A meta-analysis of syntactic priming in language production

A meta-analysis of syntactic priming in language production

Kyle Mahowald[1], Ariel James[2], Richard Futrell[1], Edward Gibson[1]

[1] Massachusetts Institute of Technology
[2] University of Illinois at Urbana-Champaign

THIS IS A DRAFT THAT IS UNDER REVIEW AND RESULTS ARE NOT FINAL. For comments and correspondence, e-mail Kyle Mahowald at kmahowald@gmail.com.

Abstract

We performed an exhaustive meta-analysis of 73 peer-reviewed journal articles from the seminal Bock (1986) paper through 2013. Extracting the effect size for each experiment and condition, where the effect size is the log odds ratio of the frequency of the primed structure X to the frequency of the unprimed structure Y, we found a robust effect of syntactic priming with an average weighted odds ratio of 1.67 when there is no lexical overlap and 3.26 when there is. That is, a construction X which occurs 50% of the time in the absence of priming would occur 63% if primed without lexical repetition and 77% of the time if primed with lexical repetition. The syntactic priming effect is robust across several different construction types and languages, and we found strong effects of lexical overlap on the size of the priming effect as well as interactions between lexical repetition and temporal lag and between lexical repetition and whether the priming occurred within or across languages. We also analyzed the distribution of p-values across experiments in order to estimate the average statistical power of experiments in our sample and to assess publication bias. Analyzing a subset of experiments in which the primary result of interest is whether a particular structure showed a priming effect, we did not find evidence of major p-hacking and the studies appear to have acceptable statistical power: 82%. However, analyzing a subset of experiments that focus not just on whether syntactic priming exists but on how syntactic priming is moderated by other variables (such as repetition of words in prime and target, the location of the testing room, the memory of the speaker, etc.), we found that such studies are, on average, underpowered with estimated average power of 53%. Using a subset of 45 papers from our sample for which we received raw data, we estimated subject and item variation and give recommendations for appropriate sample size for future syntactic priming studies.

Keywords: syntactic priming, meta-analysis, statistical power

Introduction

When someone is primed with a syntactic structure X and is then asked to produce a new sentence, it is claimed that they are more likely to use that same structure X than if they had instead heard some other structure Y. This phenomenon, *syntactic priming* (also sometimes called *structural priming* or *syntactic persistence*), has been an important topic of study in psycholinguistics since Bock (1986). Syntactic priming has been used to test theories of event structure (Bunger, Papafragou, & Trueswell, 2013), social interaction (Branigan, Pickering, McLean, & Cleland, 2007), bilingualism (Bernolet, Hartsuiker, & Pickering, 2007, 2013; Schoonbaert, Hartsuiker, & Pickering, 2007), syntactic surprisal (Jaeger & Snider, 2013), childhood linguistic representations (Messenger, 2010), amnesia (Ferreira, Bock, Wilson, & Cohen, 2008), autism (Slocombe et al., 2013), aphasia (Verreyt et al., 2013), implicit learning (Kaschak, Kutta, & Jones, 2011), and human mating behavior (Coyle & Kaschak, 2012). Perhaps most critically, syntactic priming has been used as evidence for the abstractness of syntactic operations (Bock, 1986, 1989). See Ferreira & Bock (2006), Heydel & Murray (2000), and Pickering & Ferreira (2008) for critical reviews of this literature.

As a phenomenon that has become central to the field of psycholinguistics, syntactic priming is ripe for a cumulative quantitative analysis. One of the goals of this meta-analysis is to assess the current state of knowledge in the field by aggregating data and evaluating it quantitatively. All else being equal, how big is the syntactic priming effect? What is the range of variation one could expect? How much bigger should it be when there is lexical overlap between the prime and target? Can the existing literature be trusted, or does it suffer from publication bias?

While there have been several large-scale critical reviews of syntactic priming, there has not been a systematic, large-scale quantitative meta-analysis (but see Jaeger & Snider (2013) for a meta-analysis of three earlier experiments). Meta-analyses, whereby a group of studies are gathered and quantitatively analyzed together, can be useful for assessing what we have learned through a large body of distinct studies and for exploring whether these studies are exploring the same underlying phenomena (Lipsey & Wilson, 2001). Meta-analyses dramatically increase statistical power–that is, the probability of detecting a true effect–by pooling data together. In our meta-analysis, for instance, we included data from over 5,000 unique participants, whereas no single experiment in our sample used more than 144. For these reasons, increased use of meta-analysis in the social sciences has been widely recommended as a way to investigate the reliability of published results (Button et al., 2013b; Cumming, 2013; Simonsohn, Nelson, & Simmons, 2014b).

In this paper, we report three results: a standard meta-analysis, an analysis of publication bias, and recommendations for sample size in future priming studies. We define effect size as the log odds ratio of the proportion of target structure produced in the prime condition to the proportion of target structure produced in the no-prime condition. For 45 of the 73 papers in our sample, we obtained raw data from the authors and used it to derive estimates of effect size and standard error. From the remaining papers, we estimated the effect size and standard error using the published estimates. Along with effect sizes and their associated standard errors, we also collected information on several key manipulations that can potentially modulate the priming effect, including the construction used, lexical repetition, lag, and whether the priming is within or across languages. Using these variables, we estimated the average effect size of syntactic priming given various experimental conditions.

As a secondary analysis, we assessed the extent to which the set of papers in our study suffer from publication bias and low power. Indeed, there have been meta-analyses in other branches of psychology alleging widespread publication bias (Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014; Landy & Goodwin, 2014), low reproducibility (Open Science Collaboration, 2015), and low statistical power (Button et al., 2013b, 2013a). Low statistical power can lead to inflated false-positive rates in the literature and unreliable results (Gelman & Carlin, 2014). To assess publication bias and statistical power, we used p-curve, a tool developed for that purpose which works by analyzing the distribution of significant p-values in the literature (Simonsohn et al., 2014b; Simonsohn, Nelson, &

Simmons, 2014a). Using the raw data gathered from the study authors, we did a power analysis and give guidelines on how to run syntactic priming studies with sufficient statistical power.

In addition to quantifying the state of the field, there are a number of open questions in syntactic priming that we can investigate using this method. For instance, Pickering & Ferreira (2008) describe conflicting evidence as to just how long-lived syntactic priming is. Here, we provide evidence that, as Hartsuiker, Bernolet, Schoonbaert, Speybroeck, & Vanderelst (2008) suggest, syntactic priming decays relatively slowly but the lexical overlap decays quickly. We also show that, when there is lexical overlap between the prime and target, syntactic priming is very strong in a speaker's second language–much stronger than any observed priming within a first language. This collection of priming results, analyzed together for the first time, yields the strongest support yet to claims that priming is an abstract process largely independent of modality or task.

Meta-analysis

Method

For our main meta-analysis, we exhaustively searched for a set of papers on syntactic priming in production. We then extracted the measures of effect size along with details of the experimental set-up. Finally, we performed several regressions to assess (a) the size of the overall priming effect and (b) how it is affected by variations in the experimental conditions.

Inclusion and exclusion criteria. We included only controlled experiments that were focused on syntactic priming in production, in which prime sentences were designed to elicit participant-generated productions of the same syntactic structure and in which the dependent variable was the production itself (thus excluding studies where the dependent variable is reaction time or some other psychometric measure). We defined "syntactic priming" as priming above the level of the word and as not including priming of inflectional or derivational morphology or metrical structure. For that reason, we will refer to the phenomenon mostly as "syntactic priming" (as opposed to "structural priming" or "structural persistence"). While there are many arguments that could be made as to what constitutes a *syntactic* alternation as opposed to a lexical, semantic or pragmatic one, we restrict our investigation to pairs of materials with different word order but close to the same meaning. Classic syntactic alternations are the active/passive alternation ("The boy chased the ball." vs. "The ball was chased by the boy.") and the dative alternation ("The man gave the boy the ball." vs. "The man gave the ball to the boy."). We also include constructions that differ only in the simple context-free rules that would be used to generate them (such as a complex noun phrase vs. a simple noun phrase, as in Bunger et al. (2013)). We excluded experiments like Vernice, Pickering, & Hartsuiker (2012) that involved the priming of thematic roles because this priming did not prime a word order. While studies like these likely tap into syntax, we chose to be narrow in our inclusions in order to have a more homogeneous sample.

We further constrained our sample to experiments with healthy adult participants. When a study focuses on a non-healthy or child population but also presents data from a control group of healthy adult participants, we included the control group in our sample. We required that the results be published in English-language, peer-reviewed journals (not including conference proceedings or dissertations) in 2013 or earlier. Finally, we only considered papers where the proportion of productions matching the primed structures was included as a dependent measure.

Our criteria exclude some studies which are sometimes classified as syntactic priming. Specifically, we chose not to include comprehension priming studies, which includes any study where the dependent measure is not a linguistic production but a measure of how a priming manipulation affects participants' comprehension of sentences (see Tooley & Traxler (2010) for a review of the extensive literature on comprehension priming). We excluded recall studies (e.g., Potter & Lombardi (1998) in which participants are asked to recall a memorized sentence; the dependent measure is whether they make errors) because these studies do not involve free sentence production, and there is a correct answer on each trial. We similarly excluded cross-linguistic priming studies in which the task is to directly translate a sentence from one language to the other. We also excluded studies that were not strictly controlled–including "syntactic alignment" studies whereby the dependent measure is how well the use of a particular structure X predicts the use of that structure at a later time in free-form conversation. While alignment of this nature is arguably a subset of priming, it is beyond the scope of this meta-analysis.

These criteria were applied to 2,096 records returned during the search process, resulting in 73 records.

Moderator analysis. In addition to the main priming effect, studies often investigate other questions about the mechanisms underlying structural priming. Because it is not possible to model all possible differences among experiments, meta-analysis requires choosing a number of experimental variables to consider as *moderators* of the priming effect. For instance, a body of literature has investigated whether there are medium-to-long term effects of structural priming; in this case the moderator is temporal lag. There are a variety of such moderator variables, and we will include some of these variables as predictors in the meta-analysis.

We extracted information from each paper for the moderators listed below. Each bullet point is a particular variable, and each sub-bullet point represents possible values for that variable. For some variables, we recorded more detail but collapsed them into the possible values shown below.

- Language

- Construction type

  - active/passive: "The boy kicked the ball." vs. "The boy was kicked by the ball." See, e.g., Bock (1986).
  - dative: "The girl gave the boy a ball." vs. "The girl gave a ball to the boy." See, e.g., Bock (1986).
  - genitive: "The man's car" vs. "The car of the man." See, e.g., Bernolet, Hartsuiker, & Pickering (2012).
  - transitive/intransitive: "The man was driving." vs. "The man was driving the car." See, e.g., R. P. van Gompel, Arai, & Pearson (2012).

  - locative inversion: “A cat lies on the table.” vs. “On the table lies a cat.” See, e.g., Hartsuiker (1999).
  - modifier order (preferred vs. dispreferred): “The big, red chair” vs. “The red, big chair.” See, e.g., Goudbeek & Krahmer (2012).
  - NP modifier type (adjective vs. relative clause): “The red book” vs. “The book that’s red.” See, e.g., Cleland (2003).
  - Relative Clause attachment (high or low): Relative clause attachment, as in “The men with the kids who plays the piano” vs. “The men with the kids who play the piano.” See, e.g., Scheepers (2003).
  - verb-participle order: “De man belde de politie omdat zijn portemonnee was gestolen/The man called the police, because his wallet was stolen.” vs. “De man belde de politie omdat zijn portemonnee gestolen was/The man called the police, because his wallet stolen was.” (from Hartsuiker & Westenberg (2000)).
  - VP syntax: “The woman entered a cave.” vs. “The woman drove into a cave.” See, e.g., Bunger et al. (2013).
  - complex NP: “The man in the car. . . ” vs. “The man. . . .” See, e.g., Bunger et al. (2013).

- Temporal lag between the prime and the target.

  - none: Target appears after prime with no intervening linguistic material (there can be a fixation screen).
  - filler: Some number of filler items appear between the prime and target.
  - cumulative block priming: In this category, we included studies in which the prime does not proceed the target in an alternating fashion, but rather the priming round occurs followed by a target round (i.e. all the primes occur, then all the targets). Many of the Kaschak, et al. studies (e.g., Kaschak et al. (2011), Kaschak, Loney, & Borreggine (2006)) fall into this category.
  - cumulative block priming (long): A priming round occurs more than 10 minutes before a target round (e.g., Kaschak (2007)). We chose to separate this category from other types of cumulative priming since we hypothesized that long delays (including several days in some studies) could have a qualitatively different effect than priming on the scale of seconds or minutes.

