^2_p = 0.002^2$, with the effect being smaller than the lower bound of the original effect's confidence interval. Lastly, the authors conducted a Condition × Block ANOVA of just the written label and location conditions, as they were the highest- and lowest-performing conditions in the original paper, and found a significant interaction. In the replication data, the interaction was not significant, $F(8, 324) = 0.276$, $p = .974$, $\eta^2_p = .007^2$. This effect was also smaller than the lower bound of the original effect's confidence interval.

Next, we conducted analyses on the verification trials from the auditory label, written label, and location conditions (Figure 3). In the original paper, they found that verification accuracy correlated with performance on the training trials, which was replicated, $r(55) = .47$, $p < .001$, 95% CI [.24, .65]. They also found that an ANOVA with Condition as a between-subjects factor was significant, which was not the case in our replication data, $F(2, 54) = 1.10$, $p = .339$, $\eta^2_p = .04$, 90% CI [.00, .13]. Both of these effects lie within the confidence intervals of the original effects.

Figure 3

Verification Trial Results



Note. A) Training accuracy correlates with verification performance. B) auditory label, written label, and location conditions do not vary significantly in verification performance.

Original key statistical result and interpretation ("claim")

The key statistical result was a mixed ANOVA with condition as a between-subjects factor and block as a within-subjects factor, from which the original paper reported a significant Condition × Block interaction, $F(24, 568) = 1.66$, $p < .01$, $\eta^2_p = .07$, and a main effect of condition, $F(3, 71) = 3.18$, $p < .03$, $\eta^2_p = .08$. This result is interpreted as showing that labels improve categorization, as participants learn the categories faster in the auditory and written label conditions than the no-label and location conditions (Lupyan et al., 2007, pp. 1080–1081).

Table 3

Lupyan et al. (2007) Results

Hypothesis / Item / Sample	*n*	Test statistic	*p*-value	Tails	Effect Size Value	Effect Size Type	CI	Type of analysis
Condition × Block interaction	75	$F(24, 568) = 1.66$	$p < .01$ (recalculated $p = .026$)	1	0.07	Partial η²	[0.00, 0.06][3]	Mixed ANOVA
Main effect of condition	75	$F(3, 71) = 3.18$	$p < .03$	1	0.08	Partial η²	[0.01, 0.22]	Mixed ANOVA
Auditory Label and Written Label comparison	36	$F(1, 34) = 0.26$	Not available/ Not possible to infer	Not available/ Not possible to infer	Not available/ Not possible to infer	Not available/ Not possible to infer	Not available/ Not possible to infer	Not available/ Not possible to infer
Location and No Label comparison	39	$F(1, 37) = 0.0$	Not available/ Not possible to infer	Not available/ Not possible to infer	Not available/ Not possible to infer	Not available/ Not possible to infer	Not available/ Not possible to infer	Not available/ Not possible to infer
Condition × Block with pooled data	75	$F(8, 584) = 2.99$	$p < .001$ (recalculated $p = .003$)	1	0.04	Partial η²	[0.01, 0.06]	Mixed ANOVA
Condition × Block with just written label and location	75	$F(8, 264) = 3.58$	$p < .001$	1	0.09	Partial η²	[0.03, 0.14]	Mixed ANOVA
Verification and category learning correlation	52	$r(50) = 0.34$	$p = .02$	2			[0.07, 0.56]	Correlation
Difference in verification performance across conditions	52	$F(2, 50) = 5.28$	$p < .01$	1	0.17	Partial η²	[0.03, 0.32]	One-way ANOVA

Note. All present values were reported in the paper except for the tails and confidence intervals, which were inferred based on the available information. CIs are 90% for *F*-tests and 95% for all other tests. There are also two recalculated *p*-values included from the statcheck report found here: https://pubpeer.com/publications/EB3FD6CF5EB7E2EA9C9A1E424CB18D (Nuijten et al., 2016).

[3] With large numerator df, the sample η^2_p is positively biased and can fall outside the noncentral-F CI; the interval was computed with effectsize::F_to_eta2.

