UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

de Leeuw, J. R., Andrews, J., Barney, L., Bigler, M., Bruna, P. J., Chen, Y., Cherry, R., Dowie, D. R., Forbes, E., Haffey, B., Hu, X., Jaklitsch, M., Leopold, N., Lewis, C., MacDonald, D., McShaffrey, C., Monroy-Spangenberg, E., Nakayama, K., Olstad, W., ... Zhang, L. (2021). Words May Jump-Start Meaning More Than Vision: A Non-Replication of Early ERP Effects in Boutonnet and Lupyan (2015). *Collabra: Psychology*, *7*(1). https://doi.org/10.1525/collabra.29763

Cognitive Psychology

Words May Jump-Start Meaning More Than Vision: A Non-Replication of Early ERP Effects in Boutonnet and Lupyan (2015)

Joshua R. de Leeuw[1] [a], Jan Andrews[1], Lori Barney[1], Margaret Bigler[1], Polyphony J. Bruna[1], Yijing Chen[1], Ryan Cherry[1], Davon R. Dowie[1], Eden Forbes[1], Ben Haffey[1], Xinyue Hu[1], Michael Jaklitsch[1], Nicole Leopold[1], Caitlin Lewis[1], Dylan MacDonald[1], Connor McShaffrey[1], Erik Monroy-Spangenberg[1], Karen Nakayama[1], Wesley Olstad[1], Rebecca Peng[1], Griffin Scott-Rifer[1], Allison Wan[1], Logan Willans[1], Lingxiu Zhang[1]

[1] Cognitive Science, Vassar College

Keywords: replication, perception, language, top-down effects, erp, eeg, p100, p200, n400

https://doi.org/10.1525/collabra.29763

Collabra: Psychology

Vol. 7, Issue 1, 2021

We report a replication of Boutonnet and Lupyan's (2015) study of the effects of linguistic labelling on perceptual performance. In addition to a response time advantage of linguistic labels over non-linguistic auditory cues in judging visual objects, Boutonnet and Lupyan found that the two types of cues produced different patterns in the early perceptual ERP components P1 and P2 but not the later, semantics-relevant N4. This study thus adds an important piece of evidence supporting the claim of genuine top-down effects on perception. Given the controversy over this claim and the need for replication of key findings, we attempted to replicate Boutonnet and Lupyan (2015). We replicated their behavioral findings that response times to indicate whether an auditory cue matches a visual image of an object were faster for match than mismatch trials and faster for linguistic than non-linguistic cues. We did not replicate the main ERP effects supporting a positive effect of linguistic labels on the early perceptual ERP components P1 and P2, though we did find a congruence by cue type interaction effect on those components. Unlike Boutonnet and Lupyan, we found a main effect of cue type on the N4 in which non-linguistic cues produced more negative amplitudes. Exploratory analyses of the unpredicted N4 effect suggest that the response time advantage of linguistic labels occurred during semantic rather than early visual processing. This experiment was pre-registered at https://osf.io/cq8g4/ and conducted as part of an undergraduate cognitive science research methods class at Vassar College.

Introduction

Are there genuine top-down effects of language on perception? This is a longstanding issue in cognitive science and its component disciplines, most familiar perhaps in the form of the so-called Whorf hypothesis of linguistic relativity (Whorf, 1956) which states that lexical and grammatical differences between languages produce differences in non-linguistic processes in their speakers. The top-down claim has been controversial from its inception right up to the present. For example, Lupyan et al. (2020) review a large body of empirical evidence suggesting effects of language on visual perception, yet the interpretation of such evidence is far from clear. In an important critique of alleged top-down effects of cognition on perception, Firestone & Scholl (2016) argued that thus far all such effects are susceptible to plausible alternative explanations, and thus don't demonstrate compelling empirical support for top-down effects. They present six pitfalls that undermine the validity of published research demonstrating top-down effects, one of which is that the effects could just as well be on higher-level judgement rather than perception.

One kind of top-down effect of language on perception that has seen recent support in the literature is the label advantage for object recognition. Verbal labels (e.g., "dog") produce faster recognition of visually presented objects compared to non-verbal sounds (e.g., a dog's bark) (Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012). However, like most claims of top-down effects on visual perception, there is disagreement about the mechanism behind the behavior. For example, the label advantage could be due to top-down effects on visual processing as a result of expectations (Lupyan & Clark, 2015) or it could be due to language causing enhanced recognition memory rather than perception (Firestone & Scholl, 2016) in which case language is not affecting a lower-level process but rather, another aspect of higher-level cognition.

By using electroencephalography (EEG) to measure the

a jdeleeuw@vassar.edu

tiny voltage changes on the scalp that reflect brain activity, researchers can investigate the timing of different cognitive processes at the level of milliseconds. In particular, event-related potentials (ERPs) are typical brain wave patterns that occur with characteristic timing in response to specific time-locked events such as the presentation of a certain type of stimulus. For example, the "P1" is the first positive-going ERP component that occurs when a visual stimulus is shown, occurring approximately 100 milliseconds after its presentation. ERPs provide one possible way to address the distinction between perception and judgment (or other "back-end" processes) because of this precise timing information: if earlier components like the P1, N1, and P2 are affected, this would provide evidence beyond potentially ambiguous behavioral data that the effect is at least partly perceptual. The P1 component, for example, has been shown to be modulated by early visual processes involved in object recognition (e.g., Herrmann & Knight, 2001; Luo et al., 2013; Tanaka, 2018). In addition, "demand characteristics," the worry that participants are adjusting their responses in order to produce the "right" behavior for the experimenter, would no longer be a plausible alternative explanation of apparent top-down effects because of the covert nature of EEG measures and the speed of the response. On the other hand, if only later components like the N4 are involved, then the effect is more likely to be semantic rather than perceptual (e.g., Kutas & Federmeier, 2011), though the nature of the processing that N4 involves is not fully understood (e.g., Lau et al., 2008) and there is evidence that multiple processes take place during the N400 time window (Nieuwland et al., 2019).

Boutonnet & Lupyan (2015) used EEG to examine ERP correlates of the label advantage effect (see Figure 1 for their experimental procedure). They replicated the response time advantage of verbal labels versus nonverbal sounds when judging whether an auditory cue matches a visually presented object and offered ERP evidence to support the claim that this label advantage operates during an early, perceptual stage and not a post-perceptual semantic stage. They found that the P1 component had higher amplitudes and earlier latencies when visual stimuli were cued by linguistic labels rather than specific sounds. In addition, the P1 peak latency was sensitive to cue-target congruence only in the label condition. They report similar effects for the P2. However, the N4 ERP showed no distinction between linguistic and non-verbal auditory cues and only an overall effect of congruence, as would be expected given prior findings on semantic congruity and the N4 (e.g., Kutas & Federmeier, 2011). The lack of differences in N4 makes sense assuming that labels and non-verbal cues both activate the same high-level semantic representations once they have been processed. Earlier P1 latencies also predicted shorter RTs, suggesting that the low-level visual processes indexed by the P1 were related to the behavioral responses. Overall, their data suggest that the enhancement in visual processing due to verbal labels occurred only in the early perceptual stage, providing key support for top-down effects on perception.

We think that even studies that avoid Firestone and Scholl's (2016) six pitfalls and find evidence of top-down effects must also meet a seventh criterion: the demonstration of reliability, e.g., through direct replication. Replication is especially needed when the initial sample size is small and the analysis provides numerous researcher degrees of freedom, problems that are especially prevalent in the EEG literature (Clayson et al., 2019; Luck & Gaspelin, 2017). Boutonnet & Lupyan (2015) had a sample size of 14, and while there were some reported efforts to avoid data-driven analysis in the paper (e.g., in the selection of time windows for ERP analysis), the analysis plan was not pre-registered and it is unknown to what degree the analysis is sensitive to the specific modeling choices that were made. By carrying out a direct replication of Boutonnet & Lupyan (2015), our goal is to clarify the role that their findings should play in this ongoing controversy concerning top-down effects of higher-level knowledge, in this case, specifically linguistic knowledge, on visual object recognition.

Methods

Stimuli and experiment scripts are available on the Open Science Framework at https://osf.io/cq8g4/. A pre-registration for this study is available at https://osf.io/gkq7b. Due to recruitment difficulties, we amended our pre-registered data collection plan part way through data collection, extending our window for data collection by one week. The amendment is registered at https://osf.io/5a8bz.