- Bilingualism

  - L1 → L1 (priming within first language)
  - L2 → L2 (priming within second language)
  - L1 → L2 priming (cross-linguistic with first-language prime, second-language target)
  - L2 → L1 priming (cross-linguistic with second-language prime, first-language traget)

- Lexical overlap between the prime and target
  - No repeat of critical words between prime and target.
  - Yes otherwise (if word is repeated, if semantically related word is repeated, if translated version of word is repeated, etc.).
- Year of publication
- Target task
  - Picture description (participant orally describes a picture).
  - Written sentence completion (participant completes a preamble like "The boy gave. . . ").
  - Auditory sentence completion (participant speaks the completion to a preamble like "The boy gave. . . ").
  - Sentence from words (participant is given target words and told to assemble them into a sentence).
- Modality of prime
  - Auditory prime (including both recordings as well as spoken primes by a "live" interlocutor).
  - Visually presented prime.
  - Prime is visually presented but read aloud by the participant.
- Whether the prime is repeated by the participant
  - Yes (includes cases where the prime is delivered auditorily and then repeated as well as cases where the participant self-primes by being required to complete a prime sentence in a particular way)
  - No
- Confederate
  - Yes if a second person is using structures intended to prime the participant, but the participant is not aware
  - No otherwise

Note that many studies manipulated variables in addition to the moderators listed here, but including those was beyond the scope of this study.

Search strategies. The literature search was conducted using three primary methods: recording references listed in relevant review papers (Ferreira & Bock, 2006; Heydel & Murray, 2000; Pickering & Ferreira, 2008); searching for records which cite relevant work (Bock, 1986; Ferreira & Bock, 2006; Pickering & Ferreira, 2008); and searching ProQuest, Scopus, and Web of Science databases using natural language terms and controlled vocabulary. The third method was by far the most exhaustive, identifying 71 of the 73 papers included in the final list. The remaining two were found through the forward citation method (Goudbeek & Krahmer, 2012; Kantola & Gompel, 2011). The first literature retrieval effort was conducted in July 2013. This yielded 70 of the total 73 records in the final list. A second retrieval was conducted in June 2015 in order to include records that had been updated since the first retrieval, up to the end of 2013. As a result, the final list of 73 includes all recovered records with journal publication dates prior to 2014 (2013 and earlier). Three additional papers were included in our initial analyses, but are not included in the final list of 73 papers: Biria, Ameri-Golestan, & Antón-Méndez (2010) which is about indirect questions and not strictly syntactic; Kootstra, Hell, & Dijkstra (2010) in which priming is only indirectly tested by comparing different experiments; and Shin & Christianson (2012) in which priming is used as a teaching aid making it not directly comparable to other production studies. An additional two papers in our final sample of 73 were added after a reviewer noticed the omissions; these two papers all appeared in the initial literature search and were erroneously excluded. For more details on the search procedure, see the supplementary material.

Coding procedures. From each experiment within each paper, one of the authors extracted the number of unique subjects, the number of unique items, and the number of items per subject per experimental cell, along with the population characteristics and task characteristics needed for the moderator analysis. This coding was subsequently re-checked by K.M. For each experimental condition (e.g. verb repeated vs. not repeated), we extracted the mean proportion of productions matching the primed structure (e.g. prepositional-object dative; PO) and the alternative structure (e.g. double-object dative; DO). For a small number of papers, this information was not available. In these cases, the raw data (obtained from the original study authors) was used to obtain the estimates.

For every paper in the initial sample except the two added after the first round of peer review, one or more of the original authors was contacted by e-mail and asked for the raw data. Each author was contacted at least twice. For 45 of the 71 original papers, we received the raw data from the authors. For 25 of the 71 papers, the authors responded that the data was unavailable either because it was lost, corrupted, or otherwise inaccessible. For only 1 paper, we received no response from the authors.

Because information on "other" responses is not always available (both in raw data and in published estimates), our analysis excluded Other responses whenever possible. Thus, for all studies, the proportion of X responses (e.g., DO) and the proportion of Y responses (e.g., PO) add up to 1.

Statistical methods. To make meaningful comparisons across different studies, we need a uniform notion of "effect size" (J. Cohen, 1992; Lipsey & Wilson, 2001). For our purposes, we want the effect size to answer the question "how big is the effect of syntactic priming." To that end, the effect size measure we use is the log odds ratio of the prime condition compared to the no-prime condition (see Equation 1). That is, if the proportion of trials using the passive is .34 after a passive prime and .20 after an active prime, the log

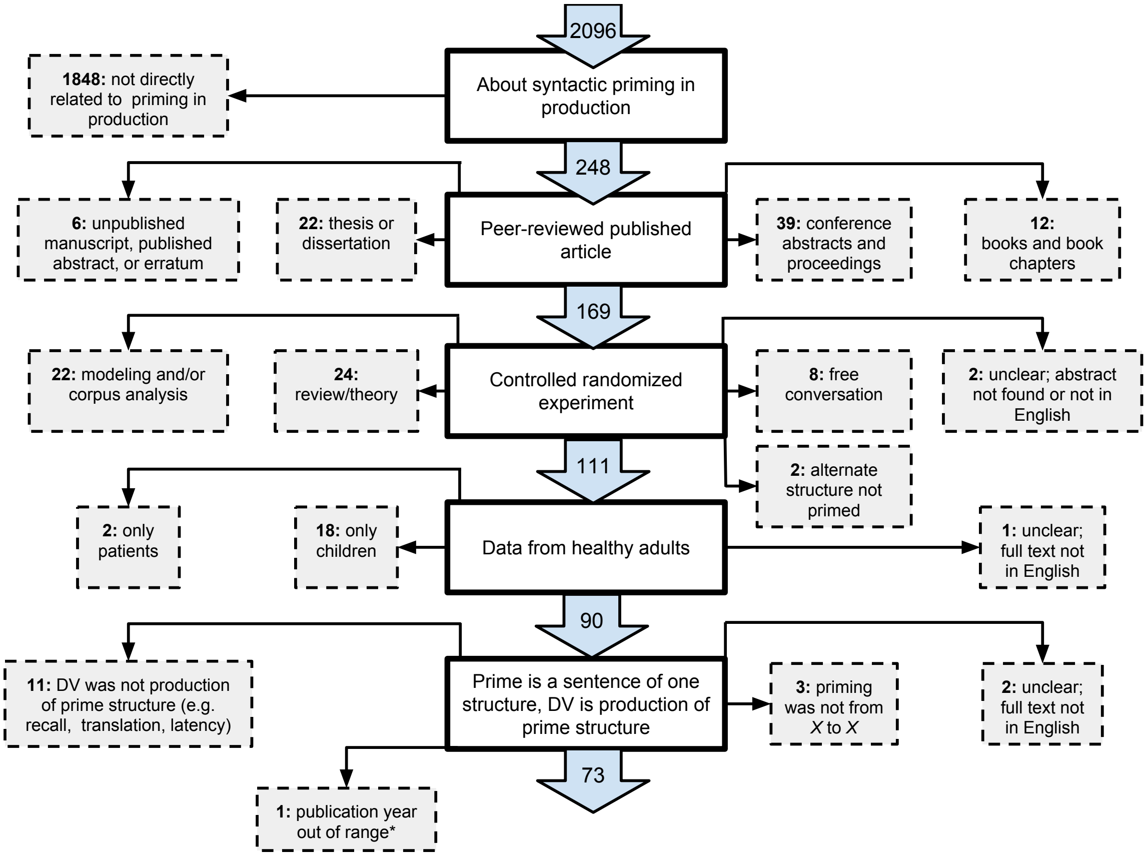

2096
About syntactic priming in production
1848: not directly related to priming in production
248
Peer-reviewed published article
6: unpublished manuscript, published abstract, or erratum
22: thesis or dissertation
39: conference abstracts and proceedings
12: books and book chapters
169
Controlled randomized experiment
22: modeling and/or corpus analysis
24: review/theory
8: free conversation
2: unclear; abstract not found or not in English
2: alternate structure not primed
111
Data from healthy adults
2: only patients
18: only children
1: unclear; full text not in English
90
Prime is a sentence of one structure, DV is production of prime structure
11: DV was not production of prime structure (e.g. recall, translation, latency)
3: priming was not from X to X
2: unclear; full text not in English
1: publication year out of range*
73

Figure 1. Flowchart showing literature search.

odds ratio would be $\log \frac{.34}{1-.34} - \log \frac{.20}{1-.20} = 0.72$.[1]

(1) $\text{LogOddsRatio} = \log(\frac{p(X|\text{Prime})}{1-p(X|\text{Prime})})$ - $\log(\frac{p(X|\text{NoPrime})}{1-p(X|\text{NoPrime})})$.

In a meta-analysis, we do not want to give each study equal weight. Rather, studies which have smaller standard error (perhaps because of more subjects and items) should be weighted more. In order to know how much to weight each study in the meta-analysis, we need the standard error. We computed the standard error on the log odds ratio using the formula for standard error on a log odds ratio:

(2) $SE = \sqrt{\frac{1}{n_{\text{Prime}X}} + \frac{1}{n_{\text{NoPrime}X}} + \frac{1}{n_{\text{Prime}Y}} + \frac{1}{n_{\text{NoPrime}Y}}}$,

where $n_{\text{Prime}X}$ is the number of individual data points for which Structure X is primed and Structure X is used in the target, $n_{\text{NoPrime}X}$ is the number of data points for which Structure Y is primed and Structure X is used in the target, $n_{\text{Prime}Y}$ is the number of data points where Structure Y is primed and Structure Y is used in the target, and $n_{\text{NoPrime}Y}$

[1] We use the log odds ratio instead of the odds ratio because it has better distributional properties, but it can easily be converted to an odds ratio by exponentiating.

is the number of data points where Structure X is primed and Structure Y is used in the target.

When there was a baseline condition in addition to two prime conditions (i.e. DO prime, PO prime, and baseline), we ignored the baseline condition for our main meta-analysis (although it could be used in the p-curve analysis). We excluded studies for which either of the condition means was above .98 or below .02 or in which both condition means were either above .90 or below .10, since the log odds ratio is inflated near 0 and 1. In addition to being problematic for the quantitative analysis, we do not believe that these studies are directly comparable to the studies in which participants are less categorical in their use of particular constructions. Results from 43 experimental conditions of an original 386 were excluded for this reason, including all the data from 3 papers. We also excluded Coyle & Kaschak (2012) since item and prime condition are confounded in that experiment, and thus it is not possible to obtain an estimate of the size of the priming effect.

These methods for computing log odds and standard error are not necessarily optimal given the within-subject, within-item designs common in psycholinguistics. The current standard for estimating the effect size and standard error in categorical data like this, given the latest statistical thought (Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015; Jaeger, 2008), is to extract size and its associated standard error from a linear mixed effect logistic regression with random effects for subject and item. But this information is not always available in published reports for a variety of reasons. First, before 2008, researchers typically reported ANOVAs instead of mixed effect regressions and the reported results of those ANOVAs are not usually sufficient for computing standard error that is comparable to a standard error from a mixed effect regression. Moreover, in both mixed effect regressions and ANOVAs, there is variability as to how the random effects are structured across papers that may make them incomparable. And often the hypothesis being tested is not just whether priming exists but about some other variable–in which case sufficient test statistics for the actual priming manipulation may or may not be reported.

Thus, we believe that the best ways to ensure that estimates are consistent across papers is to a) use raw trial-level data (which we obtained for a subset of papers) and analyze all experiments together in one model and b) use the published means and design characteristics (which are almost always available) to compute the log odds ratio and the standard error. We used both of these techniques and, as we report below, found similar results using each.

Results

Characteristics of the remaining studies. From our initial 73 papers, we analyzed a total of 343 data points (i.e., experimental conditions) from 138 experiments from 69 papers. The median number of participants per condition was 32. The median number of items seen by each participant in each experimental cell was 6. The full list of included studies is provided in the appendix. Summary statistics (unweighted mean effect size and number of data points per moderator) can be found in Figures 2 and 3.