Table 4

Replication Results

Hypothesis / Item / Sample	*n*	Test statistic	*p*-value	Tails	Effect Size Value	Effect Size Type	CI	Type of analysis	Within original CI?	Calc
Condition × Block interaction	76	$F(24, 648) = 0.235$	$p > .99$	1	0.009	Partial η²	[0.00, 0.00]	Mixed ANOVA	Yes	https://doi.org/10.17605/OSF.IO/XJFPU
Main effect of condition	76	$F(3, 648) = 6.774$	$p = .0002$	1	0.03	Partial η²	[0.01, 0.05]	Mixed ANOVA	Yes	https://doi.org/10.17605/OSF.IO/XJFPU
Auditory Label vs Written Label comparison	38	$t(648) = 4.247$	$p < .0001$	2	0.459	Cohen's *d*	[0.25, 0.67]	Emmeans contrast	N/A	https://doi.org/10.17605/OSF.IO/XJFPU
Location vs No Label comparison	38	$t(648) = -1.094$	$p = .474$	2	-0.118	Cohen's *d*	[-0.33, 0.09]	Emmeans contrast	N/A	https://doi.org/10.17605/OSF.IO/XJFPU
Condition × Block with pooled data	76	$F(8, 666) = 0.205$	$p = .990$	1	0.002	Partial η²	[0.00, 0.00]	Mixed ANOVA	No	https://doi.org/10.17605/OSF.IO/XJFPU
Condition × Block with just written label and location	38	$F(8, 324) = 0.276$	$p = .974$	1	0.007	Partial η²	[0.00, 0.00]	Mixed ANOVA	No	https://doi.org/10.17605/OSF.IO/XJFPU
Verification and category learning correlation	57	$r(55) = 0.468$	$p = .0002$	2			[0.24, 0.65]	Pearson's correlation	Yes	https://doi.org/10.17605/OSF.IO/XJFPU
Difference in verification performance across conditions	57	$F(2, 54) = 1.105$	$p = .339$	1	0.04	Partial η²	[0.00, 0.13]	One-way ANOVA	Yes	https://doi.org/10.17605/OSF.IO/XJFPU

Note. CIs are 90% for *F*-tests and 95% for all other tests. CIs of [0.00, 0.00] were degenerate at zero.

Further results

We conducted a couple of exploratory analyses to better understand the patterns in our replication data. First, since our planned condition contrasts did not pattern in the same way as the original study, we did pairwise contrasts with a Tukey correction on the category training data to see which conditions differed from each other. These contrasts revealed that the auditory label condition performed better than the written label, $t(648) = 4.247$, Cohen's $d = 0.459$, 95% CI [0.25, 0.67], location, $t(648) = 3.41$, Cohen's $d = 0.369$, 95% CI [0.16, 0.58], and no-label conditions, $t(648) = 2.315$, Cohen's $d = 0.250$, 95% CI [0.04, 0.46][4]. They also revealed that the no-label condition performed better than the written label, $t(648) = 1.932$, Cohen's $d = 0.209$, 95% CI [0.00, 0.42], and location conditions, $t(648) = 1.094$, Cohen's $d = 0.118$, 95% CI [-0.09, 0.33]. This is a different pattern than was shown in the original study, where the auditory label and written label conditions performed better than the location and no-label conditions.

The original authors of the paper also gave the helpful suggestion to account for interdependence between multiple observations from the same participant by incorporating a random effect of participant in our analysis. Therefore, we conducted a mixed-effects logistic regression on trial accuracy with predictors of block and condition and random intercepts of participant and alien stimulus (the image used for each trial). In our model, block was a weak predictor of trial accuracy, but was estimated most precisely ($\beta = 0.100$, SE = 0.008, z = 11.797, 95% CI [0.08, 0.12]). Meanwhile, the condition predictors had stronger estimates but were not precise (Table 5).

[4] *p*-values are not reported for any of the analyses in this section since they are exploratory analyses.