Participants

Thirty-five Vassar College students participated in the study. Participants were native English speakers aged 18–23. Our pre-registered target minimum of 35 participants was 2.5x larger than the original sample (N = 14). This target was chosen based on Simonsohn's (2015) heuristic: Studies with 2.5x larger samples have about 80% power to detect an effect size that the original study had 33% power to detect. Participants were compensated with candy. This study was approved by the Vassar College Institutional Review Board and all participants provided informed consent.

Stimuli

We used the original stimuli from Boutonnet & Lupyan (2015), which were graciously provided by Lupyan. The stimuli consisted of 10 categories: cat, car, dog, frog, gun, rooster, train, cow, whistle, and motorcycle. For each category, there were five images (three photographs, one color drawing, and one cartoon image) and two audio cues (a non-verbal sound and a word label). Pictures and non-verbal sounds were easily identifiable, validated through norming studies conducted by Lupyan & Thompson-Schill (2012), and the labels and non-verbal sounds were matched on "imagery concordance" (Rossion & Pourtois, 2004).

We used version 6.1.0 of the jsPsych library (de Leeuw, 2015) for stimulus presentation, with a custom addition to allow for recording timing events in the EEG data stream via TCP/IP communication with the NetStation recording software. We conducted a timing test prior to launching the study to determine the relative offset between recorded events and actual stimulus display. We found that the average offset was 60ms (*SD* = 2.8 ms) between the event and the stimulus display. 84.6% of events were within 2 ms of

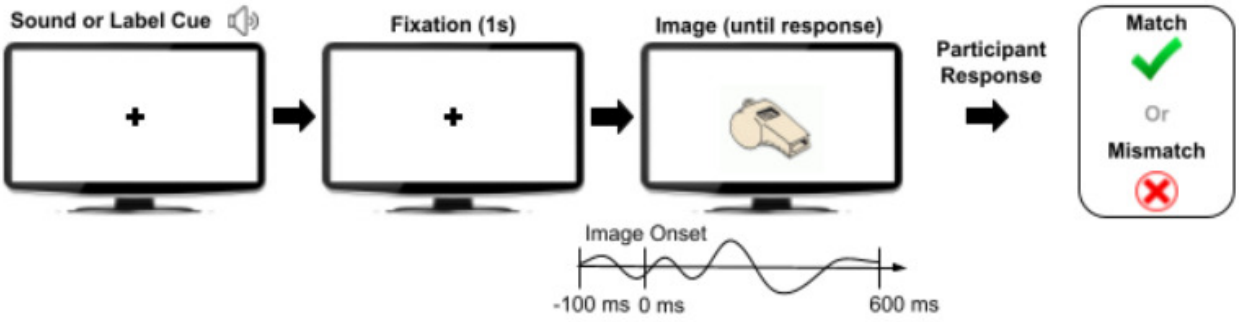

Figure 1. Trial Procedure.

At the start of each trial, an audio file (sound or label) played while a fixation cross appeared on the screen. After the completion of the audio, the fixation remained on the screen for one second. The image appeared immediately after and participants indicated whether the image matched audio or not by pressing one of two keys. ERPs were synchronized with the onset of the image.

the mean. We corrected for this offset in the segmentation phase of data preprocessing. Data from the timing test are available at https://osf.io/y8js6/.

Procedure

Each trial began with the participant hearing either a verbal or a non-verbal cue, followed by a one-second delay. Participants then viewed a visual stimulus, which remained on-screen until a response was made. Participants decided if the audio cue was congruent or incongruent with the visual stimuli, which they indicated by pressing either a "Match" or a "Mismatch" labeled key on the keyboard (see Figure 1 above). We randomized the key labels between participants to account for potential left/right biases in responses. We instructed participants to keep their gaze on the fixation cross when it was present and to keep their head still and blink only between trials to minimize interference with the EEG recording.

Participants began with 10 practice trials, randomly sampled from all possible trial types. Participants completed 500 trials of the cue-recognition task, organized into 5 blocks of 100 trials, with breaks between blocks to rest their eyes. Half of the trials were congruent (the audio cue matched the image) and the other half were incongruent (the audio cue did not match the image). Half of the trials were cued by a non-verbal sound (e.g., a train whistle), and the other half were cued by a linguistic label (e.g., "train"). Our randomization procedure ensured that each image appeared as a match and mismatch equally often. In total, each participant completed 125 trials of each combination of congruence and cue type.

Data collection and EEG preprocessing

The EEG was recorded using a 128 Ag/AgCl electrode net. (Boutonnet and Lupyan used a 64-channel net.) The sampling rate used was 1,000 samples/s referenced to Cz, and the data were amplified using a Net Amps 400 Amplifier (Electrical Geodesics Inc.). We monitored 17 individual sensors that were analogous to those monitored in the original study. To measure the P1 and P2 we used the following electrodes (EGI sensor numbers shown in parentheses): PO3(67), PO4(77), PO7(65), PO8(90), PO9(68), PO10(94), O1(70), O2(83). For the N4, we used FC1(13), FC2(112), FCz(6), C1(30), C2(105), Cz(129), CP1(37), CP2(87), CPz(55). Eye movement and blinks were monitored using electrodes placed above, below, and to the side of each eye. Data were filtered offline using Netstation 5.4 waveform tools by first applying a high pass filter at 0.1 Hz and a low pass filter at 30 Hz. Data were split into 700 ms segments which started 100 ms before the stimulus onset and ended 600 ms after stimulus onset.[1] Stimulus onset time was corrected based on our timing test (see Stimuli, above). An artifact detection function applied to the eye electrode channels was included to remove epochs where activity exceeded the default max - min ranges with an 80 ms moving average for eye blinks (150μV) and horizontal eye movements (55μV). Epochs with more than 20 bad channels (defined as exceeding a moving average range of 200μV) and incorrect trials were also excluded. Remaining good epochs had the Netstation bad channel replacement tool applied to the EEG data which were re-referenced using an average reference and baseline corrected to the 100 ms prior to stimulus onset. This processing pipeline is similar to steps used by Boutonnet & Lupyan (2015) but varies in some respects due to the use of a different EEG system. Most notably, we removed epochs with ocular artifacts while they used ICA to remove components associated with eye movements. After exclusions, the number of included segments per subject ranged from 213 to 485 (out of 500 possible), with an average of 403 (*SD* = 74.1).

1 Our pre-registration specified an epoch length of 1,100 ms, from -100 to 1,000 ms relative to the visual stimulus onset. We found that many participants blinked or moved immediately after pressing the response key, creating a high rate of artifacts in the 600 - 1,000 ms portion of segments. These segments were then rejected by the artifact detection tool. Given that none of the planned analyses required data from this late in the segment, we elected to generate shorter segments to preserve more data.

Results

Analysis was conducted in the R environment (v. 4.0; R Core Team, 2020), using the lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), patchwork (Pedersen, 2019), and pracma (Borchers, 2019) packages, as well as several packages from the tidyverse (Wickham et al., 2019). Our analysis script, with the full set of model results, is available at https://osf.io/u3ygb/.

Analysis Strategies

We utilized three analysis strategies. In each section we first report our pre-registered analysis, which follows the same modeling choices that Boutonnet and Lupyan made. These analyses predominantly used linear mixed-effects models with a subset of the possible random effects. For these analyses we report p-values, with the goal of seeing whether the pattern of statistically significant results is the same in our sample as in the original. Because our replication is designed to be adequately powered to detect effects that are large enough for the original study to have detected, we can treat null findings in our replication as evidence that the original study does not provide compelling evidence of a detectable effect (Simonsohn, 2015).

Our other two approaches were conducted as exploratory analyses based on the suggestion of a reviewer. For these analyses, we fit mixed effects models with a maximal random effects structure using Bayesian estimation. The maximal random effects structure of these models better captures possible sources of variability in the data, including, for example, the possibility that different images produce systematically different ERPs or behavioral responses (Barr et al., 2013).

Fitting the models using Bayesian estimation allows for a different, and arguably more direct, approach to examining the evidence for/against replication. By setting the priors of the fixed effects in the model to match what Boutonnet and Lupyan found and estimating posterior distributions after updating the model with our sample of data, we can quantify the change in beliefs about specific parameter values using Bayes Factors. For example, when our pre-registered analysis finds non-significant results that conflict with the original results, we quantify the change from prior to posterior for the null hypothesis. This is what we did for our second analysis approach.