Weighted mean results. To facilitate meaningful comparisons across studies, we computed a weighted mean effect size as described in Lipsey & Wilson (2001). The clearest and most consistent moderator of the size of the priming effect was lexical overlap between the prime and target (i.e., whether the same word, a semantically or a phonologically related

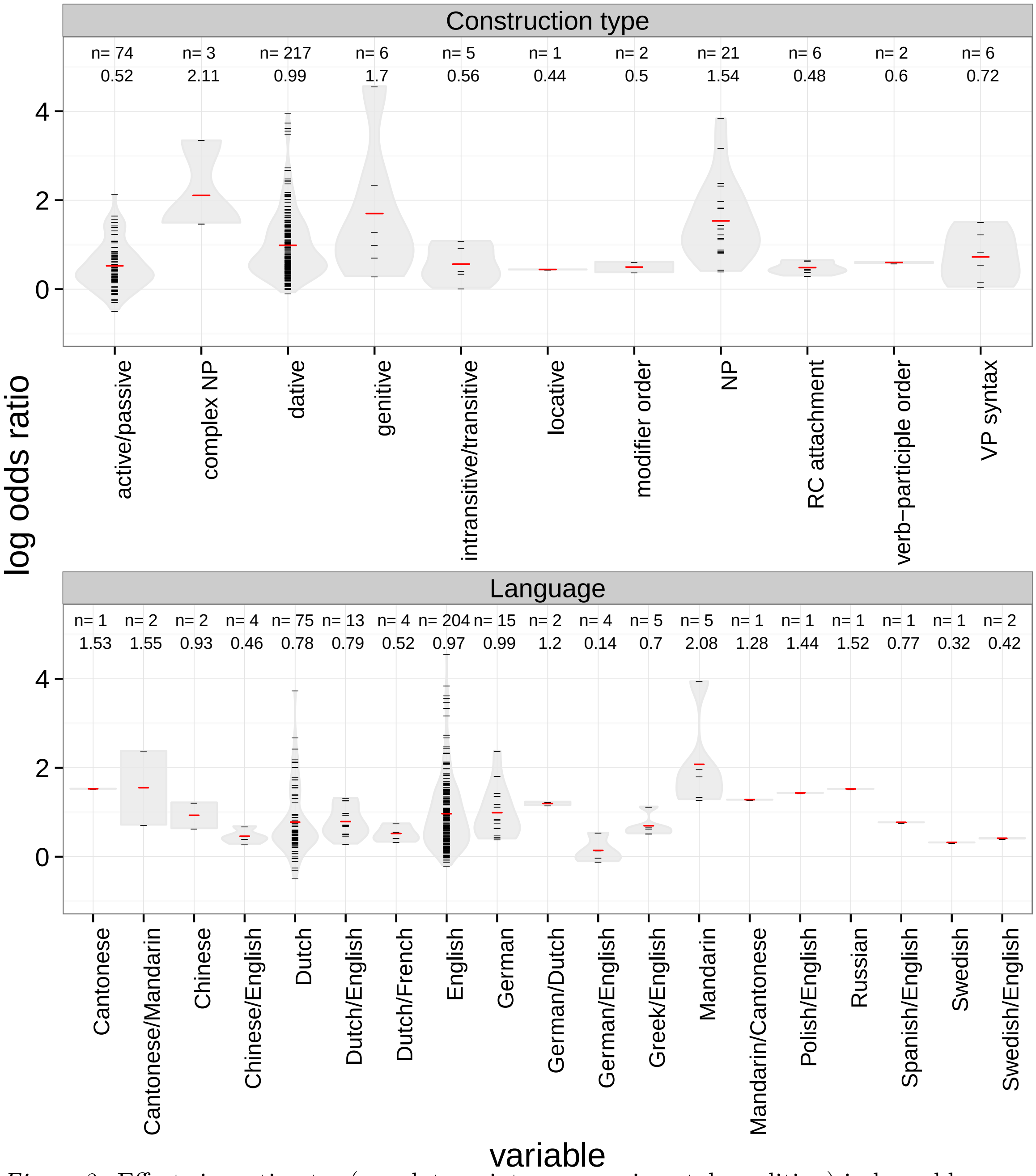


Figure 2. Effect size estimates (one data point per experimental condition) in log odds space by language and construction type are represented by the individual horizontal lines and are not weighted by sample size or standard error. The horizontal red line represents the mean, and the gray blobs represent smoothed density estimates such that fatter parts of the blob represent more likely value.

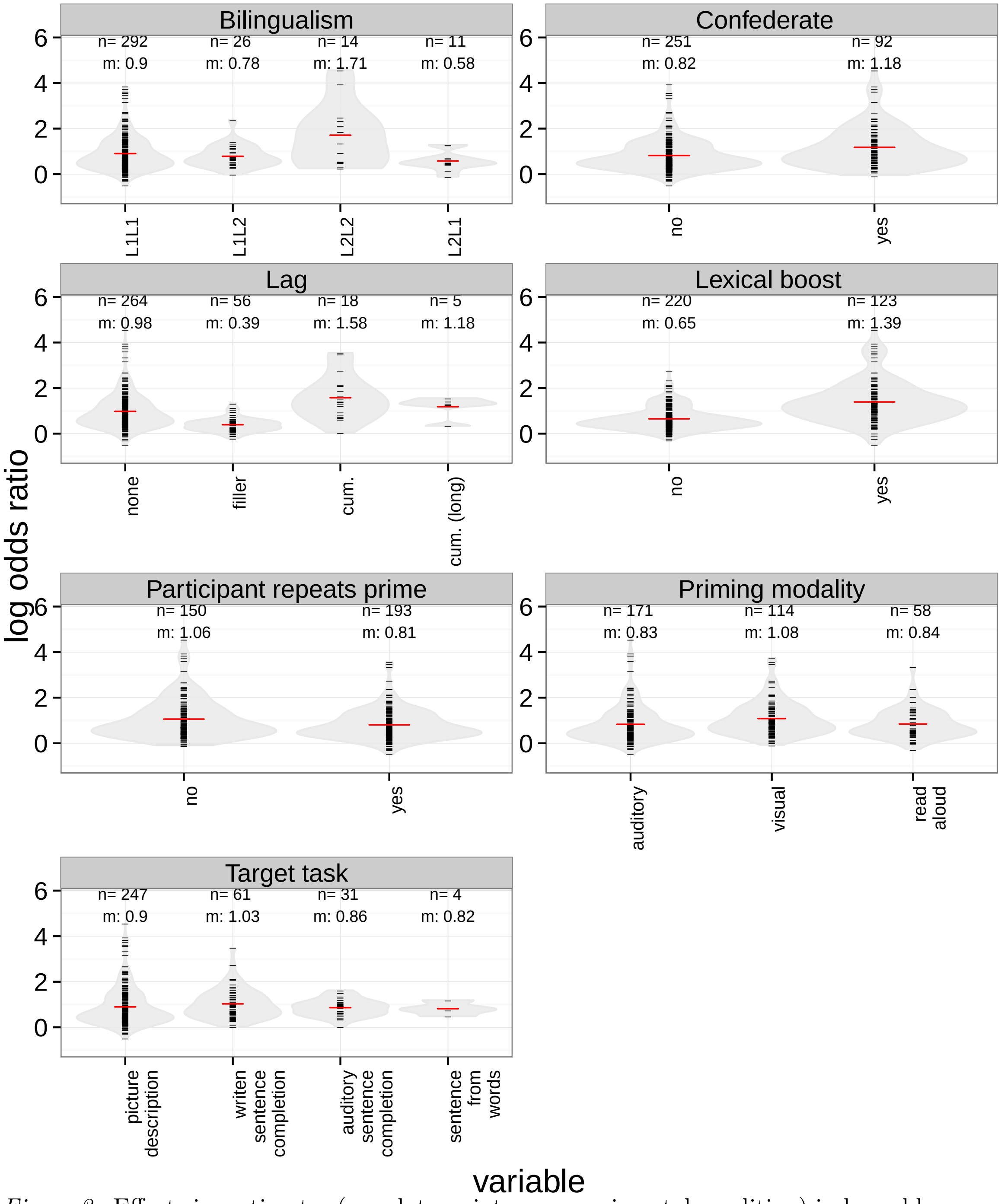


Figure 3. Effect size estimates (one data point per experimental condition) in log odds space by moderator are represented by the individual horizontal lines and are not weighted by sample size or standard error. The horizontal red line represents the mean, and the gray blobs represent smoothed density estimates such that fatter parts of the blob represent more likely value.

word, or a translation-equivalent word was repeated in the prime and the target). For 220 studies with no lexical overlap, the weighted mean odds ratio was 1.67 with a 95% CI of [1.63, 1.72], $p < .0001$, such that the odds of a construction occurring are 1.67 times greater when it is primed than when it is not primed. This means that, if a construction occurs 50% of the time when it is not primed, it would occur 63% of the time when primed. Multiplying the log odds ratio by $\frac{\sqrt{3}}{\pi}$ to convert it to an estimate of a Cohen's d standardized effect size (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hasselblad & Hedges, 1995), estimated d is 0.28 (a small-to-medium-sized effect per Cohen's original rubric (J. Cohen, 1977)).

There were 123 studies with lexical overlap. The weighted mean odds ratio was 3.26 with a 95% CI of [3.13, 3.40], $p < .0001$, such that the odds of a construction occurring are 3.26 times greater when it is primed with lexical overlap than when it is not primed. If a construction occurred 50% of the time when not primed, it would occur 77% of the time when primed with lexical overlap. Converting the odds ratio to Cohen's d, we estimate an effect size of d = 0.65 (a medium-to-large-sized effect).

We show means and 95% CIs (in log odds space) for the studies and conditions in Figures 4-6. These means and standard errors are based on the published estimates for each paper.

A simple weighted mean does not account for additional structure in the data, such as the correlations between conditions of the same experiment, the correlation between experiments in the same paper, and the various moderators of syntactic priming that are manipulated within and across experiments. Therefore, we next present results from a mixed effect meta-regression.

Model results. We first fit a random effect, intercept-only meta-analysis model to all data points and found a significant intercept of 0.87 [95% CI 0.80, 0.94], $p < .0001$. This indicates a significant effect of syntactic priming in our sample. Since these studies sometimes include very different experimental conditions, there was unsurprisingly significant heterogeneity in this estimate, as measured by a Q-test comparing the variability among effect size estimates to the expected sampling variability: Q(342)=2748.60, $p < .0001$.

None of the estimates reported above account for moderators of syntactic priming. For instance, perhaps priming exists for certain constructions but not others. Using the `metafor` package (Viechtbauer, 2010) in `R` (R Core Team, 2015), we fit a mixed effect meta-analysis regression to the data. The mixed effect meta-analysis differs from a standard mixed effect model in that the standard error of each data point is assumed to be known instead of estimating it from the data. But the underlying logic is the same in that we are asking what underlying parameter values could plausibly give rise to the observed effect sizes obtained in the published studies. Here, each individual data point is an effect size (change in log odds ratio) extracted from an experimental condition with an associated standard error (where standard error is estimated as described above, using the number of subjects, items, and the condition means).

In this meta-analysis regression, the intercept is the size of the priming effect. We included fixed effects of lag, year, lexical overlap between prime and target, within/between language condition, target task, mode of presentation, whether the participant repeated the prime, and whether there was a confederate. Since it has been widely posited in the literature that lexical overlap interacts with temporal lag and that lexical overlap exerts a stronger effect on priming in L2 populations, we included interactions between lexical overlap

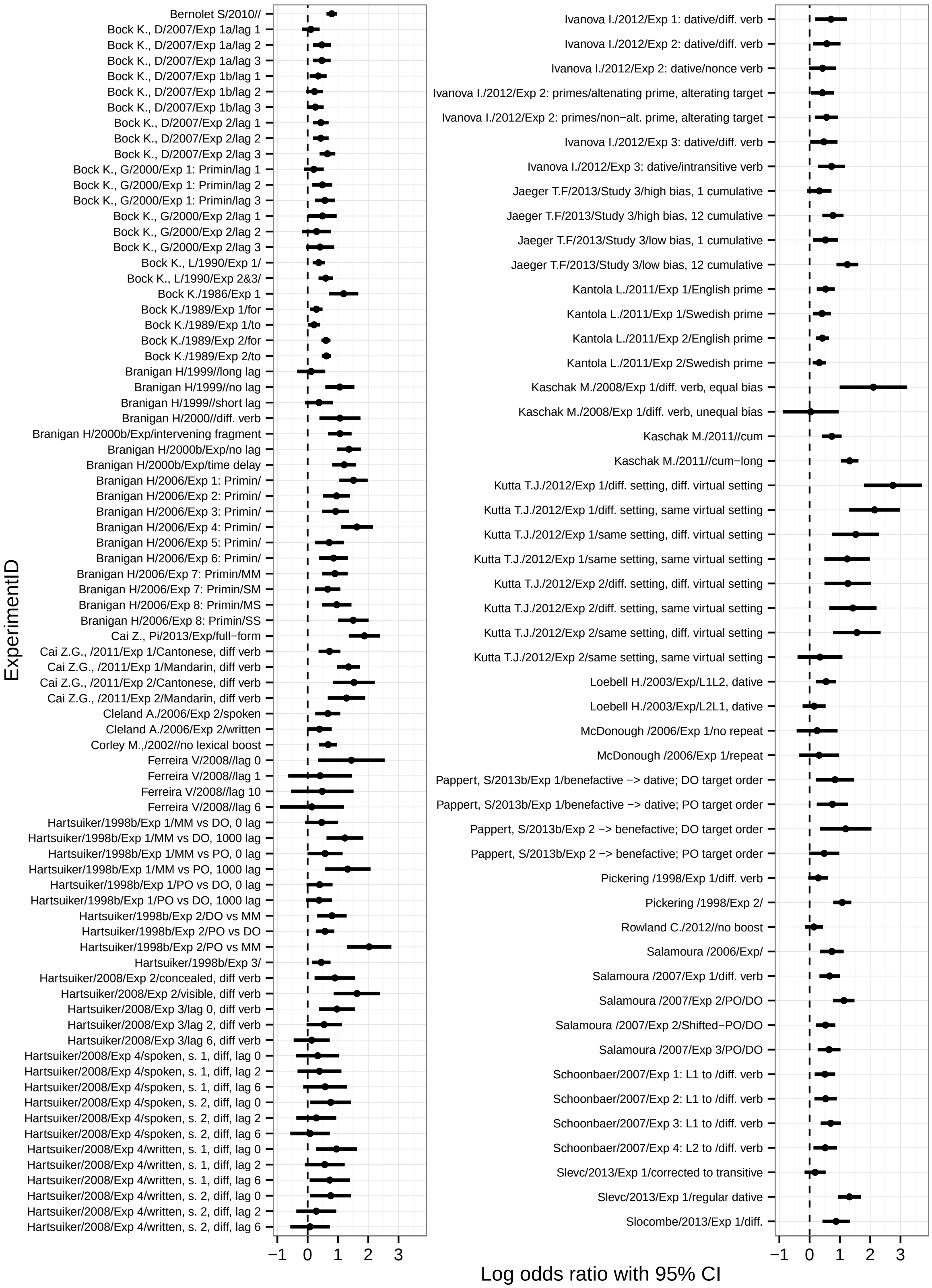


Figure 4. Forest plot with 95% CIs for dative studies with no lexical overlap in log odds space.