Table 5

Fixed Effect and Variance Estimates for LME Model of Trial Accuracy

Fixed effect	**Estimate**	***SE***	***z*-value**	**95% CI**
Intercept	0.606	0.199	3.054	[0.21, 1.00]
Condition (No Label)	-0.202	0.251	-0.804	[-0.70, 0.30]
Condition (Location)	-0.342	0.250	-1.366	[-0.84, 0.15]
Condition (Written Label)	-0.430	0.250	-1.720	[-0.93, 0.07]
Block	0.100	0.008	11.797	[0.08, 0.12]
Random intercept	s^2			
Participant	0.556			
Alien stimulus	0.102			

Note. Auditory label is used as the reference level for condition.

When conducting pairwise contrasts between the conditions with this model, we found that the confidence intervals for all contrasts included 0 (Table 6).

Table 6

Pairwise Condition Comparisons of Estimated Marginal Means

Contrast	**Estimate**	***SE***	***z*-ratio**	**95% CI**
Auditory Label – No Label	0.202	0.251	0.804	[-0.29, 0.69]
Auditory Label – Location	0.342	0.250	1.366	[-0.15, 0.83]
Auditory Label – Written Label	0.430	0.250	1.720	[-0.06, 0.92]
No Label – Location	0.140	0.250	0.561	[-0.35, 0.63]
No Label – Written Label	0.229	0.250	0.914	[-0.26, 0.72]
Location – Written Label	0.088	0.250	0.354	[-0.40, 0.58]

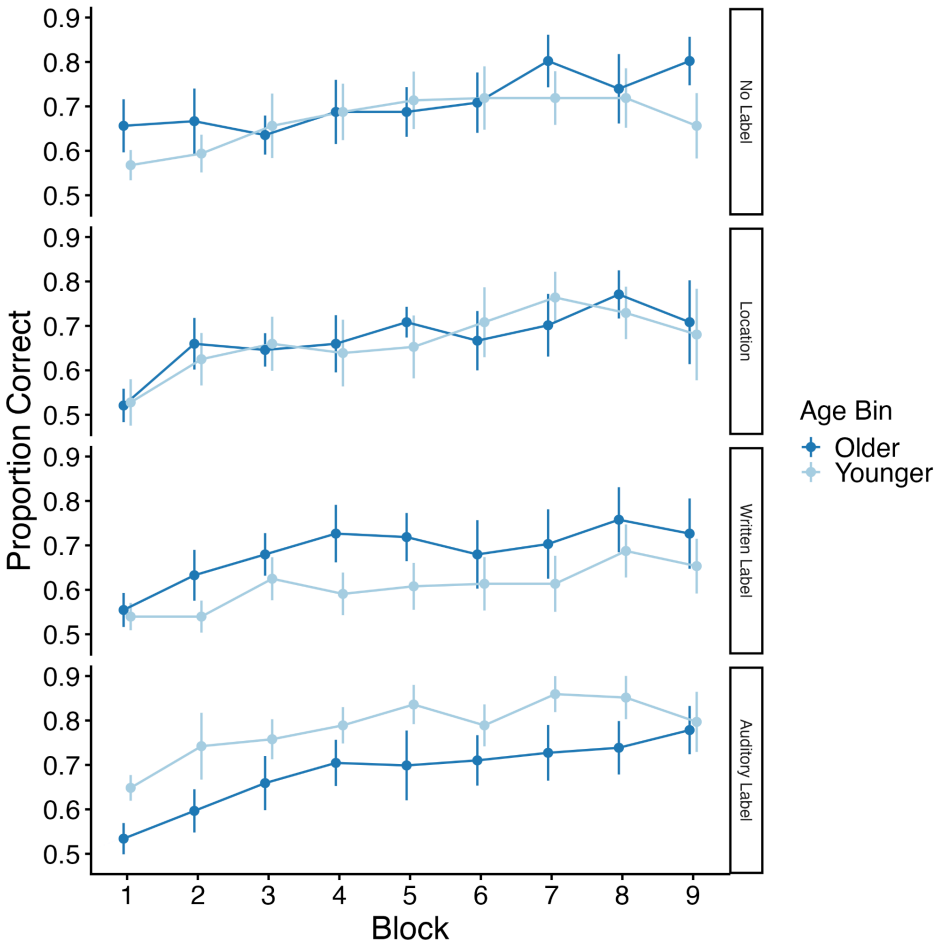
Additionally, the original authors suggested treating block as numeric rather than as a factor in the Condition × Block ANOVA. We redid the ANOVA with this adjustment and found that the results replicated our previous analysis such that there was a negligible Condition × Block interaction, $F(3, 676) = 0.254$, $\eta^2_p = .001$, 90% CI [.00, .004], but there was a small effect of condition, $F(3, 676) = 6.914$, $\eta^2_p = .03$, 90% CI [.01, .05], and stronger effect of block, $F(1, 676) = 50.345$, $\eta^2_p = .07$, 90% CI [.04, .10].

