Supplementary Materials

Supplementary Information

Anonymized

2017-05-11

Contents

The present document provides further information to accompany the paper "Assessing methodological practices in language acquisition research through meta-analyses".

Instructions to reproduce the analyses

To reproduce the analyses we show here, the following steps are necessary: 1. Clone from github the metalab repository: https://github.com/langcog/metalab 2. Download the scripts for this paper and put them in a subfolder. 3. Adjust the location of `global.R` in `initial_data.R` to read in the correct file. 4. Run the scripts. To load data, first run `initial_data.R`.

Data preprocessing

We preprocess the data to exclude data testing special populations, non-critical conditions, and to remove outliers. We also exclude a longitudinal dataset from these analyses, because it is not suitable for our purposes.

```
## CLEAN DATA ####
all_data = all_data %>%
  filter(is.na(condition_type) | condition_type == "critical") %>%
  filter(dataset!="Pointing and vocabulary (longitudinal)") %>%
  filter(infant_type == "typical")


#Remove outliers

clean_data = all_data %>%
  group_by(dataset) %>%
  mutate(mean_es = median(d_calc)) %>%
  mutate(sd_es = sd(d_calc)) %>%
  ungroup() %>%
  mutate(no_outlier = ifelse(d_calc < mean_es+3*sd_es, ifelse(d_calc > mean_es-3*sd_es, TRUE, FALSE), FA
  filter(no_outlier)

#Comment out if you do not want to remove outliers
all_data = clean_data
```