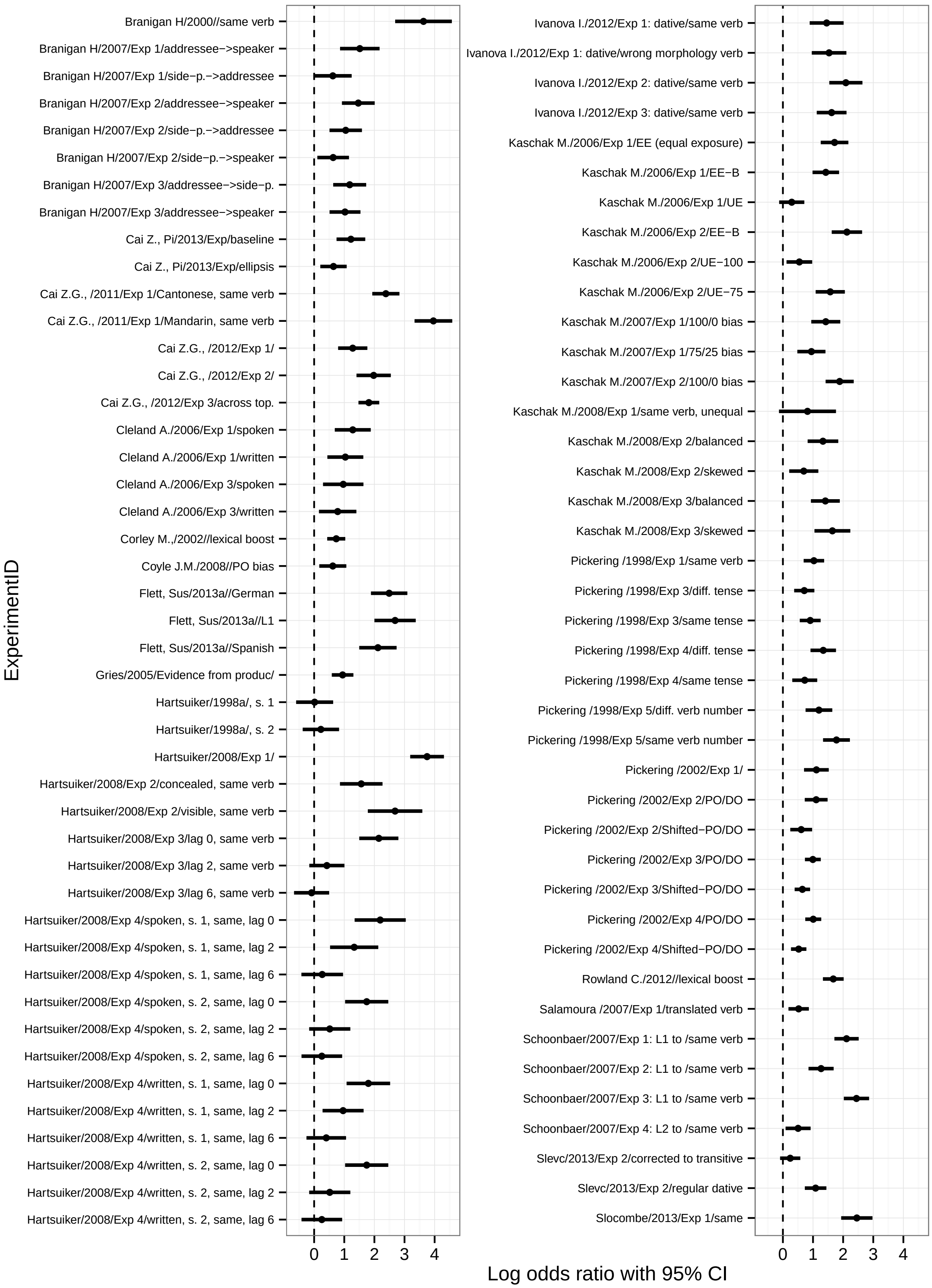


Figure 5. Forest plot with 95% CIs for dative studies with lexical overlap in log odds space.

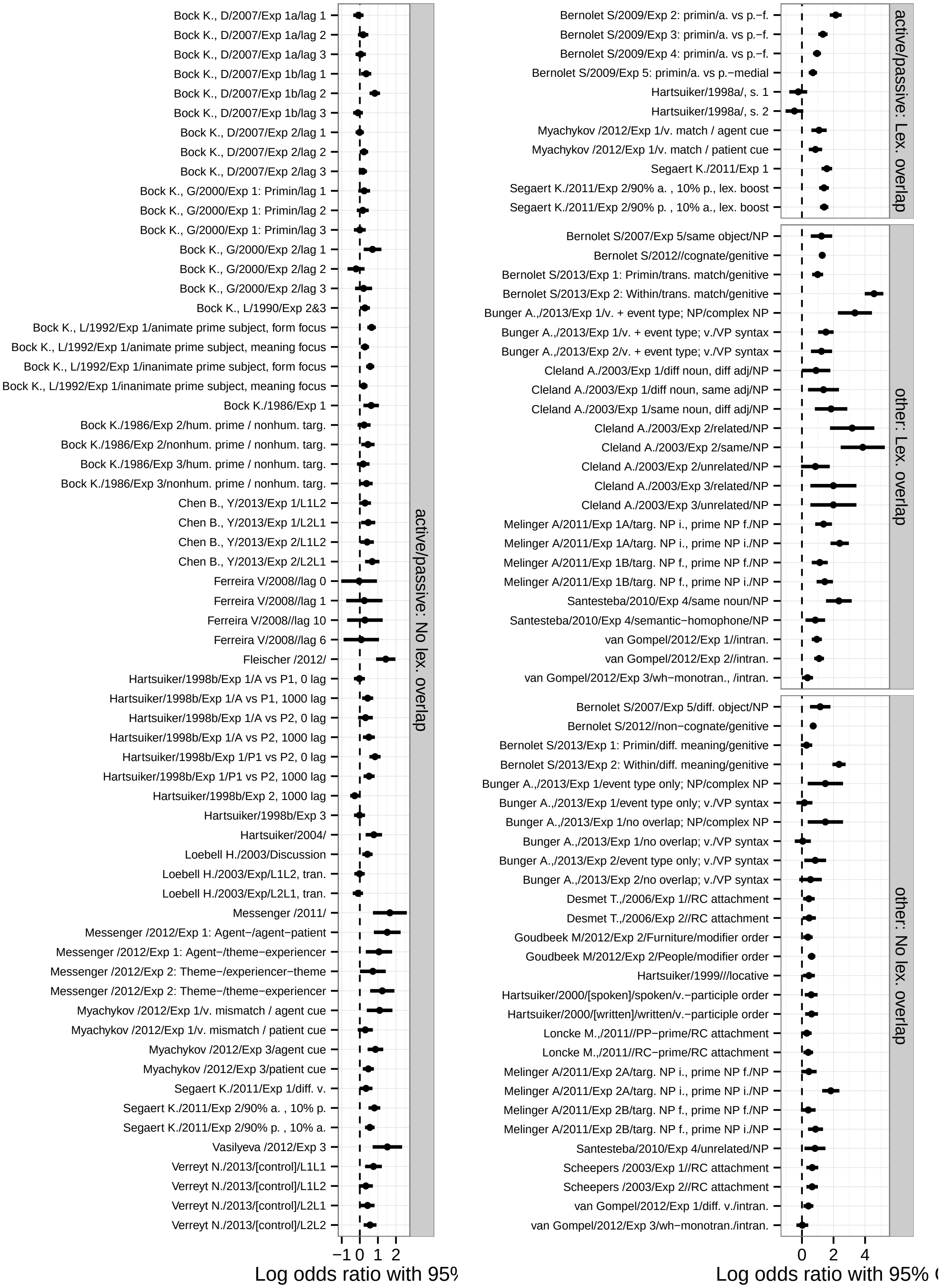


Figure 6. Forest plot for active/passive and other studies with and without boost.

and lag and between lexical overlap and bilingualism. To avoid overparameterizing the model, we did not include further interactions–especially since few papers vary by more than one factor and thus the interaction terms would be ill-defined. We included random effects for construction type (dative, active/passive, etc.), paper, experiment (nested within paper), and condition (nested within paper and experiment). We did not include random slopes since none of these grouping factors consistently vary by anything but prime condition.

Across 69 papers that survived all exclusions, consisting of 138 experiments and 343 unique conditions, we found a significant baseline priming effect (no lexical overlap, no lag) corresponding to a change in odds ratio of 1.68 [95% CI 1.25, 2.27]; Cohen's d = 0.29 when there is no lexical overlap between prime and target and 3.67 [95% CI 2.53, 5.31]; Cohen's d = 0.72 when there is lexical overlap between prime and target. Thus, as with the simple weighted mean analysis, the model suggests that the size of the priming effect is small-to-medium without lexical overlap between prime and target and medium-to-large with lexical overlap.

Moderators. The moderators are shown with their estimates and 95% CI's in log-odds space in Figure 7. Terms significant in either analysis at $p < .05$ are starred, with two stars for $p < .01$ and three stars for $p < .001$. The intercept in this model is a priming study with no lexical boost, no lag, no confederate, an auditory picture description task with an auditory prime that is not repeated by the participant, performed in the year 2000. Besides the main effect of priming, the coefficients in Figure 7 represent the change in log odds ratio associated with adding that moderator. As a rough rule of thumb, Gelman & Hill (2006) suggest dividing by 4 to convert the log odds coefficients to changes in probability space.

The presence of the moderators significantly reduces the heterogeneity in the data: QM(20) = 230.01, $p < .0001$. Even with the moderators included in the model, though, there is still significantly more variance than expected by sampling variability alone: Q(305) = 1490.53.

Lexical overlap between prime and target. Lexical overlap is the most consistent moderator of syntactic priming (this is the "leixcal boost" effect first demonstrated in Pickering & Branigan (1998)). It significantly enhances the priming effect (beta=0.76, z=9.9, p<.0001), and the effect of lexical overlap is actually stronger than the priming effect itself (i.e., the change in participant response tendency between priming without lexical overlap and priming with lexical overlap is greater than the change from no priming at all to priming). See Pickering & Ferreira (2008) for a clear summary of the many reasons that have been proposed for the strength of the lexical overlap effect.

Modality of prime. Mode of prime was analyzed in three separate coefficients: modality of prime, whether the participant generates the prime (either by having to produce it herself or by simply rotely repeating it), and whether there was a confederate. The baseline here was an auditorily presented prime that was not repeated, with no confederate. We found no clear effect of modality. Although we did not test cross-modal priming since only a small number of papers in our sample studied it, we take this to be broadly consistent with findings in the literature (Cleland & Pickering, 2006; Hartsuiker & Westenberg, 2000) that the modality of the prime does not strongly affect the size of the syntactic priming effect. There was also no clear effect of having a confederate.