We also conducted an exploratory analysis to see if age might have affected the results, as the original study had participants aged 18 to 24 and the replication participants ranged from 22 to 78. We conducted another ANOVA with age added as a covariate, along with an error term for participants. When age is added as a covariate in the ANOVA, the Condition × Block interaction is small, $F(24, 528) = 0.512$, $\eta^2_p = 0.02$[5], and the auditory label versus no-label contrast remains weak, $t(66) = 0.795$, Cohen's $d = 0.335$, 95% CI [-0.51, 1.18]. However, there is a moderate Condition × Age interaction, $F(3, 66) = 1.60$, $\eta^2_p = 0.07$, 90% CI [0.00, 0.16]. To probe this interaction, we split the data by the median age (42) into older and younger age bins, shown in Figure 4. The median split shows that younger participants performed better than older participants in the auditory label condition, as confirmed by an estimated marginal means contrast, $t(66) = 1.495$, Cohen's $d = 0.85$, 95% CI [-0.29, 1.99]. Additionally, younger participants seem to be performing worse than older participants in the written label condition, $t(66) = -1.148$, Cohen's $d = -0.65$ 95% CI [-1.79, 0.49]. Both contrasts were performed with Tukey correction.

[5] The reported 90% confidence interval was degenerate at zero [0,0].

Figure 4

Category Learning Performance by Condition, Split by Median Age

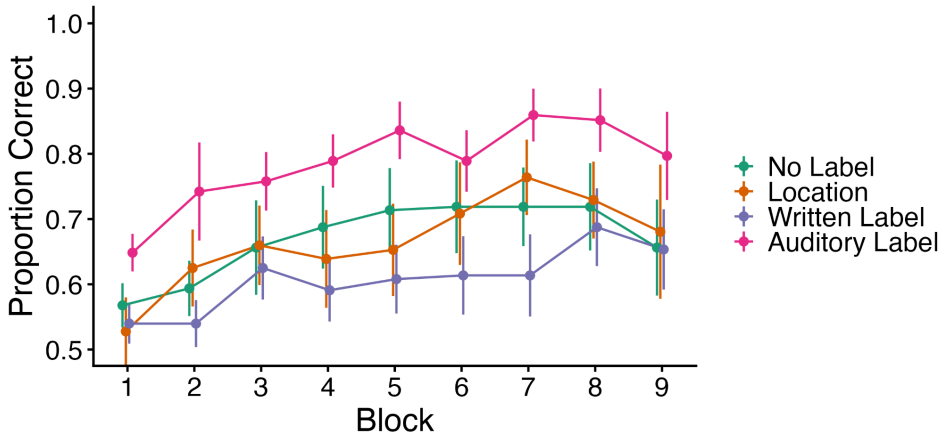


As a further analysis, we wanted to see whether in the younger group, performance in the auditory label condition was greater than the no-label condition. Category learning performance for the younger group (n = 40) is shown in Figure 5. When the contrasts are conducted with Tukey correction on just the younger group, the auditory label versus no-label contrast is stronger, $t(36) = 1.614$, Cohen's $d = 1.00$, 95% CI [-0.28, 2.29]. The auditory label group also performs better than the written label group, $t(36) = 2.439$, Cohen's $d = 1.54$, 95% CI [0.21, 2.88], and location group, $t(36) = 1.582$, Cohen's $d = 1.05$, 95% CI [-0.32, 2.41]. Meanwhile, the Condition × Block interaction is slightly stronger, $F(24, 288) = 0.575$, $\eta^2_p = .05$[6].