To set the priors we used the β and t values reported by Boutonnet and Lupyan for the fixed effects in their models to calculate the standard error of the coefficient (SE = β/t). We set the prior as a normal distribution with mean equal to β and standard deviation equal to the SE. In cases where β and t values were not reported by Boutonnet and Lupyan, we assumed that the effect was null and set the prior as a normal distribution with mean equal to 0 and standard deviation equal to the smallest coefficient that was reported as a significant result in the model. We reasoned that this put the bulk of the prior in a range that we can be reasonably confident would have been a non-significant result in the original analysis. For the priors on the random effects and intercept we used the default priors from brms, which are weakly informative (Bürkner, 2017).

The main goal of this analysis is to quantify the change from prior to posterior of the probability of the null hypothesis, i.e., calculate a Bayes Factor for the point null hypothesis (Wagenmakers et al., 2010). We did this using the hypothesis() method of brms (Bürkner, 2017). We only report Bayes Factors for fixed effects in which Boutonnet and Lupyan report their estimates of β and t.

Additionally, the posterior distributions, when using priors that reflect the results found by Boutonnet and Lupyan, are an estimate of the parameter values taking into account the data from both studies. While this would seem to be a nearly ideal way to quantify our beliefs about the effects post-replication, we are cautious about interpreting these estimates because (1) we are using a different random effects structure than Boutonnet and Lupyan for these models, and (2) we are approximating the prior from an incomplete set of results from the original paper. Both of these differences mean that there are sources of uncertainty in the prior that we are not attempting to model, and thus the posteriors from our model will overweight Boutonnet and Lupyan's data relative to our data.

Thus, for our third analysis approach, we also fit the models using a set of moderately informative priors that reflect knowledge about the scale of possible effects but still assume that smaller effects are more likely than larger effects. We set the prior on each fixed effect as a normal distribution with mean equal to 0 and a standard deviation equal to twice the size of the largest effect reported by Boutonnet and Lupyan for that model. These priors capture the plausible scale of the effect and reflect the belief that, absent other knowledge, small effects are more likely than large effects. We think this approach is particularly useful as a comparison point for the first analysis approach, since the first approach uses only a subset of the maximal random effects structure. This analysis may capture some additional sources of variability in the data and give us a clearer picture of the fixed effects.

For all of the Bayesian models, we utilized brms (Bürkner, 2017) and Stan (Stan Development Team, 2019). We used 8 chains, each with 1,000 steps of warmup and 3,000 steps of sampling. The $\widehat{R}$ value for all parameters was less than or equal to 1.01 and the effective sample size (bulk and tail) was greater than 1,300.[2] There were no divergent transitions in the samplers after warm up.

Behavioral Analyses

Boutonnet and Lupyan found that participants responded slightly more accurately when cued by a label as

2 It was typically much larger than this, and exact values can be found in our analysis notebook on the OSF. This is the smallest value across all parameters and all models.

Table 1.1. Predicting response times from cue type and congruence. Fixed-effects estimates for pre-registered replication model fit using *lmerTest*.

Parameter	Estimate	SE	t	*p*
Intercept	592.7	24.2	24.502	< 2e-16
Cue Type (Sound)	40.7	7.2	5.687	2.14e-07
Congruence (Mismatch)	31.8	8.5	3.718	0.000454
Cue Type:Congruence	8.1	8.3	0.973	0.330323

Table 1.2. Predicting response times from cue type and congruence. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and moderately informative priors.

Parameter	Estimate	95% Credible Interval	BF_{01}
Intercept	590.7	541.3 to 639.6	-
Cue Type (Sound)	40.1	22.8 to 56.9	-
Congruence (Mismatch)	30.3	12.4 to 47.9	-
Cue Type:Congruence	8.2	-11.8 to 28.0	-

opposed to a sound. We replicated this result in our sample. Overall accuracy was 97.1% (*SD* = 3.1%) for label cued trials and 96.1% (*SD* = 3.3%) for sound cued trials, a statistically significant difference, $t(34) = 3.00$, $p = 0.005$. Boutonnet and Lupyan reported accuracies of 97% and 95% for the two conditions, making the overall accuracy very similar across the two experiments.

We also replicated Boutonnet and Lupyan's finding of faster response times to label cues and faster response times to congruent trials (Figure 2). We fit a linear mixed effects model to the response time data (correct responses only), with cue type, congruence, and their interaction as fixed effects and random slopes and intercepts for the main effects of cue type and congruence by participant. Table 1.1 summarizes the model's estimates of the fixed effects. Consistent with Boutonnet and Lupyan's results, reaction time was faster for label cues as opposed to sound cues; congruent trials showed faster reaction times as compared to incongruent trials; and no significant interaction effect was found between cue type and congruence.

Following the analysis strategy described above, we fit an exploratory model with the maximal random effects structure using Bayesian estimation and two different sets of priors. This model had fixed effects of cue type, congruence, and their interaction, and random effects of cue type, congruence, and their interaction for both subjects and images.

With the moderately informative priors (Table 1.2), this model estimated that reaction times were faster for label cues as opposed to sound cues and faster for match trials than mismatch trials. The interaction between cue type and congruence was not credibly different from 0.

With the Boutonnet and Lupyan priors (Table 1.3), we estimated Bayes Factors for the point null hypotheses that there is no effect of cue type on response time and no effect

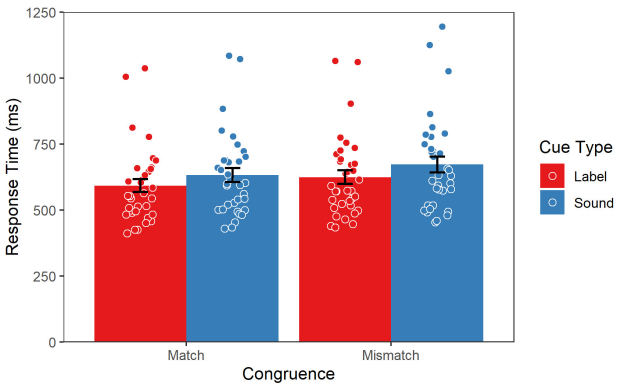

Figure 2. Effects of Cue Type and Congruence on Mean Correct Reaction Time (Error Bars Indicate +/- 1 S.E.M.)

of congruence on response time. The estimated Bayes Factor for both hypotheses was 0 because the posterior sample didn't contain any values close to 0. In this case we can say that the Bayes Factor is very small, though it is not strictly 0. This means that our replication should substantially increase our belief against the null for both effects.

Across all three analysis methods and all measures our behavioral results are consistent with the results of Boutonnet and Lupyan. The biggest deviation is that our estimate of the effect of cue type on response time (40ms advantage for label-cued pictures) was much larger than their estimate (10ms advantage for label-cued pictures).

Table 1.3. Predicting response times from cue type and congruence. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and Boutonnet and Lupyan priors.

Parameter	Estimate	95% Credible Interval	BF_{01}
Intercept	570.2	516.6 to 621.6	-
Cue Type (Sound)	12.0	8.2 to 15.8	0*
Congruence (Mismatch)	29.7	26.1 to 33.4	0*
Cue Type:Congruence	15.24	1.5 to 28.9	-

* The BF_{01} is estimated to be 0 when the sampler never visits values sufficiently close to 0 to estimate the density of the posterior at 0. We can treat these Bayes Factors as being very small, even though they are not strictly 0.

ERP Analysis

ERP Time Windows

Boutonnet & Lupyan (2015) used post-stimulus onset time windows of 70–125 ms for the P1, 190–230 ms for the P2, and 300–500 ms for the N4.[3] When we applied these time windows to our data it was clear that our timing was shifted from the original time windows (see Figure 3). In accordance with our pre-registration, which specified that if we failed to replicate the results using these time windows we would use the procedure described in the original paper to identify new time windows,[4] we created grand average waveforms for P1, P2, and the difference wave for N4 (mismatch - match) across all subjects, conditions, and waveform-relevant electrodes. We showed these waveforms to a colleague familiar with ERP analysis but unfamiliar with our particular study and asked him to select time windows that matched the duration of the windows reported by Boutonnet and Lupyan. Note that this procedure is unbiased with respect to the factors of interest in the study, since we have collapsed across all conditions to perform this selection. He identified time windows of 35–90 ms for P1, 170–210 ms for P2, and 170–370 ms for N4. We used these time windows for our analysis.