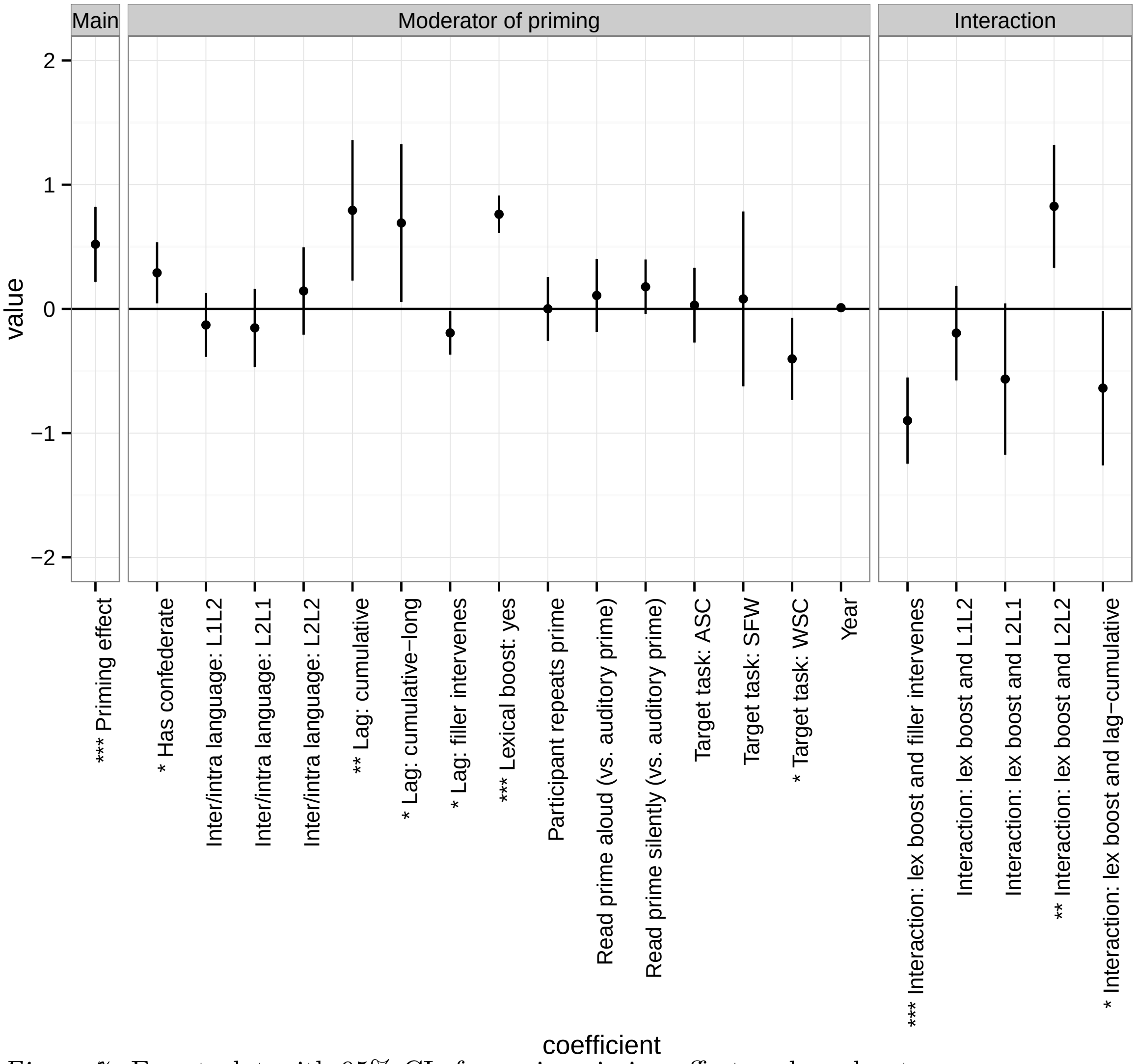


Figure 7. Forest plot with 95% CIs for main priming effect and moderators

Target task. In our model, the baseline target task is the classic picture description task (Bock, 1986). Consistent with Hartsuiker & Westenberg (2000), relative to picture description, auditory sentence completion (Branigan, Pickering, Stewart, & McLean, 2000; Hartsuiker & Westenberg, 2000) produced similar-sized priming effects, as did the task where participants are asked to generate sentences from a list of words (although only one paper, Pappert & Pechmann (2013), used that strategy so little can be concluded about it besides what is concluded there). We found a marginally significant negative effect for written sentence completion (Pickering & Branigan, 1998), such that written sentence completion results in less priming than picture description. Every written sentence completion task in our sample, however, does find a numerically positive priming effect.

Bilingualism. We consider three types of bilingual priming: priming within a second language, priming from a first language to a second language, and priming from a second language to a first language. We did not find significant main effects for priming to be stronger or weaker for bilingual priming relative to classic priming within a native language, although cross-linguistic priming was numerically weaker than within-language priming. This is consistent with Pickering & Ferreira (2008), who suggest that the effect of cross-linguistic priming is similar to that of L1-L1 priming. The trend towards priming being weaker cross-linguistically may be, in part, driven by the inclusion of cross-linguistic priming effects in which the languages used have different word orders. For instance, Bernolet et al. (2007) found little to no priming between Dutch and English for complex noun phrases, possibly since the structures are so different as to be represented differently.

Importantly, we found that, relative to L1→L1 priming, there was a strong enhancement of the lexical overlap effect when priming took place in a second language (beta=0.83, z=3.27, p<.01). This is consistent with past results (e.g., Kim & McDonough (2007), Schoonbaert et al. (2007), Bernolet et al. (2013)), which suggest that L2 speakers are highly sensitive to lexical boost and maintain strong item-specific representations. Specifically, Bernolet et al. (2013) found that less proficient speakers are more susceptible to lexical boost effects–perhaps suggesting that less proficient speakers rely on item-specific representations.

We also found a significant trend for the lexical overlap to be smaller for L2→L1 priming than for the L1→L1 baseline (beta=-0.57, z=-1.82, p= 0.07). This is perhaps not surprising since, in the former case, "lexical overlap" typically refers to a translation-equivalent word, whereas in within-language priming it is often an identical word. There was a smaller, not significant trend for L1→L2 priming (beta=-0.19, z=-1, p= 0.32) to also show less of a lexical boost effect than within-language priming. These findings–an asymmetry in lexical boost between within-language and between-language priming and a further assymetry in lexical boost between L1→L2 priming and L2→L1 priming–are consistent with what Schoonbaert et al. (2007) found and used to argue in favor of the lexical-syntactic model described in Hartsuiker, Pickering, & Veltkamp (2004).

Lag. We considered three types of lag: the inclusion of fillers between the prime and target, cumulative priming, and long-term cumulative priming (more than one day between prime and target). The latter two are techniques developed and used mostly by Kaschak and colleagues (e.g., Kaschak (2007), Kaschak & Borreggine (2008)) whereby a group of prime sentences are presented in a block (priming either a construction or a particular construction with a particular verb), which is then followed by a target phase. While we treat this as a type of lag, temporal delay is not the only difference between these paradigms and other syntactic priming paradigms.

The baseline condition here is no temporal lag between prime and target. There has been some debate in the literature as to how much, if at all, temporal lag reduces the priming effect. We found a small but significant negative main effect of including filler material between prime and target (beta=-0.19, z=-2.17, p<.05), such that the priming effect was smaller when there was filler material. We also found that including one or more fillers between the prime and target significantly reduces lexical boost (beta=-0.9, z=-5.08, p<.0001). This reduction of lexical boost is strikingly large relative to the main effect of lexical boost and the main effect of filler, suggesting that the lexical boost effect essentially disappears entirely when there is filler material between prime and target. Hartsuiker et al. (2008) found something similar and argued that it reflects evidence for both a long-term implicit learning account of syntactic priming as well as a short-term lexical component.

There was a strong main effect for priming to increase relative to the no-lag standard priming condition when the paradigm uses cumulative priming (beta=0.79, z=2.75, p<.01) or long-term cumulative priming (beta=0.69, z=2.13, p<.05). But, as with filler lag, the effect of lexical overlap is reduced using this paradigm (beta=-0.64, z=-2.01, p<.05). Note that cumulative lag, in our meta-analysis, is largely confounded with primes that are presented in blocks as opposed to as single sentences. So the increased effect of priming using this paradigm, as seen in the large main effect, is plausibly the result of the paradigm as a whole and not because of the temporal delay between the primes and the target.

Year of publication. Centering year at 2000, we found no effect of year of publication on the size of the priming effect (beta=0.01, z=1.2, p= 0.23).

Validating the model using raw data. To test whether the meta-analysis regression described above is appropriate and whether the effect sizes and standard errors used are good estimates of the data, we fit a mixed effect logistic regression to the 71194 trial-level data points from the 43 non-excluded papers for which we have raw data.

For each experiment, we set the dependent variable to be the less frequent of the two syntactic constructions and predicted the dependent variable from fixed effects of prime condition and the interactions of prime condition with the same fixed effects as for the meta-regression above, with the exception that the intercept here is the response variable and thus there is a main effect of prime condition. In this regression, moderators are represented not as main effects but as interactions with prime condition. But the underlying logic is the same, and we therefore treated these interactions straightforwardly as moderators of the priming effect. We included random effects for construction type (dative, active/passive, etc.), paper, experiment (nested within paper), condition (nested within paper and experiment), unique subject, and unique item. Subject and item were necessarily nested within experiment, and although it is likely that some subjects participated in multiple experiments and even more likely that some items were re-used across experiments, we do not have data to model this. Following Bates et al. (2015) and using the `RePsychLing` package, we first fit the maximal justified model and then simplified it to obtain convergence and avoid overparameterization. We found that including the correlation parameters did not significantly improve the model. We included random slopes for prime condition by subject, item, experiment, and condition. Other random slopes did not significantly improve model fit or led to lack of convergence.

We compared the results of this regression to a meta-regression on the same set of data (i.e. just the subset of experiments for which we have raw data). We found that the results are qualitatively and quantitatively similar to those obtained using the mixed effect model on the raw data and that, for most parameters, the 95% confidence interval includes the point estimate from the other method. Figure 8 shows the fixed effect coefficients with 95% CIs from the model for this data from the raw data regression (in gray) and the meta-analysis regression (in black). We thus believe that the meta-analysis technique used here, based off published estimates, is sufficient to give similar results as having all the published data.

Having said that, the meta-analysis regression using all the data gives an odds ratio (1.68) that is higher than the margin of error of the same regression run on just the subset of studies for which we have raw data (1.46). While it is possible there are systematic differences between the studies for which we were able to obtain raw data and studies for which we were not, the estimates are similar enough that the difference is possibly just noise.

Interim conclusions. Using the published estimates of the priming effect across 69 papers, we found a robust effect of syntactic priming that becomes dramatically larger when

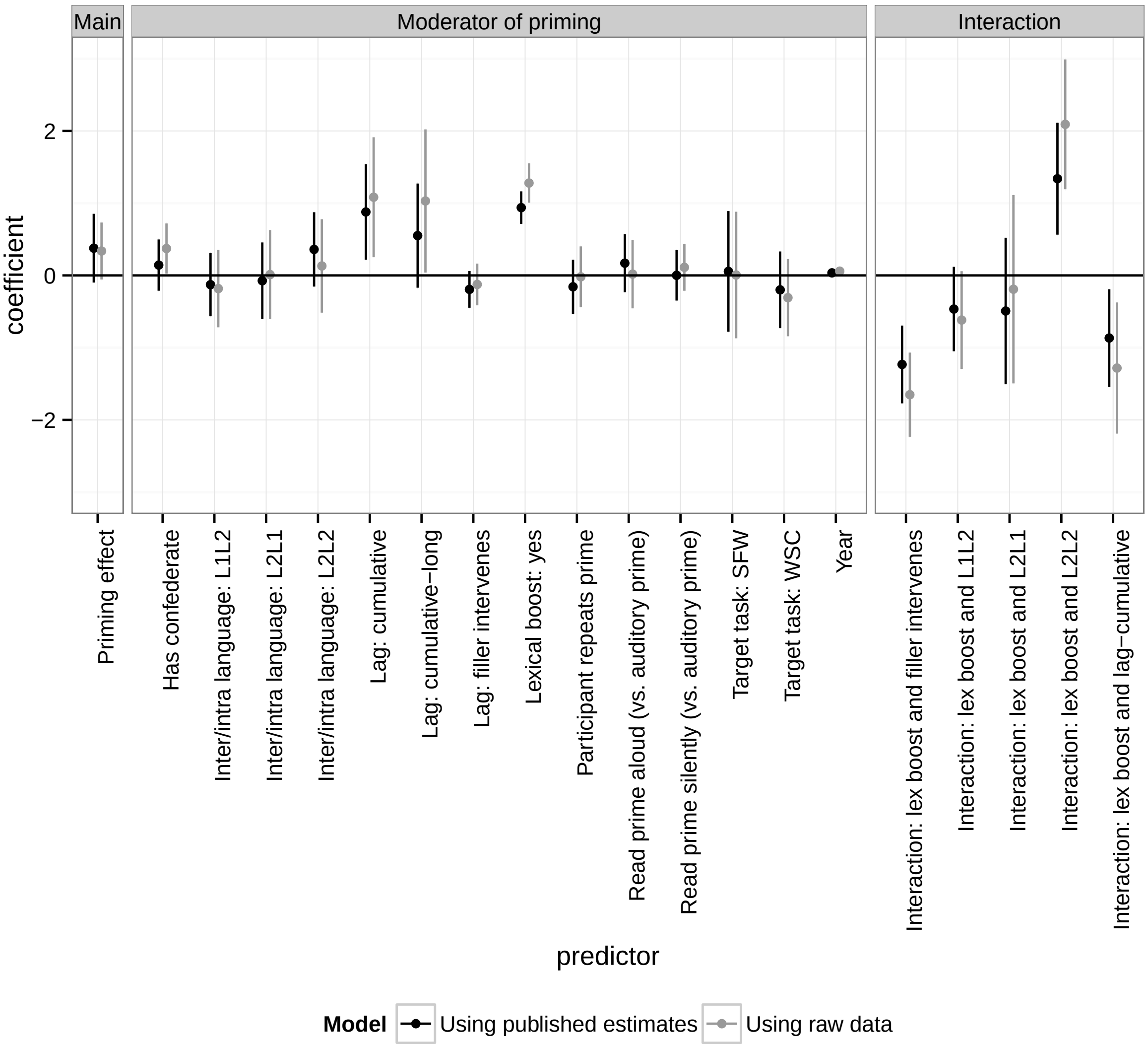


Figure 8. Forest plot with 95% CIs for main priming effect and moderators

there is lexical boost. We also found effects of temporal lag and bilingualism. We validated the model on a subset of papers for which we have raw data and found similar results.