[6] The reported 90% confidence interval was degenerate at zero [0,0].

Figure 5

Category Learning Performance by Condition for the Younger Participants



Note. Only participants aged 42 years or younger are included (N = 40). Number of participants per condition: auditory label, n = 8; written label, n = 11; location, n = 9; no-label, n = 12.

Dataset and Materials link

https://doi.org/10.17605/OSF.IO/XJFPU

Object/file name

all_participants.csv

Limits to sharing

All materials are openly available at https://osf.io/xjfpu/overview. These include the analysis scripts, data, and preregistration.

FAIR data/Codebook

The processed and raw datasets including a codebook and all code used for the analyses are available online.

- compiled processed data: https://osf.io/xjfpu/files/w5v92
- raw and processed datasets: https://osf.io/xjfpu/files/osfstorage
- codebook: https://osf.io/xjfpu/files/sr5gv
- code: https://osf.io/xjfpu/files/xqv5j

Deviations from the Preregistration

In the preregistration, the ANOVAs are described as "mixed ANOVAs", but the implementation does not include an error term for participants. Therefore, in our Results section we stuck with the preregistered analysis code, and in our Further Results section we addressed this issue. Šidák correction was used instead of Tukey correction in instances where two contrasts were performed at once. Additionally, effect size and confidence interval calculations were added to the preregistered analysis code. The preregistered analyses use data labeled as "Pilot B" which was a set of 7 pilot participants used to verify that the task and analysis code were working correctly.

Evaluation

Even though the replication data show a partial replication of the primary analysis of interest (the main effect of condition), the lack of a Condition × Block interaction fails to support the main claim made in the original paper: adults learn categories quicker if they have labels versus when they do not. While the auditory label group performed better than the no-label group, the written label group had lower performance than the auditory label and location groups. This suggests that having a label is not always beneficial in this category learning task. Lastly, while performance on the verification trials was numerically higher for the auditory label group compared to the written label and location groups, this was not a statistically significant difference. The benefit of having a label may be sensitive to task implementation and participant population differences.

Feedback from the original study's authors

Comment provided by Gary Lupyan:

We commend the authors for undertaking this replication of our 2007 study and thank the authors for providing this opportunity to comment. Our original study has been the subject of numerous extensions, as we review below. However, to our knowledge there have not been prior exact replications. The closest is Lupyan and Casasanto (2015) who successfully replicated the main effect of labels (though with slightly different stimuli), and Brojde et al. (2011), who showed that labels significantly interfere with category learning when the category contrast is not based on shape (see also Tolins & Colunga, 2013; for a partial replication, see Mertens et al., 2025). When it is based on shape (as in the original study), labels were found to lead to numerically, but not significantly better learning. Other work aiming to investigate effects of labels on learning different types of categories, found labels to help learning of probabilistic categories (Fotiadis & Protopapas, 2014) as well as learning hard-to-name shape-based categories in a within-subject design (Fotiadis & Protopapas, 2023). The facilitating effect of labels has also been extended to other modalities, for example, in learning olfactory (Vanek et al., 2021) and tactile (Miller et al., 2018) categories. An attempt to find effects of labels on emotion-categories was unsuccessful (Hoemann et al., 2022). Consistent with our proposal that labels affect category learning by highlighting category-diagnostic features (e.g., Lupyan, 2012), a number of findings have confirmed that labels selectively increase looking toward diagnostic features in both children and adults (Althaus & Mareschal, 2014; Barnhart et al., 2018).

One factor that may help explain the present non-replication is the lower overall categorization accuracy observed here compared to the original experiments. It is possible that online testing may have contributed to a lower general engagement with the task, contributing to a larger proportion of participants (in both conditions) who struggled with the basic task. It is also worth noting that the sample sizes in the original study were small by current standards; we cannot rule out that the original effect size was therefore an overestimate. We hope this replication attempt, together with the broader literature, will help the field develop a clearer picture of the conditions under which labels facilitate category learning and home in on the factors that modulate this effect.