Effect of Cue Type and Congruence on Early ERP Components

We examined whether the mean amplitudes[5] of the P1 and P2 were affected by cue type and congruence using the three approaches described in our analysis strategy (above).

First, we matched the analysis of Boutonnet and Lupyan by fitting a linear mixed effects model with fixed effects of cue type, congruence, hemisphere, all interactions of those three terms, and a random intercept and slopes of cue type, congruence, and hemisphere by subject. We fit this model twice, once to predict P1 amplitude and once to predict P2 amplitude. Table 2.1 summarizes the model's estimates.

For the P1, the primary finding is that we did not replicate Boutonnet and Lupyan's result of a more positive P1 amplitude for label trials than sound trials; numerically, our results were in the other direction. Results for the P2 were similar. Sound-cued trials elicited numerically more positive P2 amplitudes, but this effect was not statistically significant . Both of these results are in the opposite direction as those found by Boutonnet and Lupyan.

We found more positive mean amplitudes for incongruent trials than congruent trials in both the P1 and the P2. This replicates Boutonnet and Lupyan's finding for the P2, but they found no effect of congruence on the amplitude of the P1.

We also found an interaction between cue type and congruence for both the P1 and P2 , with label cues eliciting larger mean amplitudes than sound cues in the incongruent condition (Figure 4). These interactions were not found in Boutonnet and Lupyan's sample.

Finally, we found no effect of hemisphere and no interactions involving hemisphere for either the P1 or the P2, consistent with Boutonnet and Lupyan.

Next we fit exploratory linear mixed effects models with the maximal random effects structure using Bayesian estimation. These models had fixed effects of cue type, congru-

3 Boutonnet and Lupyan found no effects for N1 so we did not analyze that time window.

4 We did not conduct any statistical analysis on the original time windows since it was obvious from inspecting the waveforms that we would be measuring something completely different than the original experiment given the differences in the timing of the P1, P2, and N4. Thus, conducting our pre-registered analysis on the original time windows wouldn't produce any valid inferences, as the models would be examining a different part of the ERP than the original study.

5 It is unclear whether Boutonnet and Lupyan used *mean* or *peak* amplitude in their analyses of P1 and P2 group-level effects. The methods section states that the "P1 ... and P2 analyses [used] mean ERP amplitudes" (pg. 9330) while the results section refers to "P1 peak amplitudes" (pg. 9331). We used mean amplitude, given that it is generally regarded as the preferred method (Luck, 2014). Additionally, when Boutonnet and Lupyan clearly use peak amplitude in the single-trial analysis of the P1, they compute the peak averaging over hemispheres. Since these models include hemisphere as a factor, we think it is most likely that they used the mean amplitude in this analysis. As a sensitivity check, we ran a modified version of these models (dropping hemisphere and including only a random intercept by subject due to convergence problems with a more complex random effects structure) and found approximately the same pattern of results, though the effect of cue type that was borderline non-significant in both models switched sides over the $p = 0.05$ threshold. However, the effect is in the opposite direction to that found by Boutonnet and Lupyan. Even if we were to treat these jumps from one side of the $p = 0.05$ threshold to the other as theoretically meaningful, which we do not, none of these changes would impact our conclusions.

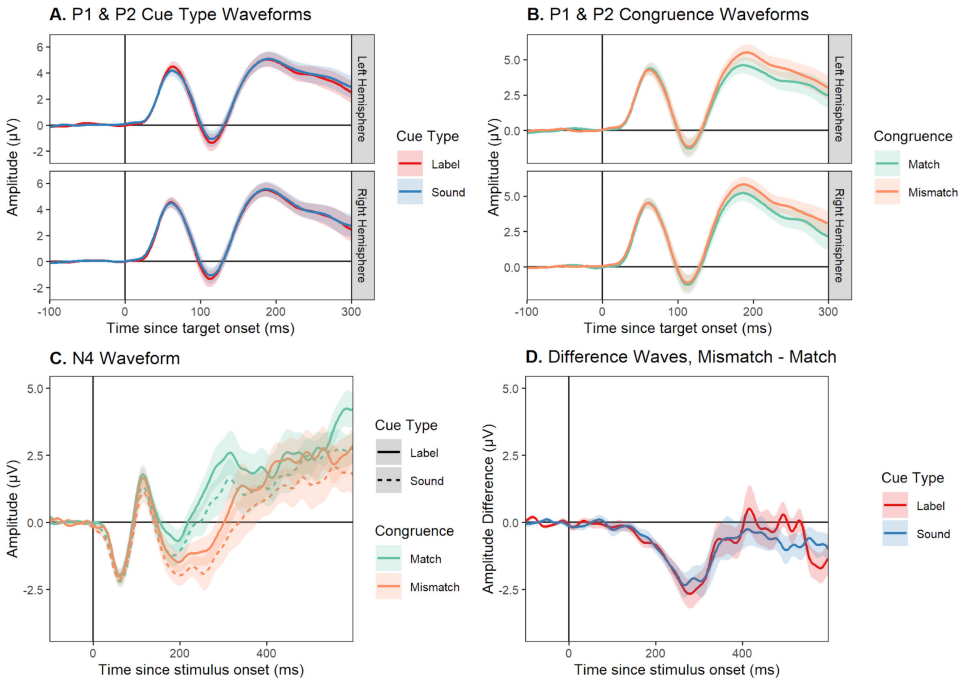

Figure 3. ERP Waveforms. The figure shows the impact of cue type (A) and congruence (B) on P1 and P2. (C) shows the effect of both factors on the N4. (D) shows mismatch - match difference waves of the N4, illustrating the typical semantic incongruity effect.

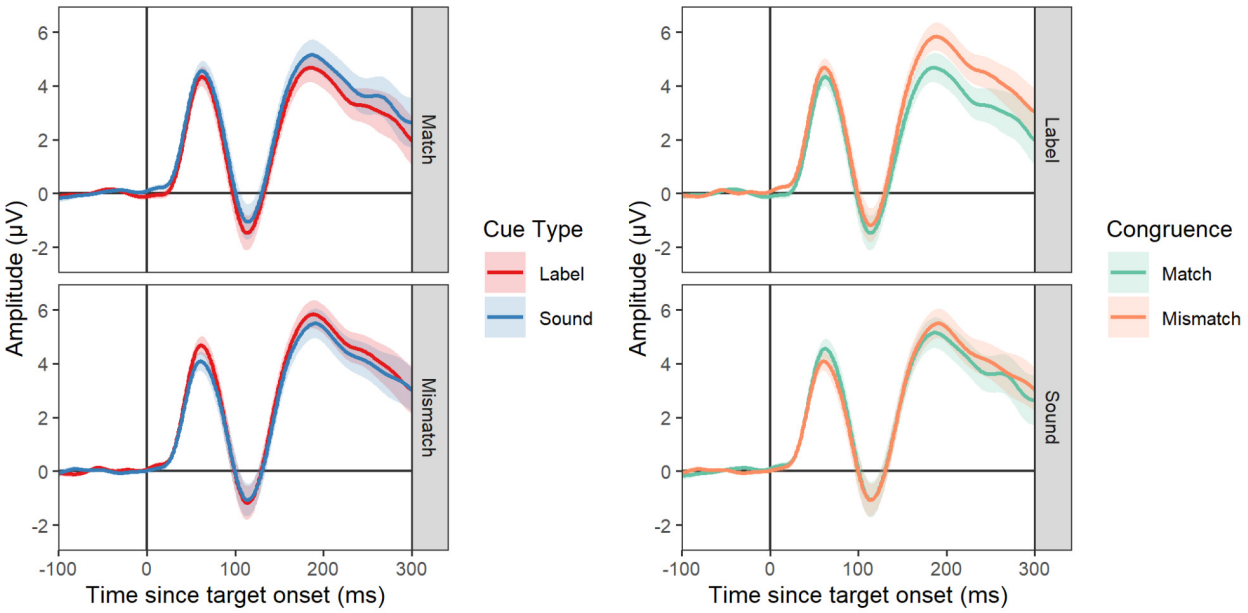

Figure 4. Interaction of Cue Type and Congruence on P1 and P2, shown two ways. Label-cued trials appear to produce more positive P1s than sound-cued trials only when the audio cue does not match the image.

ence, and hemisphere, plus all of the interactions between these terms, and a random intercept and slopes of all of these terms, including the interactions, by subject and by stimulus (image). We fit the model with both moderately informative priors and priors based on the effects that Boutonnet and Lupyan reported (see analysis strategy, above).