Effect size calculation

Effect sizes were calculated based on standard formulae, below we showcase commented code from the MetaLab platform at http://metalab.stanford.edu and the associated github repository https://github.com/langcog/metalab. In the source script `compute_es.R` a number of special cases are also computed, we leave them out of this document for brevity.

```
  #start of decision tree where effect sizes are calculated differently based on participant design
  #depending on which data is available, effect sizes are calculated differently
  if (participant_design == "between") {
    es_method  <- "between"
    #effect size calculation
    if (complete(x_1, x_2, SD_1, SD_2)) {
      pooled_SD <- sqrt(((n_1 - 1) * SD_1 ^ 2 + (n_2 - 1) * SD_2 ^ 2) / (n_1 + n_2 - 2)) # Lipsey & Wil
      d_calc <- (x_1 - x_2) / pooled_SD # Lipsey & Wilson (2001)
    } else if (complete(t)) {
      d_calc <- t * sqrt((n_1 + n_2) / (n_1 * n_2)) # Lipsey & Wilson, (2001)
    } else if (complete(f)) {
      d_calc <- sqrt(f * (n_1 + n_2) / (n_1 * n_2)) # Lipsey & Wilson, (2001)
    }
    if (complete(n_1, n_2, d_calc)) {
      #now that effect size are calculated, effect size variance is calculated
      d_var_calc <- ((n_1 + n_2) / (n_1 * n_2)) + (d_calc ^ 2 / (2 * (n_1 + n_2)))
    } else if (complete(r)) {
      #if r instead of d is reported, transform for standardization
      d_calc <- 2 * r / sqrt(1 - r ^ 2)
      d_var_calc <- 4 * r_var / ((1 - r ^ 2) ^ 3)
    } else if (complete(d, d_var)) {
      #if d and d_var were already reported, use those values
      d_calc <- d
      d_var_calc <- d_var
    }

  } else if (participant_design == "within_two") {
    if (is.na(corr)) {
      #if correlation between two measures is not reported, use an imputed correlation value
      corr <- corr_imputed
    }
    #effect size calculation
    if (complete(x_1, x_2, SD_1, SD_2)) {
      pooled_SD <- sqrt((SD_1 ^ 2 + SD_2 ^ 2) / 2) # Lipsey & Wilson (2001)
      d_calc <- (x_1 - x_2) / pooled_SD # Lipsey & Wilson (2001)
      es_method  <- "group_means_two"
    } else if (complete(x_1, x_2, SD_dif)) {
      within_SD <- SD_dif / sqrt(2 * (1 - corr)) # Lipsey & Wilson (2001); Morris & DeShon (2002)
      d_calc <- (x_1 - x_2) / within_SD # Lipsey & Wilson (2001)
      es_method  <- "group_means_two"
    } else if (complete(x_dif, SD_1, SD_2)) {
      pooled_SD <- sqrt((SD_1 ^ 2 + SD_2 ^ 2) / 2) # Lipsey & Wilson (2001)
      d_calc <- x_dif / pooled_SD # Lipsey & Wilson (2001)
      es_method  <- "subj_diff_two"
    } else if (complete(x_dif, SD_dif)) {
      wc <- sqrt(2 * (1 - corr))
      d_calc <- (x_dif / SD_dif) * wc #Morris & DeShon (2002)
      es_method  <- "subj_diff_two"
    } else if (complete(t)) {
      wc <- sqrt(2 * (1 - corr))
      d_calc <- (t / sqrt(n_1)) * wc #Dunlap et al., 1996, p.171
      es_method  <- "t_two"
    } else if (complete(f)) {
      wc <- sqrt(2 * (1 - corr))
      d_calc <- sqrt(f / n_1) * wc
      es_method  <- "f_two"
    }
    if (complete(n_1, d_calc)) {
      #now that effect size are calculated, effect size variance is calculated
      #d_var_calc <- ((1 / n_1) + (d_calc ^ 2 / (2 * n_1))) * 2 * (1 - corr) #we used this until 4/7/17
      d_var_calc <- (2 * (1 - corr)/ n_1) + (d_calc ^ 2 / (2 * n_1)) # Lipsey & Wilson (2001)
    } else  if (complete(r)) {
      #if r instead of d is reported, transform for standardization
      d_calc <- 2 * r / sqrt(1 - r ^ 2)
      d_var_calc <- 4 * r_var / ((1 - r ^ 2) ^ 3)
      es_method  <- "r_two"
    } else if (complete(d, d_var)) {
      #if d and d_var were already reported, use those values
      d_calc <- d
      d_var_calc <- d_var
      es_method  <- "d_two"
    }

  } else if (participant_design == "within_one") {
    if (complete(x_1, x_2, SD_1)) {
      d_calc <- (x_1 - x_2) / SD_1
      es_method  <- "group_means_one"
    } else if (complete(t)) {
      d_calc <- t / sqrt(n_1)
      es_method  <- "t_one"
    } else if (complete(f)) {
      d_calc <- sqrt(f / n_1)
      es_method  <- "f_one"
    }
    if (complete(n_1, d_calc)) {
      #d_var_calc <- (2/n_1) + (d_calc ^ 2 / (4 * n_1)) #we used this until 4/7/2017
      d_var_calc <- (1 / n_1) + (d_calc ^ 2 / (2 * n_1)) #this models what is done in metafor package,

    } else if (complete(r, n_1)){
      # this deals with pointing and vocabulary
      # Get variance of transformed r (z; fisher's tranformation)
      SE_z = 1 / sqrt(n_1 - 3) # from Howell (2010; "Statistical methods for Psychology", pg 275)
      var_z = SE_z ^ 2

      # Transform z variance to r variance
      var_r = tanh(var_z)  # from wikipedia (https://en.wikipedia.org/wiki/Fisher_transformation) for c

      # Transform r to d
      d_calc = 2 * r / (sqrt(1 - r ^ 2)) # from (Hunter and Schmidt, pg. 279)
      d_var_calc = (4 * var_r)/(1 - r ^ 2) ^ 3 # from https://www.meta-analysis.com/downloads/Meta-anal

      es_method  <- "r_one"

    } else if (complete(r)) {
      #if r instead of d is reported, transform for standardization
      d_calc <- 2 * r / sqrt(1 - r ^ 2)
      d_var_calc <- 4 * r_var / ((1 - r ^ 2) ^ 3)

       es_method  <- "r_one"

    } else  if (complete(d, d_var)) {
      #if d and d_var were already reported, use those values
      d_calc <- d
      d_var_calc <- d_var
      es_method  <- "d_one"
```

Readers interested in calculating effect sizes for single papers might want to try out the shiny app or spreadsheet based on @lakens2013 https://katherinemwood.shinyapps.io/lakens_effect_sizes/ https://osf.io/ixGcd/