Limitations and Recommendations for Future Replications

The major difference between this replication study and the original study is that the replication was done online and the original was in-person. It is possible that this influenced the findings, especially given that the participants' overall performance was lower than the original study.

Some work has suggested that switching to online for a replication study when the original was done in person, compared to using the same modality for both the original and replication, results in lower replicability (Boyce et al., 2023). We speculate that this may be for a couple of reasons: 1) conducting a study online introduces more flexibility for participants in the amount of attention paid to the task, meaning that the participants may not be reliably engaged throughout the task, or 2) it could be due to the differences introduced to the task when converting it to an online task. In our case, any validity concerns would most likely coincide with the first possibility. This is because the differences in the tasks across studies are minor, as participants completed the task at a computer in the original study as well (just in-person rather than online). Technical differences include uncontrolled screen size, visual angle, and audio delivery (e.g., volume, headphone vs speakers), as well as potential noise introduced to the stimulus presentation timing as a result of using browser-based timing. These differences can be reduced in future replications through additional checks such as standardized audio checks and timing validation procedures.

Ideally, the effect would be robust enough to survive potential differences in level of participant engagement – however, it is difficult to know whether the non-replication of the primary analysis of interest, the Condition x Block interaction showing a benefit of having labels, was due to this difference. The present study does not include attention checks, which would be helpful in determining how the level of participant engagement affected the results (Cuskley & Sulik, 2024). Additionally, in the present study there are wide confidence intervals for some effects, especially in the exploratory analyses using the mixed effects model. This indicates that some estimated effects are not precise. It is also worth noting, as the original author does above, that the original study may have overestimated the size of the effect. In addition, our inferred confidence intervals from the original study include 0 for both the Condition x Block interaction and main effect of condition. In light of these considerations, we suggest that future attempts to replicate this effect should aim to detect smaller effect sizes and include more participants. Additionally, we recommend including attention checks to ensure sufficient data quality.

Another difference between this study and the original study is the age range of participants. Our study had participants with ages ranging from 22–78, while the original study had college student participants ranging from ages 18–24. The greater age range in our replication may have introduced more variation that resulted in our smaller effect size. Our exploratory analyses showed that age did not substantially influence our main analysis of interest, the Condition x Block interaction. However, we did see that for the younger half of participants, the difference in performance on the auditory label and no-label conditions was much greater than in our full sample. This could perhaps be due to a higher prevalence of hearing loss in older adults, thus hindering their performance in the auditory label condition (Agrawal et al., 2008).

With our participant sample, we are not able to subsample to the original study age range, as only 3 participants fall in the range of 18–24. Additionally, since the original study participants were all college students, this could be another relevant difference that would remain even if age was equated. Therefore, while it seems that age in general did not influence our ability to detect the main analysis of interest, it is not possible to know whether the effect would have been replicated if the participant samples were more closely matched on age and other demographics such as education status. Additionally, while both samples were English speakers, another potential difference between samples could be first language and language(s) spoken, which are relevant to language processing. We do not have this information from our participants, and the original study did not report this information.

As detailed above, there are a number of demographic differences between participant samples in the original study and our replication. Future work could help gain further clarity as to what extent these differences influence the label benefit in category learning, and inform how researchers can investigate this phenomenon in children.

Brief Evaluation

□ success
X failure
□ mixed
□ inconclusive

Social Impact & Responsibility

Fine-grained category knowledge is often acquired with the help of linguistic labels in adults, and is a form of expertise across many domains. While this project aimed to replicate the finding that linguistic labels boost fine-grained category learning for adults, this was intended as a first step before extending the experimental paradigm to children. Examining the development of this ability in children helps us understand how surprisingly sophisticated world knowledge can develop at a young age. Additionally, the broader goal of understanding the mechanisms supporting category learning in children can be translated to the educational context in the form of pedagogical applications.

Declarations

Contribution Statement

Caroline Kaicher: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing
Ellen M. Markman: Conceptualization, Writing – review & editing
Laura Gwilliams: Conceptualization, Writing – review & editing

Ethical Statement

This research was approved by the IRB of Stanford University under protocol #IRB-23274. The procedures used in this study adhere to the tenets of the Declaration of Helsinki.