In any meta-analysis, though, the conclusions that can be drawn are only as sound as the data that goes into it. In our case, our meta-analysis was restricted to only published studies in peer-reviewed journals.[2] As a result, to assess the validity of our results, we must ask whether the sample suffers from low statistical power, publication bias, or "p-hacking". If the only articles that are accepted to journals are ones that include significant results or if study authors performed multiple analyses and only reported the significant ones, the effect sizes here would be inflated. In the next section, we analyze the distribution of p-curves in the sampled studies in order to evaluate the evidential value of the results

[2] We also solicited unpublished studies, but we received only a handful of responses including studies that were not likely to be published in the next year. Therefore, we did not include these in the current report.

Assessment of publication bias and statistical power

The distribution of p-values used to support or refute the hypotheses of a particular set of studies can be used to test for evidence of publication bias or “p-hacking”" in those studies (Francis, Tanzman, & Matthews, 2014; Simonsohn et al., 2014b, 2014a). Recall that a p-value is the probability of a null hypothesis having generated data as extreme as the data observed. In psychology studies, p-values less than .05 are taken as sufficient evidence to reject the null hypothesis. In studies investigating effects of syntactic priming, the null hypothesis is usually that there is no difference between the two prime conditions. If a study is investigating whether or not there is lexical boost, the null hypothesis would be that lexical overlap between prime and target has no effect on the strength of priming.

If a group of studies is investigating a real, robust effect, the distribution of p-values will be right-skewed such that there are more p-values between 0 and .01 than between .02 and .03, more p-values between .02 and .03 than between .03 and .04, and so on. Just how sharply skewed the “p-curve” (Simonsohn et al., 2014b, 2014a) is will be a function of statistical power: high power leads to more right skew. This is perhaps simplest to think about in the extreme case. With infinite sample size and a true effect (even if that effect is small), the p-values would be arbitrarily small (certainly all less than .01). If a group of studies is investigating a *null* effect, though, the p-values will be uniformly distributed between 0 and 1. Now imagine that only p-values less than .05 were published in journals and so we only have access to p-values less than .05. If the underlying effect is real, we will still see a right-skewed p-curve. If the underlying effect is *not* real (i.e., the distribution of p-values for experiments is uniform between 0 and 1), when we censor all p-values greater than .05, we will be left with a flat distribution of p-values between 0 and 1. Note that, if statistical power is low (i.e., the probability of correctly rejecting the null when the null is false is too small), that could also lead to a flat distribution of p-values. In either case, a flat p-curve would suggest a lack of evidential value in the data. If researchers are actively p-hacking–that is, re-running variants of analyses until a significant result (p<.05) is obtained, then the curve will actually be left-skewed such that there are more p-values between .04 and .05 then between .03 and .04.

The goal of our p-curve analysis is to assess whether the collection of experiments identified in the meta-analysis contain evidential value for the claims they make. The papers in our sample make two fundamentally different types of claims, however. Some argue that syntactic priming exists or does not exist given various experimental conditions. Others argue that some moderator significantly affects the size of the syntactic priming effect. To that end, we made a pre-analysis decision to split the studies for p-curve analysis into two groups to be analyzed separately: those in which the main effect of interest was a main effect of syntactic priming and those in which the main effect of interest was a moderator of syntactic priming (e.g. whether using a temporal lag impacts the syntactic priming effect).

Method

Inclusion and exclusion criteria for p-curve. In p-curve, we can use only one p-value from each experiment. To decide which p-value to use, we identified the main statistical prediction of each study. This step is important since the p-curve analysis critically depends on the fact that the set of p-values included in the p-curve actually test the experimental hypothesis. Including auxiliary p-values that do not actually relate to the hypothesis of the study could cause the p-curve to be uninformative (Simonsohn et al., 2014b; Simonsohn, Simmons, & Nelson, 2015). Consequently, studies whose central claim was not about syntactic priming were not included in the p-curve analysis. Studies which predicted and/or found null results can also not be included in p-curve. Our sample included many such experiments.

As an example of a study that we included, consider Cleland (2003) Experiment 2. This study investigated priming of noun phrases in three conditions: when the noun is the same in the prime and target (sheep/sheep), when the noun is unrelated in the prime and target (knife/sheep), when the noun is semantically related in the prime and target (goat/sheep). The question was whether the priming effect is moderated by the relationship between the prime noun and target noun. One reported result is that there is a lexical boost effect such that there is an interaction between prime and the 3-condition factor semantic relatedness. But this result could be driven by simple lexical boost since it includes the same noun condition vs. the different noun condition. The main prediction of theoretical interest–and the apparent raison d'etre for the experiment–is that the semantically related condition will differ from the unrelated condition. Thus, the p-value that we p-curve is the p-value for the planned comparison between the semantically related condition and the unrelated condition. The fact that this is the main claim of interest is made clear in the abstract, in which the contribution of Experiment 2 is distilled to one sentence: "Experiment 2 showed an enhanced priming effect when prime and target contained semantically related nouns (e.g., 'goat' and 'sheep')."

Among the studies that were included in our main meta-analysis, we used the following criteria to determine inclusion or exclusion of p-values in the p-curve analysis.

- We excluded results in which the statistical result appears across two or more experiments in the same paper. Thus one data point in our p-curve analysis corresponds to one experiment and is never a combined analysis of two or more experiments in a paper.

- We excluded experiments in which the main claim involves a comparison with a population excluded from our meta-analysis (i.e., populations like aphasic patients or children). (For these studies, the control groups may appear in our main meta-analysis, but typically there is no main claim being made about how the control group will behave.)

- We excluded experiments in which the only main claim involves a dependent variable other than elicited sentence production (e.g., eye-tracking or reaction time). If an experiment had multiple main claims, some of which were about elicited sentence production, we included only the claim about elicited sentence production.

- When we cannot determine a clear "main result", we recorded the p-values for the main results in the order in which they appear (henceforth known as results "a", "b", "c", etc.).

- When an experiment reports an ANOVA by subject and item (F1 and F2) or any other analysis that analyzes subjects and items separately, we take the higher p-value of the two since using just one or the other is very anti-conservative (Barr et al., 2013) and thus violates an assumption of a p-curve analysis: that the p-value reflects the actual false-positive rate of the statistical test. Moreover, since the criteria for significance is that both F1 and F2 give p-values less than .05, any p-hacking would take place on the higher of the two p-values.

Coding procedures for p-curve. For the p-curve analysis, we created a p-curve disclosure table (available at https://osf.io/b9zyk/), following best practices (Simonsohn et al., 2014b, 2014a, 2015). One of the authors (K.M., A.J., R.F., or E.G.) first coded the results, and they were all independently recoded by another author. K.M. then went over each result in which the coder and re-coder's results did not match and discussed them. In cases of obvious error, the correct version was kept. In cases in which there was legitimate disagreement as to which p-value best reflected the main hypothesis of an experiment, both versions were used and the re-coded version was used as a "b" result.

Statistical procedures for p-curve. Following the procedures in Simonsohn et al. (2015), observed p-curves are tested for significant right skew (most p-values near zero) using Stouffer's method. We also compared the observed p-curve to a hypothetical p-curve with 33% power. Simonsohn et al. suggest that observed curves that are flatter than the 33% power curve suggest a lack of evidential value. Further comparisons with curves of varying statistical power are the basis for the estimated average power of the included studies. Essentially, we ask what power level is most likely to produce the observed curve shape. See Simonsohn et al. (2014) for more details of the p-curve analysis.[3]

To see how robust the p-curve analysis is to different plausible decisions about how to code the papers, whenever there were multiple results coded in the p-curve table, we randomly sampled one per experiment. We did this 100 times for each experiment to sample from 100 plausible ways in which the p-curve data could have been coded. In the rest of the analyais, we use representative plots and data from the 100 samples. (And for power estimates, the variation across these random samples was not typically larger than the confidence intervals for any one simulation.)

P-curve results. After eliminating studies in which the main hypothesis was about excluded populations or in which there was no priming-related main hypothesis internal to a single study, we were left with 139 studies from 65 papers. Of those 139 studies, only 88 (63%) had a significant result used to support the main hypothesis.

Figure 9 plots the distribution of p-values for a representative choice of what the "main" hypothesis is from the set of papers, for 56 experiments that directly test the existence of some form of syntactic priming, in blue. As recommended by the p-curve guide, the green line shows the expected p-curve under 33% power. We see that most studies in our sample (59%) have p-values less than .01, and the curve shows significant right skew ($p < .0001$ by a Stouffer's z-test for skew on both the full p-curve and half p-curve, as described in Simonsohn et al. 2015). Comparing the blue line to the green line shows that there are many more small p-values than we would expect if the power were 33%. The bias-corrected

[3] Note that Simonsohn et al. (2014) cast doubt on the validity of p-curve for discrete tests since the p-value from a discrete test does not necessarily correspond to the false-positive rate. In our case, while we are ultimately concerned with a comparison of proportions, the p-values used in psycholinguistics are typically designed such that the p-value reflects the actual false-positive rate (Barr et al., 2013). Therefore, this should not be a major issue.

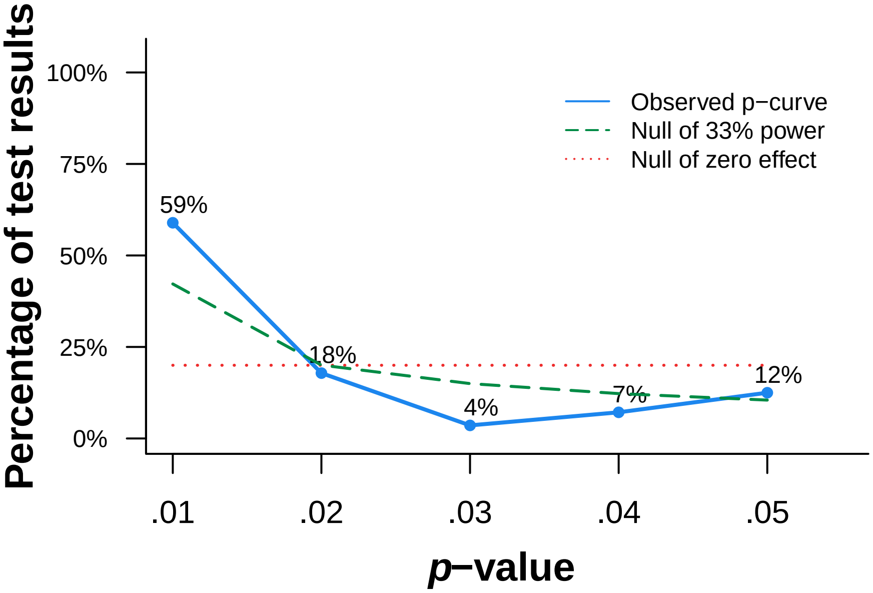


Figure 9. Default plot produced by p-curve for sample of studies in which the effect of interest is the existence of syntactic priming. The blue line is the distribution of p-values in the study. The red line shows the expected distribution of p-values if there was no underlying effect. The green line shows the expected p-curve under 33% statistical power. The right skew in the blue line shows there is evidential value in this set of studies.

average power estimate is 82% with a 95% confidence interval of [72%, 89%], which is above the recommended 80% threshold (a minimum standard for statistical power in many fields). There are numerically more p-values at the .04 and .05 levels than at the .03 levels (which is not consistent with a healthy p-curve), but the distribution is not extreme enough to conclude that there is definitive bias.

In a post-hoc p-curve analysis that we ran after a reviewer noticed that the recommended sample size for 80% power in priming studies with no lexical overlap was higher than the sample size of most papers in our sample, we ran a separate p-curve power analysis on just a subset of experiments (n=32) that contain no lexical overlap between prime and target. (If an experiment contained overlap and non-overlap conditions, it was not included in this subset.) For those papers, we found that average bias-corrected power was only 54% [32%, 72%]. This analysis suggests that, in the absence of lexical repetition between prime and target, studies in our sample may be underpowered.