With moderately informative priors, the only fixed effects with a credible interval that excluded 0 were the interaction between cue type and congruence for the P1 and the main effect of congruence for the P2. This is a similar pattern to the partial random effects model, with the main difference that the interaction between cue type and con-

Table 2.1. Predicting the amplitude of early ERP components from cue type, congruence, and hemisphere. Fixed-effects estimates for pre-registered replication model fit using *lmerTest*.

Parameter	Component	Estimate	SE	t	*p*
Intercept	P1	3.01	0.40	7.585	4.15e-09
Cue Type (Sound)	P1	0.33	0.21	1.567	0.1190
Congruence (Mismatch)	P1	0.42	0.18	2.345	0.0195
Hemisphere (Right)	P1	0.15	0.37	0.401	0.6899
Cue Type:Congruence	P1	-0.85	0.25	-3.395	0.0007
Cue Type:Hemisphere	P1	-0.14	0.25	-0.569	0.5694
Congruence:Hemisphere	P1	-0.21	0.25	-0.834	0.4041
3-way Interaction	P1	0.54	0.35	1.534	0.1250
Intercept	P2	4.19	0.59	7.136	1.84e-08
Cue Type (Sound)	P2	0.49	0.26	1.906	0.0575
Congruence (Mismatch)	P2	1.33	0.29	4.601	8.10e-06
Hemisphere (Right)	P2	0.68	0.41	1.644	0.1050
Cue Type:Congruence	P2	-0.93	0.35	-2.644	0.0082
Cue Type:Hemisphere	P2	-0.20	0.35	-0.579	0.5628
Congruence:Hemisphere	P2	-0.50	0.35	-1.435	0.1512
3-way Interaction	P2	0.49	0.49	0.980	0.3269

gruence is a weaker effect (plausibly null) in the maximal random effects model.

Next we computed Bayes Factors for the null hypotheses using Boutonnet and Lupyan's estimates as the model's priors. For the null effect of cue type on P1 amplitude, the Bayes factor (BF_{01}) is 6.11, suggesting that our replication results should increase our belief in the null hypothesis by about 6x. For the P2, the BF_{01} for the effect of cue type is 4.14, again suggesting that the replication should increase our belief in the null hypothesis by a moderate amount.

For the effect of congruence on P2 amplitude, the model estimated a BF_{01} of 0. While a BF_{01} of 0 is only possible when the posterior probability of the null hypothesis is exactly 0 and the actual BF_{01} must be greater than 0 here, we can safely infer that the replication should substantially increase our confidence that the effect of congruence on the P2 is not 0.

Finally, our first analysis approach found an interaction between cue type and congruence on both P1 and P2 amplitude. The Bayesian analysis suggests that the evidence for this interaction is mixed. With the maximal random effects structure and moderately informative priors, the estimated effects are smaller. The confidence interval for the interaction between congruence and cue type on P1 amplitude is -1.08 to -0.07, but for the P2 the confidence interval is wider and includes 0 as a plausible value (-1.37 to 0.05). In sum, while there is some evidence for this interaction there are also reasons to be cautious.

Effect of Cue Type and Congruence on the N4

We fit a model predicting N4 amplitudes by cue type and congruence and their interaction, with random slopes by participant for cue type and congruence, and a random intercept for each subject. This follows our first analysis strategy, matching what Boutonnet and Lupyan did. Given the results of the original study, and previous work showing that the N4 is stronger in response to semantic information that is unexpected or harder to integrate in the current context (e.g., Kutas & Federmeier, 2011), we expected more negative amplitudes for the N4 in mismatch trials. We were able to replicate the finding that mismatch trials elicited a more negative N4 than match trials. Additionally, sound cues elicited more negative amplitudes than label cues. We found no significant interaction effect between cue type and congruence for the N4, suggesting that the difference in the N4 amplitude between match and mismatch trials was similar for sound and label cues (Figure 3D).

The Bayesian maximal random effects models find a similar pattern of evidence. These models included random intercepts and slopes of cue type, congruence, and their interaction by subject and by image. With moderately informative priors, the model estimates that mismatch trials produce more negative amplitudes than match trials and that sound trials also produce more negative amplitudes than label trials. The interaction between congruence and cue type is not credibly different from 0. With priors from Boutonnet and Lupyan, the Bayes Factor for the point null hypothesis of congruence is approximately 0, indicating that the replication should strongly decrease our belief that there is no effect of congruence. The Bayes Factor for the point null hypothesis of cue type is 0.25, which means that our belief that there is no effect of cue type should decrease by about 4x based on the replication.

Across all three analysis strategies we replicated the effect of congruence on the N4, but we also found evidence in all three models that the N4 is more negative for sound trials than label trials. This is a non-replication of what Bou-

Table 2.2. Predicting the amplitude of early ERP components from cue type, congruence, and hemisphere. Fixed effects estimates for exploratory model with maximal random effects structure fit using *brms* and moderately informative priors.

Parameter	Component	Estimate	95% Credible Interval	BF_{01}
Intercept	P1	3.08	2.29 to 3.87	-
Cue Type (Sound)	P1	0.20	-0.22 to 0.60	-
Congruence (Mismatch)	P1	0.28	-0.07 to 0.62	-
Hemisphere (Right)	P1	0.06	-0.58 to 0.70	-
Cue Type:Congruence	P1	-0.58	-1.08 to -0.07	-
Cue Type:Hemisphere	P1	0.00	-0.42 to 0.41	-
Congruence:Hemisphere	P1	-0.06	-0.48 to 0.35	-
3-way Interaction	P1	0.28	-0.28 to 0.82	-
Intercept	P2	4.31	3.07 to 5.54	-
Cue Type (Sound)	P2	0.36	-0.20 to 0.90	-
Congruence (Mismatch)	P2	1.11	0.56 to 1.65	-
Hemisphere (Right)	P2	0.50	-0.24 to 1.23	-
Cue Type:Congruence	P2	-0.66	-1.37 to 0.05	-
Cue Type:Hemisphere	P2	-0.03	-0.64 to 0.57	-
Congruence:Hemisphere	P2	-0.30	-0.91 to 0.30	-
3-way Interaction	P2	0.19	-0.61 to 1.01	-

Table 2.3. Predicting the amplitude of early ERP components from cue type, congruence, and hemisphere. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and Boutonnet and Lupyan priors.

Parameter	Component	Estimate	95% Credible Interval	BF_{01}
Intercept	P1	3.22	2.43 to 4.01	-
Cue Type (Sound)	P1	-0.17	-0.39 to 0.05	6.11
Congruence (Mismatch)	P1	0.14	-0.15 to 0.43	-
Hemisphere (Right)	P1	0.05	-0.41 to 0.52	-
Cue Type:Congruence	P1	-0.24	-0.62 to 0.14	-
Cue Type:Hemisphere	P1	0.03	-0.26 to 0.31	2.30
Congruence:Hemisphere	P1	0.01	-0.32 to 0.34	-
3-way Interaction	P1	0.15	-0.24 to 0.54	-
Intercept	P2	4.79	3.59 to 5.97	-
Cue Type (Sound)	P2	-0.23	-0.44 to -0.03	4.14
Congruence (Mismatch)	P2	0.57	0.37 to 0.77	0*
Hemisphere (Right)	P2	0.16	-0.32 to 0.64	-
Cue Type:Congruence	P2	-0.05	-0.48 to 0.37	-
Cue Type:Hemisphere	P2	0.18	-0.21 to 0.58	-
Congruence:Hemisphere	P2	-0.01	-0.39 to 0.38	-
3-way Interaction	P2	-0.06	-0.51 to 0.41	-

* The BF_{01} is estimated to be 0 when the sampler never visits values sufficiently close to 0 to estimate the density of the posterior at 0. We can treat these Bayes Factors as being very small, even though they are not strictly 0.

tonnet and Lupyan described as an "important" null result in their experiment, as they interpreted the lack of effect of cue type on the N4 as evidence that the behavioral results were not being driven by semantic differences. We examine the relationship between the N4 and the behavioral data further in the exploratory analysis section below.

Relationship of the P1 to behavior

If perceptual processing, as indexed by the P1, is altered

Table 3.1. Predicting N4 amplitude from cue type and congruence. Fixed-effect estimates for pre-registered replication model fit using *lmerTest*.