Transparency Statement

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

AI Use Statement

Chat GPT 5.6 Terra was used to generate suggestions of how to implement confidence interval calculations for the model terms in the mixed effects logistic regression model in R. Generative AI was not used for generating any text or for review or editing of the writing in this paper.

Acknowledgements

We thank Gary Lupyan for his willingness to clarify the original stimuli used, helpful discussion and suggestions of additional analyses, and for providing a comment for section 4.1. We thank Jay McClelland for helpful discussions about the paper, Mike Frank for facilitating this replication as a class assignment, and Irmak Ergin for her helpful feedback along the way as a teaching assistant. We thank the reviewers for their thoughtful and constructive comments.

Conflicts of Interest

The authors declare no conflict of interest associated with the publication of this manuscript.

Funding Statement

L.G. was supported by BRAIN Foundation A-0741551370, Whitehall Foundation 2024-08-043, and the Esther A. & Joseph Klingenstein Fund.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2146755. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Agrawal, Y., Platz, E. A., & Niparko, J. K. (2008). Prevalence of hearing loss and differences by demographic characteristics among US adults: Data from the National Health and Nutrition Examination Survey, 1999–2004. Archives of Internal Medicine, 168(14), 1522–1530. https://doi.org/10.1001/archinte.168.14.1522

Althaus, N., & Mareschal, D. (2014). Labels direct infants' attention to commonalities during novel category learning. *PLoS ONE*, *9*(7), Article e99670. https://doi.org/10.1371/journal.pone.0099670

Barnhart, W., Rivera, S., & Robinson, C. (2018). Effects of linguistic labels on visual attention in children and young adults. *Frontiers in Psychology*, *9*, Article 358. https://doi.org/10.3389/fpsyg.2018.00358

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), Article 2815. https://doi.org/10.21105/joss.02815

Boutonnet, B., Dering, B., Viñas-Guasch, N., & Thierry, G. (2013). Seeing objects through the language glass. *Journal of Cognitive Neuroscience*, *25*(10), 1702–1710. https://doi.org/10.1162/jocn_a_00415

Boyce, V., Mathur, M., & Frank, M. C. (2023). Eleven years of student replication projects provide evidence on the correlates of replicability in psychology. *Royal Society Open Science*, *10*(11), Article 231240. https://doi.org/10.1098/rsos.231240

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Brojde, C. L., Porter, C., & Colunga, E. (2011). Words can slow down category learning. *Psychonomic Bulletin & Review*, *18*(4), 798–804. https://doi.org/10.3758/s13423-011-0103-z

Cuskley, C., & Sulik, J. (2024). The burden for high-quality online data collection lies with researchers, not recruitment platforms. *Perspectives on Psychological Science*, *19*(6), 891–899. https://doi.org/10.1177/17456916241242734

Deng, W., & Sloutsky, V. M. (2012). Carrot eaters or moving heads: Inductive inference is better supported by salient features than by category labels. *Psychological Science*, *23*(2), 178–186. https://doi.org/10.1177/0956797611429133

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fotiadis, F. A., & Protopapas, A. (2014). The effect of newly trained verbal and nonverbal labels for the cues in probabilistic category learning. *Memory & Cognition*, *42*(1), 112–125. https://doi.org/10.3758/s13421-013-0350-5

Fotiadis, F. A., & Protopapas, A. (2023). Immediate and sustained effects of verbal labels for newly-learned categories. *Quarterly Journal of Experimental Psychology*, *76*(8), 1773–1789. https://doi.org/10.1177/17470218221126659

Gauthier, I., James, T. W., Curby, K. M., & Tarr, M. J. (2003). The influence of conceptual knowledge on visual discrimination. *Cognitive Neuropsychology*, *20*(3–6), 507–523. https://doi.org/10.1080/02643290244000275

Hoemann, K., Gendron, M., & Barrett, L. F. (2022). Assessing the power of words to facilitate emotion category learning. *Affective Science*, *3*(1), 69–80. https://doi.org/10.1007/s42761-021-00084-4

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lenth, R., Piaskowski, J. (2026). emmeans: Estimated Marginal Means, aka Least-Squares Means. *R package version 2.0.2*.