Figure 10 plots the distribution of p-values for 32 experiments that do not directly investigate the existence of syntactic priming but ask questions about how syntactic priming is moderated by other variables. The p-curve for these studies is quite a bit flatter, although still significantly right skewed ($p < .01$). Although publication bias cannot be ruled out

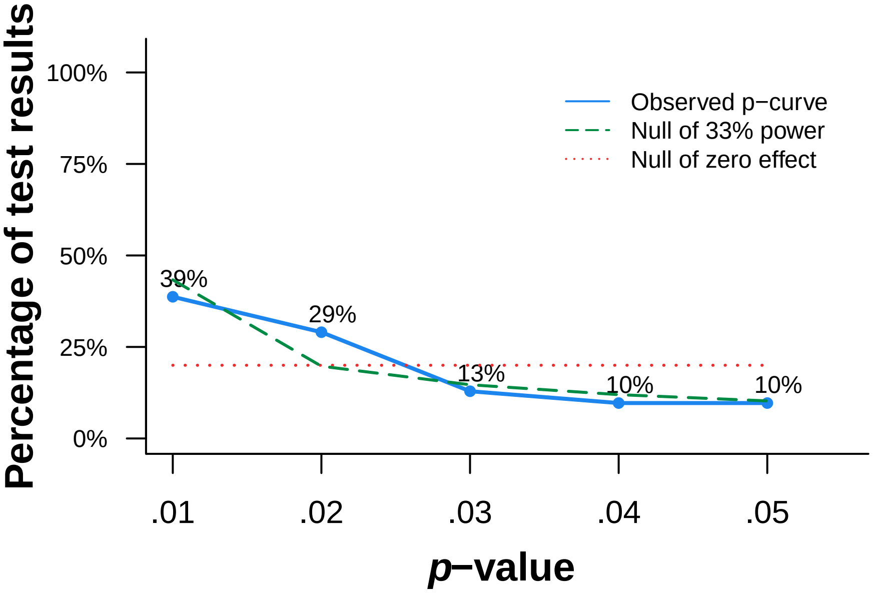


Figure 10. Default plot produced by p-curve for sample of studies in which the effect of interest is a moderator of syntactic priming. The blue line is the distribution of p-values in the study. The red line shows the expected distribution of p-values if there was no underlying effect. The green line shows the expected p-curve under 33% statistical power. The right skew in the blue line shows there is evidential value in this set of studies.

as an explanation for the flat p-curve, the p-curve is consistent with low statistical power (see 33% power curve, plotted in green, for comparison; p-curves significantly flatter than this line can be taken as evidence that the included studies lack evidential value). While the estimated power for these studies is only 53% [32%, 71%], the right skew and the fact that the curve is not significantly flatter than the 33% power curve ($p = .96$) suggests these studies do contain evidential value.

Interim conclusions. These results suggest that studies which purport to directly investigate the existence of syntactic priming do not suffer from extreme p-hacking and are moderately powered. Studies which investigate moderators of syntactic priming are likely underpowered but still contain some evidential value. The former suggests that our main meta-analysis is likely not overly influenced by publication bias or p-hacking.

One possible reason that studies investigating moderators of priming are underpowered is that these designs are more likely to involve interactions. Studies like these typically need many more subjects or items than those that are simply investigating a main effect of priming. In the next section, we will investigate just how many subjects and items need to be included to run well-powered syntactic priming studies of varying design.

Overall, the p-curve results should be interpreted with caution. Each p-curve analysis aggregates over a heterogeneous group of studies. In the p-curve analysis of studies that directly test priming, for instance, there are both studies with lexical overlap between prime and target and studies without. The ones with lexical overlap have much larger effect sizes on average than those wihtout. If the same amount of data is collected from a study with overlap as a study without, the latter would have higher power. Moreover, as we saw in the main meta-analysis, certain constructions show much stronger effects than others. In particular, constructions where one form is very infrequent are likely to show large effects in logistic regression and produce low p-values. Therefore, there is a possibility that the estimated power in the p-curve analysis may be too high for many common study designs.

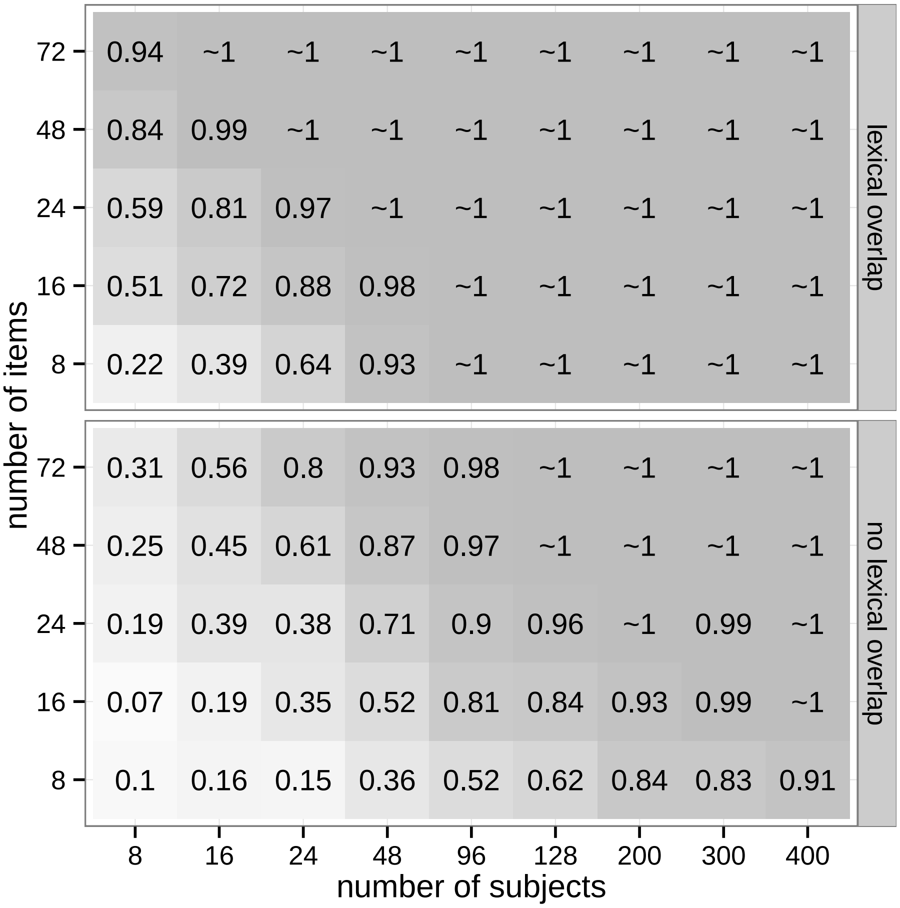
Sample size recommendations

Using the raw data collected from a subset of the papers in our sample, we can use simulation to give detailed recommendations for how to run future priming studies with sufficient statistical power. To do this, we used the mixed effect logistic regression described above in which we fit a regression to all data points from all studies for which we obtained raw data in order to simulate data for hypothetical new studies of varying designs. Specifically, we simulated 100 new experiments, each assumed to be a different random syntactic construction from a different experiment in a different paper. Each of the 100 experiments had S subjects and I items. The underlying "true" effect size was the effect size estimated using our actual data, and the effect size varied based on the paper, experiment, and subjects and items. We assume that 20% of data was lost due to "other" responses. We simulated experiments with all combinations of 8, 16, 24, 48, 96, 128, 200, 300, and 400 subjects and 8, 16, 24, 48, and 72 items.

Here, we assume that the underlying size of the "true" priming effect is .51 (change in log odds ratio) as estimated in the meta-analysis. However, using the raw data from the subset of 45 papers (or performing a meta-analysis on published results from those 45 papers), we find an effect size of only .34. Whether we assume the underlying true priming effect is .51 or .34 affects the power estimates. Here, we use the estimate of .51, which is based on more data. In the SI, we also provide sample size recommendations for when the underlying effect size is .51. For researchers performing syntactic priming studies, we recommend estimating the size of the expected effect based on the moderators (whether there is lexical boost, which construction is being used, etc.) and using an effect size appropriate to the task. The details of how we performed these simulations are in the SI.

First, we report the results of a simple two-cell design testing for the presence of syntactic priming in the absence of lexical overlap between prime and target. We show estimated statistical power in Figure 11. If power is 0, that means that when there is an underlying true effect of priming, we will detect it 0% of the time. If power is 1, that means that we will detect it 100% of the time. 80% is a standard threshold for power in experiments. To achieve that threshold comfortably for observing a simple priming effect, we recommend 96 subjects and 24 items (for 90% power).

Whereas many traditional power analyses in psychology focus on the number of subjects, psycholinguistics experiments typically have many items. As we show through these simulations, an increase in the number of items per participant substantially increases the power.



lexical overlap	8	16	24	48	96	128	200	300	400
72	0.94	~1	~1	~1	~1	~1	~1	~1	~1
48	0.84	0.99	~1	~1	~1	~1	~1	~1	~1
24	0.59	0.81	0.97	~1	~1	~1	~1	~1	~1
16	0.51	0.72	0.88	0.98	~1	~1	~1	~1	~1
8	0.22	0.39	0.64	0.93	~1	~1	~1	~1	~1

no lexical overlap	8	16	24	48	96	128	200	300	400
72	0.31	0.56	0.8	0.93	0.98	~1	~1	~1	~1
48	0.25	0.45	0.61	0.87	0.97	~1	~1	~1	~1
24	0.19	0.39	0.38	0.71	0.9	0.96	~1	0.99	~1
16	0.07	0.19	0.35	0.52	0.81	0.84	0.93	0.99	~1
8	0.1	0.16	0.15	0.36	0.52	0.62	0.84	0.83	0.91

Figure 11. Power to detect priming effect with lexical overlap (on top) and no lexical overlap (bottom) when true effect (in difference in log odds ratio) is .51 with a lexical overlap effect of .67.

When there is lexical overlap between the prime and target, the main effect of priming is much bigger and we need fewer subjects and items to have sufficient power, as shown in the bottom panel of Figure 11. Even with 16 subjects and 16 items, we have 92% power. Note that this number of subjects and items is not sufficient to detect a presence of a lexical overlap effect but merely to detect that priming exists when all the items repeat material between prime and target.

Next, we ask how many subjects and items we need to detect a moderator of priming (i.e., to detect an interaction), such as whether the priming effect is moderated by lexical overlap or lag. We try this for three different interaction sizes: coefficients of .2, .5, and 1. A coefficient of .2 is roughly the same order of magnitude as the interaction between prime and filler lag (a small but likely real interaction of theoretical interest). The coefficient of 1 is roughly the size of the lexical overlap effect. The coefficient .5 is somewhere in between. We show results for this analysis in Figure 12.

Even with 400 subjects and 72 items, we do not achieve 80% power when the underlying interaction coefficient is .2. When it is .5 (a size larger than many of the interaction effects

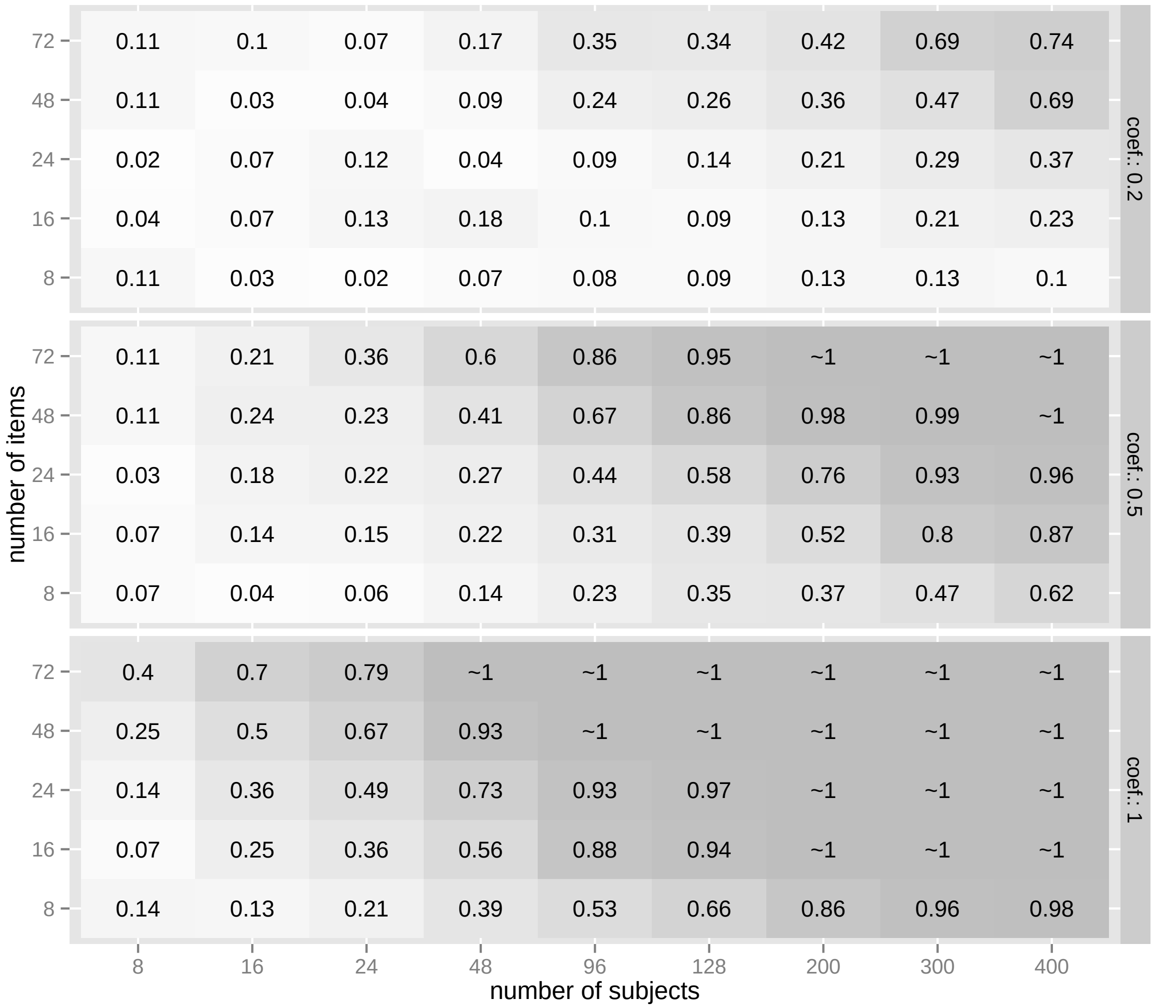


Figure 12. Power to detect an interaction with the priming effect, no lexical overlap when true effect is .51.

estimated in our meta-analysis), we need 128 subjects and 48 items to achieve 90% power. And when the coefficient of interaction is large (1), we need fewer subjects.