Parameter	Estimate	SE	t	*p*
Intercept	1.17	0.51	2.275	0.0289
Cue Type (Sound)	-0.84	0.25	-3.408	0.0007
Congruence (Mismatch)	-1.51	0.34	-4.403	4.47e-05
Cue Type:Congruence	0.11	0.35	0.332	0.7401

Table 3.2. Predicting N4 amplitude from cue type and congruence. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and moderately informative priors.

Parameter	Estimate	95% Credible Interval	BF_{01}
Intercept	1.11	0.03 to 2.15	-
Cue Type (Sound)	-0.80	-1.30 to -0.30	-
Congruence (Mismatch)	-1.43	-2.11 to -0.75	-
Cue Type:Congruence	0.04	-0.65 to 0.72	-

Table 3.3. Predicting N4 amplitude from cue type and congruence. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and Boutonnet and Lupyan priors.

Parameter	Estimate	95% Credible Interval	BF_{01}
Intercept	0.65	-0.36 to 1.65	-
Cue Type (Sound)	-0.17	-0.34 to -0.01	0.25
Congruence (Mismatch)	-0.88	-1.12 to -0.64	0*
Cue Type:Congruence	-0.57	-1.05 to -0.08	-

* The BF_{01} is estimated to be 0 when the sampler never visits values sufficiently close to 0 to estimate the density of the posterior at 0. We can treat these Bayes Factors as being very small, even though they are not strictly 0.

by the presence of a label cue and this process explains the observed differences in response time, then the P1 should show systematic relationships with response time at the single-trial level. Boutonnet and Lupyan found that the peak amplitude and peak latency of the P1 both predicted response time. While we did not find the effect of cue on P1 amplitude that Boutonnet and Lupyan did, we followed their analysis and first averaged the eight electrodes that were used to measure the P1 to create a single averaged ERP per trial. We then extracted the peak (if one existed; it is possible that the waveform only increased or decreased, or decreased and then increased) in the P1 window using the *pracma* package in R (Borchers, 2019). We measured the latency and amplitude of this peak.

Following Boutonnet and Lupyan, we fit a linear mixed effects model to predict response time from the fixed effects of peak amplitude, peak latency, cue type, and congruence. Cue type and congruence are included in the model as fixed effects because the behavioral analysis found that both of these are strong predictors of response time. We included random intercepts and slopes for cue type and congruence by participant and by image category. This model resulted in a singular fit, with high correlations in the random slopes of cue type and congruence, especially for image category, and relatively little variance in the random intercept or slopes for image category. Thus, we decided to drop the random slopes by image category, and run the model again with just a random intercept by image category (keeping the random slopes for cue type and congruence by subject). This model converged and the estimates of the fixed effects were very similar to the estimates from the more complex model. We summarize the estimates from the simpler model in Table 4.1, and the results from the more complex model can be found in our analysis code. We found that neither peak latency nor peak amplitude of the P1 was significantly predictive of response time.

The maximal random effects models again produced a similar pattern of evidence (with no evidence of poor convergence). This model included fixed effects of peak amplitude, peak latency, cue type, and congruence, with random intercepts and slopes of peak amplitude, peak latency, cue type, and congruence by subject and by image (not image category). With moderately informative priors, the model estimated that the effects of peak amplitude (95% credible interval: -1.07 to 0.36) and peak latency (95% credible interval: -0.37 to 0.46) were both plausibly 0. However, using the priors based on Boutonnet and Lupyan's estimates, we found a BF_{01} of 0.54 for the null hypothesis of peak latency,

Table 4.1. Predicting single-trial response times from P1 peak latency, P1 peak amplitude, cue type and congruence. Fixed-effects estimates for pre-registered replication model fit using *lmerTest*.

Parameter	Estimate	SE	t	*p*
Intercept	591.1	24.34	24.289	<2e-16
Peak Latency	0.001	.17	0.006	0.9952
Peak Amplitude	-0.46	.28	-1.65	0.0989
Cue Type (Sound)	43.76	6.86	6.378	5.35e-07
Congruence (Mismatch)	40.07	8.377	4.783	3.81e-05

Note that this model deviates slightly from the pre-registered model (the model that Boutonnet and Lupyan used) because of convergence problems with the original model. This model has a simpler random effects structure. See text for details.

Table 4.2. Predicting single-trial response times from P1 peak latency, P1 peak amplitude, cue type and congruence. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and moderately informative priors.

Parameter	Estimate	95% Credible Interval	BF_{01}
Intercept	584.75	528.39 to 639.94	-
Peak Latency	0.05	-0.37 to 0.46	-
Peak Amplitude	-0.35	-1.07 to 0.36	-
Cue Type (Sound)	42.57	27.20 to 57.81	-
Congruence (Mismatch)	37.84	19.27 to 56.12	-

Table 4.3. Predicting single-trial response times from P1 peak latency, P1 peak amplitude, cue type and congruence. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and Boutonnet and Lupyan priors.

Parameter	Estimate	95% Credible Interval	BF_{01}
Intercept	556.54	507.45 to 605.77	-
Peak Latency	0.23	0.06 to 0.41	0.54
Peak Amplitude	-0.49	-0.87 to -0.12	0.90
Cue Type (Sound)	11.97	8.28 to 15.71	-
Congruence (Mismatch)	29.90	26.17 to 33.58	-

suggesting that the replication should slightly reduce our belief in the likelihood of the null hypothesis. We found a BF_{01} of 0.90 for the null hypothesis of peak amplitude, indicating that belief in the null hypothesis should be essentially unchanged by the replication.

Modulation of the P1 by labels

Boutonnet and Lupyan reported a second analysis at the single-trial level, examining whether peak latency and peak amplitude of the P1 can predict the congruence of a trial on label trials. They used a generalized linear mixed model (logistic) predicting trial congruence from the peak amplitude and peak latency of the P1 and cue type, as well as all interactions of these three terms, with random slopes for cue type by participant and image category. Again, although we did not obtain the prior effect of cue on the P1, we attempted to fit this model, as per our pre-registered plan, but ran into difficulty with convergence. We explored possible modifications to the model, including (1) centering and normalizing peak time and peak amplitude as predictors, (2) dropping the random effect of cue type and the random intercept by subject, since cue type is not predictive of congruence by the nature of the experimental design because half of the trials for each subject will be congruent, (3) dropping random effects by image category, and (4) adding random effects of peak latency and peak amplitude by subject to better reflect the hierarchical structure of the data, and combinations of all of the above. None of these models adequately converged, despite trying a variety of optimization algorithms.

We then tried running these models using the maximal random effects structure and Bayesian estimation, with both moderately informative priors and priors based on Boutonnet and Lupyan's results. Note that the model structure that is justified by the design includes no main effect of

Table 5.1. Predicting congruence type from P1 peak latency, P1 peak amplitude, and cue type. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and moderately informative priors.

Parameter	Estimate	95% Credible Interval	BF_{01}
Intercept	0.0028	-0.2357 to 0.2398	-
Peak Latency	0.0003	-0.0035 to 0.0042	-
Peak Amplitude	-0.0204	-0.0481 to 0.0070	-
Peak Latency:Cue Type (Sound)	-0.0011	-0.0028 to 0.0007	-
Peak Amplitude:Cue Type (Sound)	-0.0050	-0.0379 to 0.0271	-
Peak Latency:Peak Amplitude	0.0002	-0.0002 to 0.0006	-
Three-way Interaction	0.0003	-0.0002 to 0.0008	-

Table 5.2. Predicting congruence type from P1 peak latency, P1 peak amplitude, and cue type. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and moderately informative priors.

Parameter	Estimate	95% Credible Interval	BF_{01}
Intercept	0.0022	-0.2363 to 0.2420	-
Peak Latency	0.0003	-0.0035 to 0.0042	-
Peak Amplitude	-0.0192	-0.0472 to 0.0086	-
Peak Latency:Cue Type (Sound)	-0.0010	-0.0028 to 0.0007	41,985
Peak Amplitude:Cue Type (Sound)	-0.0075	-0.0400 to 0.0248	153
Peak Latency:Peak Amplitude	0.0002	-0.0002 to 0.0006	-
Three-way Interaction	0.0003	-0.0002 to 0.0008	-

cue type, since cue type was equated across match and mismatch trials in the design. We do, however, include terms that interact with cue type. The model included fixed effects of peak latency and peak amplitude, the interaction of peak amplitude with cue type, the interaction of peak latency with cue type, the interaction of peak amplitude and peak latency, and the three-way interaction between cue type, peak amplitude and peak latency. We also included random intercepts and slopes of all these terms by subject and by image. We hoped that incorporating moderately informative priors would aid in model convergence, and indeed the basic convergence diagnostics indicated no issues. However, with both sets of priors the posterior distributions for all of the fixed effects are very concentrated at 0, and the BF_{01} for the interactions between cue type and peak latency as well as cue type and peak amplitude are strongly supportive of the null hypothesis. The model's unusual certainty, compared to the rest of our analyses, makes us skeptical of the fit despite no obvious indicators of convergence problems. We report the model coefficients in Tables 5.1 and 5.2 for the Bayesian models, and we document the full set of models and (unconverged) fits at https://osf.io/u3ygb/. Our interpretation of this set of analyses is that our data are inconclusive on this portion of the replication.