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, *3*, Article 54. https://doi.org/10.3389/fpsyg.2012.00054

Lupyan, G., & Casasanto, D. (2015). Meaningless words promote meaningful categorization. *Language and Cognition*, *7*(2), 167–193. https://doi.org/10.1017/langcog.2014.21

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083. https://doi.org/10.1111/j.1467-9280.2007.02028.x

Mertens, A. J., Colunga, E., & Kim, A. E. (2025). Labels facilitate categorical perception effects during novel category learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *47*. https://escholarship.org/uc/item/5m4272cf

Mervis, C. B. (1987). Child-basic object categories and early lexical development. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 201–233). Cambridge University Press.

Mervis, C. B., & Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development*, *53*(1), 258–266. https://doi.org/10.2307/1129660

Mervis, C. B., Johnson, K. E., & Mervis, C. A. (1994). Acquisition of subordinate categories by 3-year-olds: The roles of attribute salience, linguistic input, and child characteristics. *Cognitive Development*, *9*(2), 211–234. https://doi.org/10.1016/0885-2014(94)90004-3

Miller, T. M., Schmidt, T. T., Blankenburg, F., & Pulvermüller, F. (2018). Verbal labels facilitate tactile perception. *Cognition*, *171*, 172–179. https://doi.org/10.1016/j.cognition.2017.10.010

Murphy, G. L. (2010). What are categories and concepts? In D. Mareschal, P. C. Quinn, & S. E. G. Lea (Eds.), *The making of human concepts* (pp. 11–28). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199549221.003.02

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Posit team. (2025). *RStudio: Integrated development environment for R* (Version 2025.09.2.418) [Computer software]. Posit Software, PBC. https://posit.co

Prolific. (2026). *Prolific* [Online participant recruitment platform]. https://www.prolific.com

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X

Tanaka, J. W., Curran, T., & Sheinberg, D. L. (2005). The training and transfer of real-world perceptual expertise. *Psychological Science*, *16*(2), 145–151. https://doi.org/10.1111/j.0956-7976.2005.00795.x

Tolins, J., & Colunga, E. (2013). Exploring the role of verbal category labels in flexible cognition. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35. https://escholarship.org/uc/item/1pz3k3n9

Vanek, N., Sóskuthy, M., & Majid, A. (2021). Consistent verbal labels promote odor category learning. *Cognition*, *206*, Article 104485. https://doi.org/10.1016/j.cognition.2020.104485

Waxman, S. R. (1990). Linguistic biases and the establishment of conceptual hierarchies: Evidence from preschool children. *Cognitive Development*, *5*(2), 123–150. https://doi.org/10.1016/0885-2014(90)90023-M

Waxman, S. R., Shipley, E. F., & Shepperson, B. (1991). Establishing new subcategories: The role of category labels and existing knowledge. *Child Development*, *62*(1), 127–138. https://doi.org/10.1111/j.1467-8624.1991.tb01519.x

Open Science Badges

This work has received the following badges:



Preregistered: https://doi.org/10.17605/OSF.IO/RKTC6



Open Code, Open Materials, and Open Data: https://doi.org/10.17605/OSF.IO/XJFPU



Numerical Reproduction (Human- and AI-checked):
https://doi.org/10.5281/zenodo.24292480



Open Review Report: https://doi.org/10.5281/zenodo.24292040

Submission Data

Submission date: 2 May 2026
Publication date: 25 November 2026
Handling editor: Lydia Riedl

LICENSE | REPLICATION RESEARCH R2 | 2026
doi.org/10.17879/replicationresearch-2026-9615

Replication Research (R2, ISSN: 3052-5977) is part of the Framework for Open and Reproducible Research Training (FORRT; forrt.org) and the Münster Center for Open Science (MüCOS; uni-muenster.de/MueCOS)

MüCOS | FORRT | R2