Overall, these results suggest that, as our p-curve analysis also suggests, many of the studies in our sample are likely underpowered for the hypotheses that they are testing. Another important observation based on these simulations is that, even with a very large number of subjects, studies with fewer than 24 items are likely to be underpowered. Similarly, no matter how many items a study has, studies run with small numbers of participants (< 24) are likely to be underpowered.

Conclusion

We conclude that there is strong evidence in the literature, over the last 30 years, for syntactic priming. We estimate that the size of this effect is small to medium when there is no lexical overlap and large when there is lexical overlap. The estimated effect is not likely the result of publication bias or p-hacking since most studies that investigate syntactic priming itself have acceptable statistical power. As has been reported in the literature, there are significant effects of lexical overlap, of lag between prime and target, and of bilingual priming–especially in its interaction with lexical overlap. We have quantitatively verified these effects and, harnessing the power of our large sample, estimated their size with more precision than any previous estimate provided.

While syntactic priming appears to be a robust effect, we found that the accumulated literature that studies moderators of syntactic priming suffers from low statistical power and, in the future, we recommend using larger sample sizes to study such phenomena. For that reason, we urge caution in interpreting studies that use only modestly sized samples to investigate whether some particular factor (other than lexical overlap, which leads to large effects) significantly affects the size of the syntactic priming effect. There have been cases in the literature where there are discrepancies whether an experiment finds significant effects of moderators like temporal lag or bilingualism. Given the typical sample sizes used in these experiments, it is possible that these discrepancies are mere noise and not reflective of any meaningful difference between the experiments.

It is also important to remember the limitations of a meta-analysis like this one. The results reported here are descriptive results of the syntactic priming literature sampled here. One might wonder, for instance, whether syntactic priming effects exist in the real world or only in laboratory settings. On that question, our meta-analysis has nothing to say. Nor can our meta-analysis answer whether the studies included here are in fact providing evidence for the various theories of syntactic priming and language processing more generally. There are also many variables, such as the types of fillers used, that were not included in our meta-analysis. We encourage future researchers to use the resources we have built here to ask and answer questions of their own devising. Our spreadsheets and materials used for these analyses are available on the Open Science Framework at https://osf.io/b9zyk/.

We hope that this work can be the basis of continuing meta-analysis and aggregation of data in syntactic priming. There is an extensive parallel literature on comprehension priming (where the dependent measure is sentence interpretation, reaction time, and so on) that could benefit from a similar sort of meta-analysis. Our meta-analysis also did not include unpublished work or work that appeared only at conferences. A meta-analysis of meta-analyses found systematic differences between unpublished work and peer-reviewed journal work such that the effect size of published studies is higher (Polanin, Tanner-Smith, & Hennessy, 2015), and it would be interesting to see whether the syntactic priming literature shows this effect. Our meta-analysis predicts that we should see only a moderate difference for studies investigating whether priming exists but perhaps a larger difference for studies investigating moderators of priming.

Acknowledgments

For sharing and explaining their raw data, we thank Sarah Bernolet, Kay Bock, Holly Branigan, Ann Bunger, Zhenguan Cai, Kiel Christianson, Timmy Desmet, Vic Ferreira, Martijn Goudbeek, Zenzi Griffin, Rob Hartsuiker, Iva Ivanova, Florian Jaeger, Leila Kantola, Michael Kaschak, Gerrit Jan Kootstra, Janet McLean, Alissa Melinger, Katherine Messenger, Andriy Myachykov, Sandra Pappert, Martin Pickering, Caro Rowland, Mikel Santesteban, Christoph Scheepers, Jeong-Ah Shin, Bob Slevc, Katie Slocombe, Neal Snider, Nele Verreyt, and all their co-authors. For helpful discussions, we thank Shravan Vasishth, Uri Simonsohn, Duane Watson and members of the Communication and Language Lab, Sarah Brown-Schmidt and members of the Conversation Lab, Gary Dell and members of the Language Production Lab, Aaron Benjamin and members of the Human Memory and Cognition Lab, members of the Comprehension Joint Lab Meeting, and members of Tedlab. K.M. was supported by a Department of Defense NDSEG graduate fellowship. A.J. was supported by NSF Grant DGE-1144245.

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.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *ArXiv Preprint ArXiv:1506.04967*.

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2007). Shared syntactic representations in bilinguals: Evidence for the role of word-order repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(5), 931. Retrieved from http://psycnet.apa.org/journals/xlm/33/5/931/

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2012). Effects of phonological feedback on the selection of syntax: Evidence from between-language syntactic priming. *Bilingualism: Language and Cognition*, *15*(03), 503–516. Retrieved from http://journals.cambridge.org/abstract_S1366728911000162

Bernolet, S., Hartsuiker, R. J., & Pickering, M. J. (2013). From language-specific to shared syntactic representations: The influence of second language proficiency on syntactic sharing in bilinguals. *Cognition*, *127*(3), 287–306. doi:10.1016/j.cognition.2013.02.005

Biria, R., Ameri-Golestan, A., & Antón-Méndez, I. (2010). Syntactic priming effects between modalities: A study of indirect questions/requests among persian english learners. *English Language Teaching*, *3*(3), p111. Retrieved from http://www.ccsenet.org/journal/index.php/elt/article/view/7221

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387. doi:10.1016/0010-0285(86)90004-6

Bock, K. (1989). Closed-class immanence in sentence production. *Cognition*, *31*(2), 163–186.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

Branigan, H. P., Pickering, M. J., McLean, J. F., & Cleland, A. (2007). Syntactic alignment and participant role in dialogue. *Cognition*, *104*(2), 163–197. doi:10.1016/j.cognition.2006.05.006

Branigan, H. P., Pickering, M. J., Stewart, A. J., & McLean, J. F. (2000). Syntactic priming in spoken production: Linguistic and temporal interference. *Memory & Cognition*, *28*(8), 1297–1302. Retrieved from http://link.springer.com/article/10.3758/BF03211830

Bunger, A., Papafragou, A., & Trueswell, J. C. (2013). Event structure influences language production: Evidence from structural priming in motion event description. *Journal of Memory and Language*, *69*(3), 299–323. doi:10.1016/j.jml.2013.04.002

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013a). Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience*, *14*(8), 585–585.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013b). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Cleland, A. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, *49*(2), 214–230. doi:10.1016/S0749-596X(03)00060-3

Cleland, A., & Pickering, M. J. (2006). Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language*, *54*(2), 185–198. doi:10.1016/j.jml.2005.10.003

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (rev.* Lawrence Erlbaum Associates, Inc.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.

Coyle, J. M., & Kaschak, M. P. (2012). Female fertility affects men's linguistic choices. *PLoS ONE*, *7*(2), e27971. doi:10.1371/journal.pone.0027971

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Ferreira, V., & Bock, K. (2006). The functions of structural priming. *Language and Cognitive Processes*, *21*(7-8), 1011–1029.

Ferreira, V., Bock, K., Wilson, M. P., & Cohen, N. J. (2008). Memory for syntax despite amnesia. *Psychological Science*, *19*(9), 940–946. Retrieved from http://pss.sagepub.com/content/19/9/940.short

Francis, G., Tanzman, J., & Matthews, W. J. (2014). Excess Success for Psychology Articles in the Journal Science. *PLoS ONE*, *9*(12), e114255. doi:10.1371/journal.pone.0114255

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical* *models*. Cambridge University Press.

Gompel, R. P. van, Arai, M., & Pearson, J. (2012). The representation of mono- and intransitive structures. *Journal of Memory and Language*, *66*(2), 384–406. doi:10.1016/j.jml.2011.11.005

Goudbeek, M., & Krahmer, E. (2012). Alignment in interactive reference production: Content planning, modifier ordering, and referential overspecification. *Topics in Cognitive Science*, *4*(2), 269–289. doi:10.1111/j.1756-8765.2012.01186.x

Hartsuiker, R. J. (1999). Priming word order in sentence production. *The Quarterly Journal of Experimental Psychology: Section A*, *52*(1), 129–147. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/713755798

Hartsuiker, R. J., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition*, *75*(2), B27–B39. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027799000803

Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, *58*(2), 214–238. doi:10.1016/j.jml.2007.07.003

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in spanish-english bilinguals. *Psychological Science*, *15*(6), 409–414. Retrieved from http://pss.sagepub.com/content/15/6/409.short

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin*, *117*(1), 167.

Heydel, M., & Murray, W. S. (2000). Conceptual effects in sentence priming: A cross-linguistic perspective. In *Cross-linguistic perspectives on language processing* (pp. 227–254). Springer.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241.

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

Jaeger, T., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83. doi:10.1016/j.cognition.2012.10.013

Kantola, L., & Gompel, R. P. G. van. (2011). Between- and within-language priming is the same: Evidence for shared bilingual syntactic representations. *Memory & Cognition*, *39*(2), 276–290. doi:10.3758/s13421-010-0016-5

Kaschak, M. P. (2007). Long-term structural priming affects subsequent patterns of language production. *Memory & Cognition*, *35*(5), 925–937. Retrieved from http://link.springer.com/article/10.3758/BF03193466

Kaschak, M. P., & Borreggine, K. L. (2008). Is long-term structural priming affected by patterns of experience with individual verbs? *Journal of Memory and Language*, *58*(3), 862–878. doi:10.1016/j.jml.2006.12.002

Kaschak, M. P., Kutta, T. J., & Jones, J. L. (2011). Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic Bulletin & Review*, *18*(6), 1133–1139. doi:10.3758/s13423-011-0157-y

Kaschak, M. P., Loney, R. A., & Borreggine, K. L. (2006). Recent experience affects the strength of structural priming. *Cognition*, *99*(3), B73–B82. doi:10.1016/j.cognition.2005.07.002

Kim, Y., & McDonough, K. (2007). Learners' production of passives during syntactic priming activities. *Applied Linguistics*, *29*(1), 149–154. doi:10.1093/applin/amn004

Kootstra, G. J., Hell, J. G. van, & Dijkstra, T. (2010). Syntactic alignment and shared word order in code-switched sentence production: Evidence from bilingual monologue and dialogue. *Journal of Memory and Language*, *63*(2), 210–231. doi:10.1016/j.jml.2010.03.006

Landy, J., & Goodwin, G. (2014). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Manuscript Submitted for Publication*.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Sage Publications Thousand Oaks, CA.

Messenger, K. (2010). Syntactic priming and children's production and representation of the passive. Retrieved from http://www.tandfonline.com/doi/full/10.1080/10489220903472648

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. doi:10.1126/science.aac4716

Pappert, S., & Pechmann, T. (2013). Bidirectional structural priming across alternations: Evidence from the generation of dative and benefactive alternation structures in german. *Language and Cognitive Processes*, *28*(9), 1303–1322. doi:10.1080/01690965.2012.672752

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*(4), 633–651. doi:10.1006/jmla.1998.2592

Pickering, M. J., & Ferreira, V. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*(3), 427–459. doi:10.1037/0033-2909.134.3.427

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2015). Estimating the difference between published and unpublished effect sizes a meta-review. *Review of Educational Research*, 0034654315582067.

Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language*, *38*(3), 265–282. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X97925468

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Scheepers, C. (2003). Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, *89*(3), 179–205. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027703001197

Schoonbaert, S., Hartsuiker, R. J., & Pickering, M. J. (2007). The representation of lexical and syntactic information in bilinguals: Evidence from syntactic priming. *Journal of Memory and Language*, *56*(2), 153–171. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X06001471

Shin, J.-A., & Christianson, K. (2012). Structural priming and second language learning: Structural priming and l2 learning. *Language Learning*, *62*(3), 931–964. doi:10.1111/j.1467-9922.2011.00657.x

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve and effect size correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to ulrich and miller (2015).

Slocombe, K. E., Alvarez, I., Branigan, H. P., Jellema, T., Burnett, H. G., Fischer, A., … Levita, L. (2013). Linguistic Alignment in Adults with and Without Asperger's Syndrome. *Journal of Autism and Developmental Disorders*, *43*(6), 1423–1436. doi:10.1007/s10803-012-1698-2

Tooley, K. M., & Traxler, M. J. (2010). Syntactic priming effects in comprehension: A critical review. *Language and Linguistics Compass*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2010.00249.x/full

Vernice, M., Pickering, M. J., & Hartsuiker, R. J. (2012). Thematic emphasis in language production. *Language and Cognitive Processes*, *27*(5), 631–664. doi:10.1080/01690965.2011.572468

Verreyt, N., Bogaerts, L., Cop, U., Bernolet, S., De Letter, M., Hemelsoet, D., … Duyck, W. (2013). Syntactic priming in bilingual patients with parallel and differential aphasia. *Aphasiology*, *27*(7), 867–887. doi:10.1080/02687038.2013.791918

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/