Exploratory analysis of the relationship between P1, N4, and response times

In contrast to Boutonnet and Lupyan, we found that the N4 was more negative for sound trials than label trials. This suggests a possible semantic-level explanation of the behavioral results. If labels act as generic conceptual pointers and sounds as more specific pointers (Edmiston & Lupyan, 2015), then the semantic mismatch between a sound and image should be larger (on average) than between a label and image because the sound cues a more specific member of the category. This mismatch may be reflected by the amplitude of the N4 (Kutas & Federmeier, 2011). If semantic-level processes are driving the behavioral effects, then more negative N4 amplitude should predict longer response times given that there is ample other evidence of such a relationship (e.g., presenting a prime before a related target word reliably reduces both reaction time and negative N400 amplitude to the target, though there are also cases of dissociation as in Chwilla et al., 2000).

We tested this idea by fitting a linear mixed effects model to predict response time from the mean amplitude of the N4 at the single-trial level. We included fixed effects of cue type and congruence and random intercepts and slopes of cue type, congruence, and mean amplitude by subject and by image. Our rationale for including cue type and congruence as predictors is that we wanted to determine if N4 amplitude was predictive of response time *after* controlling for overall differences across cue and congruence conditions. If there is a relationship between N4 amplitude *within* each condition then this provides stronger evidence that the correlation is not merely due to condition-level manipulations. We fit this model using Bayesian estimation, with moder-

Table 6. Predicting single-trial response time from N4 amplitude, cue type, and congruence. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and moderately informative priors.

Parameter	Estimate	95% Credible Interval	BF_{01}
Intercept	594.76	551.95 to 637.61	-
Cue Type (Sound)	41.31	26.37 to 56.57	-
Congruence (Mismatch)	33.31	14.31 to 52.63	-
N4 Amplitude	-2.71	-4.37 to -1.09	-

Table 7. Predicting single-trial response time from normalized N4 amplitude, normalized P1 amplitude, normalized P1 latency, cue type, and congruence. Fixed-effects estimates for exploratory model with maximal random effects structure fit using *brms* and moderately informative priors.

Parameter	Estimate	95% Credible Interval	BF_{01}
Intercept	593.33	550.72 to 635.37	-
Cue Type (Sound)	39.92	24.51 to 55.18	-
Congruence (Mismatch)	33.88	14.80 to 53.01	-
Normalized N4 Amplitude	-29.30	-47.53 to -11.24	-
Normalized P1 Peak Amplitude	-4.97	-12.33 to 2.42	-
Normalized P1 Latency	0.43	-5.22 to 5.98	-

ately informative priors. Since we did not have a comparable analysis from Boutonnet and Lupyan to set the prior on N4 amplitude, we set the prior as a normal distribution centered on zero with a relatively wide standard deviation of 10. The model results are summarized in Table 6. We found that the mean amplitude of the N4 is predictive of response time, with more negative amplitudes resulting in slower response times[6].

Based on this result and the somewhat ambiguous evidence surrounding the null hypothesis that P1 peak amplitude and latency are *not* predictive of response time, we next decided to compare these factors directly by fitting a model predicting response time from the fixed effects of P1 latency, P1 amplitude, N4 amplitude, congruence, and cue type. We included random intercepts and slopes for all of these predictors by subject and by image. We decided to center and normalize the ERP predictors to ensure an apples-to-apples comparison between them, given that ERP components vary in magnitude. We fit the model using Bayesian estimation, following the same strategy as in our other analyses. Our prior on the ERP fixed effects was a normal distribution centered at 0 with a standard deviation of 100, selected to be only mildly informative about the scale of possible effects. The results are summarized in Table 7. The only ERP predictor that was definitively non-zero was N4 amplitude, further supporting a possible semantic-level explanation of the behavioral results.

Discussion

As in Boutonnet and Lupyan's (2015) original experiment, participants in our replication responded faster and more accurately to visual images when preceded by linguistic cues rather than non-verbal sound cues. These results support the "labels-as-pointers" view that applying linguistic labels helped the participants in the object recognition tasks more than unambiguous nonverbal sounds did and is not surprising given that this effect has been replicated several times (Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012). In order to determine whether this benefit of linguistic labels was operating during an early visual processing stage or a later semantic stage, Boutonnet and Lupyan employed ERP methods and found an effect of labels on the P1 but not the N4. Our close and pre-registered replication of this study did not produce the same pattern of ERP effects.

With regard to early ERP effects, our results did not show the straightforward relationship between amplitude and cue type shown in the original study that provided the main support for the claim of top-down effects of language on perception. In particular, our data did not show a main ef-

6 We also analyzed this relationship using a variety of non-Bayesian mixed effects models, but we present the Bayesian version here because we prefer interpreting the credible intervals from Bayesian models and because we were able to use the maximal random effects structure in the Bayesian version. The full set of models that we used is documented in our analysis R notebook at https://osf.io/u3ygb/. In all models that we fit, we found that N4 amplitude was predictive of response time at the single-trial level, with similar estimates for the coefficient of amplitude.

fect of cue type on mean amplitudes for P1 or P2, and numerically, the pattern was in the opposite direction. Bayes factors suggest that our replication data should increase our belief in a null effect of cue type for P1 and P2 by a moderate amount. Effects of congruency on amplitude were mixed as we found significant effects for both P1 and P2, rather than just for P2. Our single-trial analyses correlating features of the ERP components with the behavioral data also deviated from the original results in that we found no significant effect of P1 peak latency or peak amplitude on response times, though Bayes factors suggest that the evidence here is not particularly strong one way or another. Finally, whereas they found no significant interaction effects of cue type and congruence on P1 and P2, our data indicated significant interaction effects on both. However, because this result was unpredicted, we are not confident that it reflects a genuine effect of linguistic labels on early visual processing, particularly since the pattern we observed does not lend itself naturally to any clear interpretation (see Figure 4). Thus our data overall do not provide clear evidence of an early perceptual benefit of labels over sounds in a picture matching task.

With regard to later semantic processing as indexed by the N4, we replicated the main effect of congruence showing larger negative amplitudes for mismatch than match trials. This was expected based on well-established previous N4 research. We also replicated the lack of interaction of cue type and congruence on N4, meaning that the difference in N4 amplitude between match and mismatch trials was approximately the same for label cues and sound cues. However, we found a main effect of cue type on the N4 with sound cues showing larger negative amplitudes than label cues. This was unexpected, but a plausible explanation is that the more specific sound cues (e.g., a particular dog's bark that sounds like a large dog) are more likely than the more generic linguistic labels to create an expectation for the visual image that is not met regardless of whether the cue category matches the image or not. If larger N4 amplitudes reflect greater difficulty of integration and/or prediction, then these N4 results are consistent with the pattern of response times in our data and suggest a possible semantic-level explanation for the behavioral effect.

We probed the relation between N4 amplitude and response time further in a single-trial analysis patterned on the single-trial analyses in the original paper for the P1 and found that N4 amplitude predicts response time within each condition, thus reflecting something broader than condition-level effects. In other words, even after controlling for which kind of trial a participant was in, larger N4 amplitudes still predicted slower response times. Though our analysis was exploratory, this result was consistent across a variety of statistical models, suggesting that it is fairly robust. Boutonnet and Lupyan didn't report single-trial analysis of the N4, so it is possible that this relationship held in their data as well. Furthermore, we also directly compared N4 amplitude with P1 amplitude and P1 latency as predictors of behavioral responses. We found that only the N4 amplitude was predictive of response times at the single trial level. Taken in conjunction with our finding that cue type modulated the N4, this evidence suggests that labels were influencing later semantic processes and that these processes were at least partially responsible for the behavioral results.