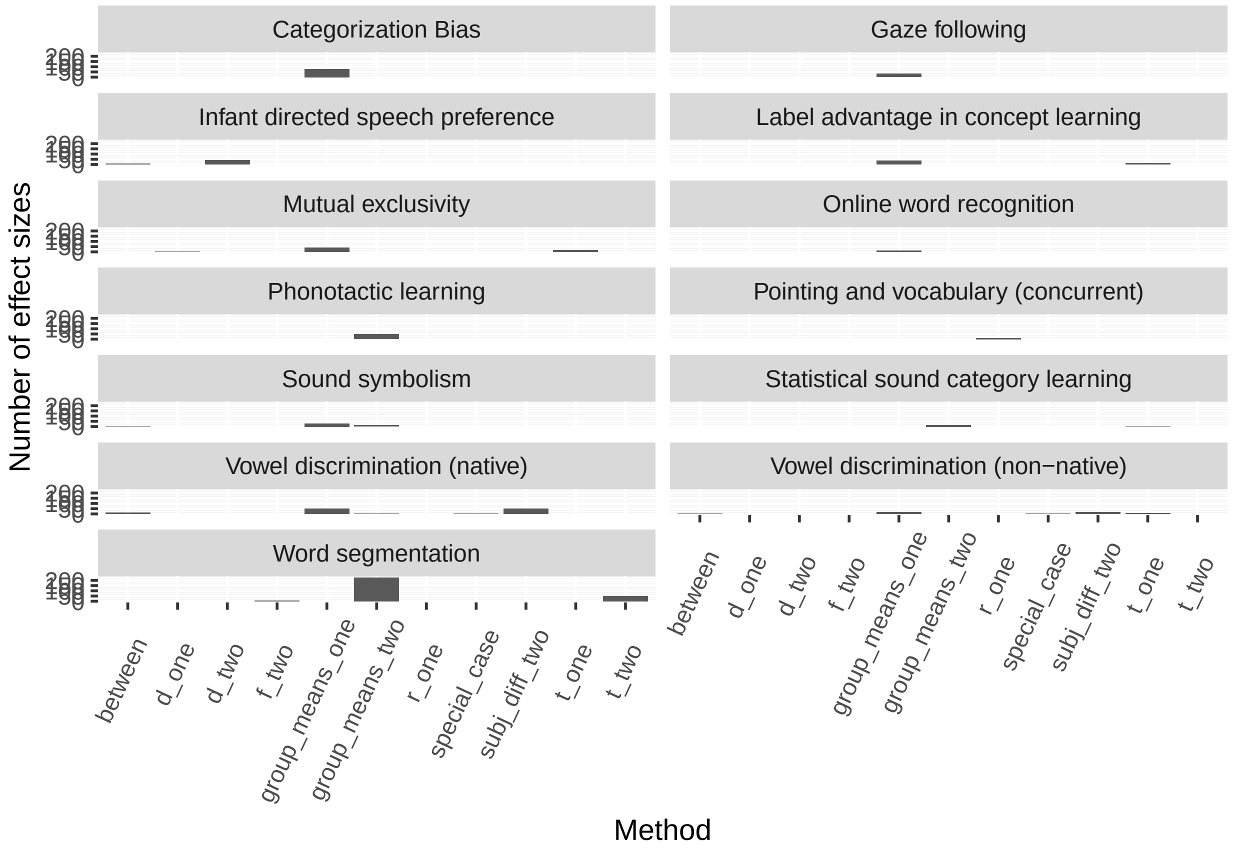
Data availability

In this section we describe the source data, with a focus on which data were used to compute effect sizes according to the formulae above.

```
counts <- all_data %>%
  select(dataset, es_method)

plot_esbasis = ggplot(counts) +
  geom_bar(aes(x = es_method)) +
  facet_wrap(~dataset, ncol=2) +
  ylab("Number of effect sizes") +
  xlab("Method") +
  theme(axis.text.x = element_text(angle=65, vjust=.8, hjust=.8))

plot_esbasis
```