To summarize, our close and pre-registered replication of Boutonnet & Lupyan (2015) yielded consistent behavioral results but ERP patterns that were not only inconsistent but actually suggest that the beneficial effect of labels vs. sounds on judgments of matching and mismatching visual images was occurring at the semantic (N4) stage rather than an earlier perceptual (P1/P2) stage of processing. However, several major caveats are in order before treating the latter as firm negative evidence concerning the existence of genuine top-down effects on visual perception. First, while the N4 results we obtained are suggestive, they were unpredicted and hence exploratory and would certainly need to be replicated. In addition, while we did not obtain the critical effect of labels vs. sounds on P1/P2 amplitude that Boutonnet and Lupyan did, our data did show interaction effects that could conceivably represent some sort of early perceptual effect, though again, needing replication in addition to an interpretive rationale.

Another caveat here concerns the challenge of conducting exact replications, which might seem like the gold standard but are in fact impossible, as noted by Nosek & Errington (2020). These authors raise the question of what changes count as minor enough to still qualify as repeating the procedure, as well as point out that the scientific claims of an experiment are always intended to generalize to some degree beyond the specific conditions initially observed. We would argue that conducting direct replications is especially challenging with ERP studies, which afford numerous additional "researcher degrees of freedom" that are often seemingly inconsequential but may create additional opportunities for false-positive findings (Luck & Gaspelin, 2017). While new efforts are emerging to help transparently address some of this flexibility (Kappenman et al., 2021), Clayson et al. (2019) point out that insufficient following of accepted ERP experiment reporting guidelines and small sample sizes, hence low statistical power, greatly reduce the methodological transparency and replicability of ERP studies. In addition, to have a basis for generalizable claims, effects have to be replicable when the methods and equipment used are similar but not necessarily identical to the original study. For example, of necessity we used a different type of EEG recording system and processing software, preventing it from being an exact replication. Because our system is a "high impedance" system it produces noisier EEG data and requires some differences in preprocessing; together these deviations from the original experiment could conceivably affect the ability to pick up on subtle ERP patterns. Furthermore, changes in the stimulus presentation and event-marking hardware between replications add additional sources of variability that are particularly impactful in EEG analysis. A striking example of this is that we had to substantially adjust the time windows for the ERP analysis from Boutonnet and Lupyan's reported windows, despite calibrating the timing accuracy of the system prior to running the experiment.

Widespread honoring of best practices for ERP methods and data handling would help support future replication efforts, but these concerns also tie in with big questions regarding generalizability that the field is starting to grapple

with directly (Yarkoni, 2019). Researchers need to be more explicit about which variables they intend the key results to generalize over (the specific EEG recording system used being one example, as well as stimuli, task procedure, participants, and a host of other variables) and find ways to systematically explore whether or not they do. This will be extremely challenging but necessary, we think, in order to make real progress going forward in the testing of important theoretical claims, including whether language can really influence early visual perception or not.

In conclusion, we endorse the proposal that replication, broadly construed, is essential for testing the predictions of theories and making progress in further development of theory (Nosek & Errington, 2020). The replication we report here suggests a number of plausible alternatives to Boutonnet and Lupyan's main theoretical claim that top-down effects explain the reaction time advantage for label-cued images in the behavioral data. These alternatives include that the P1/P2 effects reported in the original study were a false positive, or are dependent on some as yet unspecified contextual factor that happened to differ between the two studies, or require a new version of the top-down theory in order to explain why we obtained interaction effects rather than main effects. We also raise the more radically different alternative that the later stage, semantically-based N4 better explains the speed advantage of labels over sounds and the response time differences across conditions more generally. Given these alternatives, we think that our replication and Boutonnet & Lupyan (2015) do not show clear evidence of a top-down effect of language on visual perception.

Author Contributions

Conception of the experiment: JRD, JA

Design of the experiment: All authors

Acquisition of data: LB, MB, PJB, YC, RC, DRD, EF, BH, XH, MJ, NL, CL, DM, CM, EM-S, KN, WO, RP, GS-R, AW, LW, LZ

Analysis and interpretation of data: All authors

Drafting the article: LB, MB, PJB, YC, RC, DRD, EF, BH, XH, MJ, NL, CL, DM, CM, EM-S, KN, WO, RP, GS-R, AW, LW, LZ

Revising the article: JRD, JA

Final approval: All authors

Acknowledgements

We would like to thank Debra Ratchford, Cole Landolt, and Rena Lee for assisting with EEG training and supervising data collection procedures. We would also like to thank Mante Nieuwland for a variety of helpful comments during peer review.

Data Accessibility Statement

Raw and aggregated data are available on Dataverse (https://doi.org/10.7910/DVN/SWL4YQ) and also accessible through the project's Open Science Framework repository (https://osf.io/cq8g4/). The OSF repository also contains analysis code, experiment code, stimuli, and the pre-registration.

Competing Interests

The authors have no competing interests to declare.

Submitted: February 16, 2021 PST, Accepted: October 21, 2021 PST

REFERENCES

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. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Usinglme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Borchers, H. W. (2019). *pracma: Practical Numerical Math Functions* (R package version 2.2.9). https://CRAN.R-project.org/package=pracma

Boutonnet, B., & Lupyan, G. (2015). Words jump-start vision: A label advantage in object recognition. *Journal of Neuroscience*, *35*(25), 9329–9335. https://doi.org/10.1523/jneurosci.5111-14.2015

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Chwilla, D. J., Kolk, H. H. J., & Mulder, G. (2000). Mediated priming in the lexical decision task: Evidence from event-related potentials and reaction time. *Journal of Memory and Language*, *42*(3), 314–341. https://doi.org/10.1006/jmla.1999.2680

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological reporting behavior, sample sizes, and statistical power in studies of event - related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*(11), e1437. https://doi.org/10.1111/psyp.13437

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition*, *143*, 93–100. https://doi.org/10.1016/j.cognition.2015.06.008

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, *39*, e229. https://doi.org/10.1017/s0140525x15000965

Kappenman, E. S., Farrens, J. L., Zhang, W., Stewart, A. X., & Luck, S. J. (2021). ERP CORE: An open resource for human event-related potential research. *Neuroimage*, *225*, 117465. https://doi.org/10.1016/j.neuroimage.2020.117465

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

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. https://doi.org/10.18637/jss.v082.i13

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (De)constructing the N400. *Nature Reviews Neuroscience*, *9*(12), 920–933. https://doi.org/10.1038/nrn2532

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). MIT Press.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Lupyan, G., Abdel Rahman, R., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in Cognitive Sciences*, *24*(11), 930–944. https://doi.org/10.1016/j.tics.2020.08.005

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, *24*(4), 279–284. https://doi.org/10.1177/0963721415570732

Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General*, *141*(1), 170–186. https://doi.org/10.1037/a0024904

Nieuwland, M. S., Barr, D. J., Bartolozzi, F., Busch-Moreno, S., Darley, E., Donaldson, D. I., Ferguson, H. J., Fu, X., Heyselaar, E., Huettig, F., Husband, E. M., Ito, A., Kazanina, N., Kogan, V., Kohút, Z., Kulakova, E., Mézière, D., Politzer-Ahles, S., Rousselet, G., … Von Grebmer Zu Wolfsthurn, S. (2019). Dissociable effects of prediction and integration during language comprehension: evidence from a large-scale study using brain potentials. *Philosophical Transactions of the Royal Society B*, *375*(1791), 20180522. https://doi.org/10.1098/rstb.2018.0522

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Pedersen, T. L. (2019). *patchwork: The Composer of Plots* (R package version 1.0.0).

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: the role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217–236. https://doi.org/10.1068/p5117

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Stan Development Team. (2019). *Stan Modeling Language User's Guide and Reference Manual, 2.27*. https://mc-stan.org

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Whorf, B. L. (1956). *Language, thought, and reality: Selected writings*. Technology Press of MIT.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Yarkoni, T. (2019). The generalizability crisis. *PsyArXiv*. https://doi.org/10.31234/osf.io/jqw35

SUPPLEMENTARY MATERIALS

Peer Review History

Download: https://collabra.scholasticahq.com/article/29763-words-may-jump-start-meaning-more-than-vision-a-non-replication-of-early-erp-effects-in-boutonnet-and-lupyan-2015/attachment/74873.docx?auth_token=EvhUdEjd_0d0_HM79rul