Additional analyses

To illustrate the disparity between the oldest effect size and the meta-analytic effect, and consequently the difference in power, we plot the difference between both against the oldest effect. This difference is larger as oldest effect size increases, with an average of 1.08 compared with an average effect size of 0.6 (note that we based this on the absolute value). The plot showcases that researchers might want to be wary of large effects, as they are more likely to be non-representative of the true phenomenon compared to smaller initial effects being reported. Especially when making decisions about sample sizes, large effect might thus not be the best guide. Taking the above-mentioned mean values as example, a realistic sample size to ensure 80% power would be 45 participants, instead of 7 participants suggested by the first paper. While these numbers average over research questions and methods, which all influence the specific number of participants necessary, this example showcases that experimenters should take into account as much evidence as available to be able to plan for robust and reproducible studies.

How does effect size relate to *p*-values?

Single experiments are often evaluated by their associated *p*-value, despite the frequent criticisms and well-documented shortcomings of that measure [@Ioannidis2005]. One of them is particularly relevant here: In the Pearson-Neyman model (implicitly) underlying most current empirical research, *p*-values should be used to inform a binary decision, namely to either reject the null hypothesis or fail to do so; in contrast, effect sizes are a continuous measure. Researchers sometimes believe that significant *p*-values are equivalent to very high effect sizes, and that non-significant *p*-values are due to very low effect sizes near zero lead; and more generally that *p*-values have a continuous interpretation such that very low *p*-values necessarily indicate very

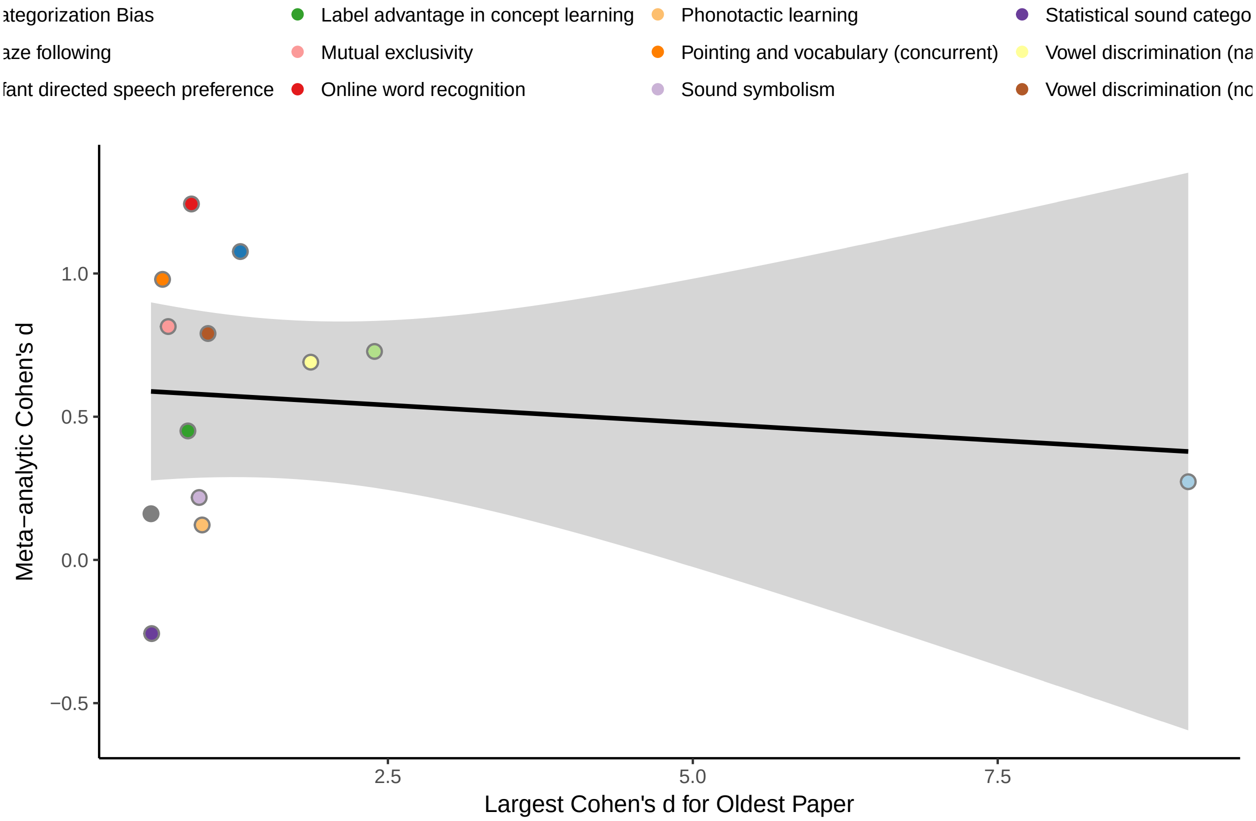


Figure 1: Correlation of largest d from oldest paper and difference between oldest d and meta-analytic d.

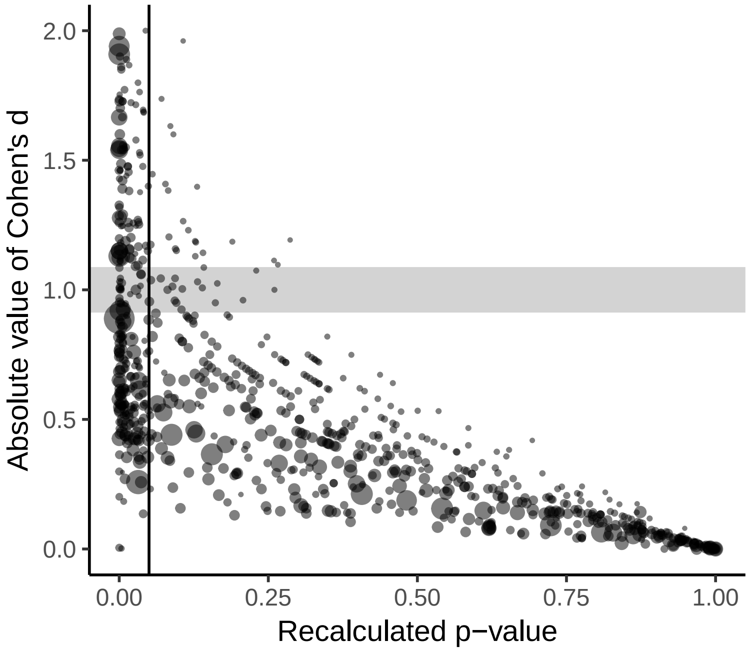


Figure 2: Comparison of a study's effect size and the according *p*-values. Point size reflects sample size. The typical significance threshold of .05 is indicated by a vertical line.

strong evidence for differences, whereas values near .05 indicate weaker evidence for a difference.

Figure 2 addresses these intutions by showing p-value and effect size data within MetaLab. The vertical line marks the usual binary decision threshold for p-values at .05. We see that, for extreme values, namely effect sizes near zero and above 2, the intuition that p-values will be respectively large and small is borne out. However, the majority of effect sizes we observe falls in a range that with sufficient power, in other words a sufficiently large sample, can lead to a significant outcome. Underpowered studies, in contrast, might tap into a similar sized effect but fail to reach significance. In Figure 2, the grey horizontal band illustrates such a region where both significant and non-significant results are observed.

Power over time

The comparison of initial and meta-analytic effect size has a number of caveats, for example, as we will lay out in the next section, methods might be different between initial reports and our overall sample; the availability of methods changes over time, as new approaches are being developed and automated procedures become more common. Further, the largest effect size from a seminal paper might have been spurious, and the research community could well be aware of that. In additional, as infant research becomes more common, recruitment and obtaining funds might both become easier, thereby increasing typical sample size over the years. For a more continuous approach, we thus investigate power (which is determined by effect size and sample size) as follows. We first generate a meta-analytic model for each dataset that takes into account infant age and method and then derive the respective to be expected effect size base on those data for each entry in this dataset. Power is then estimated based on the sample size actually tested. Across datasets we observe a general negative trend, with varied steepness. The only positive trends occur in the upper ranges of estimated power and for older children.

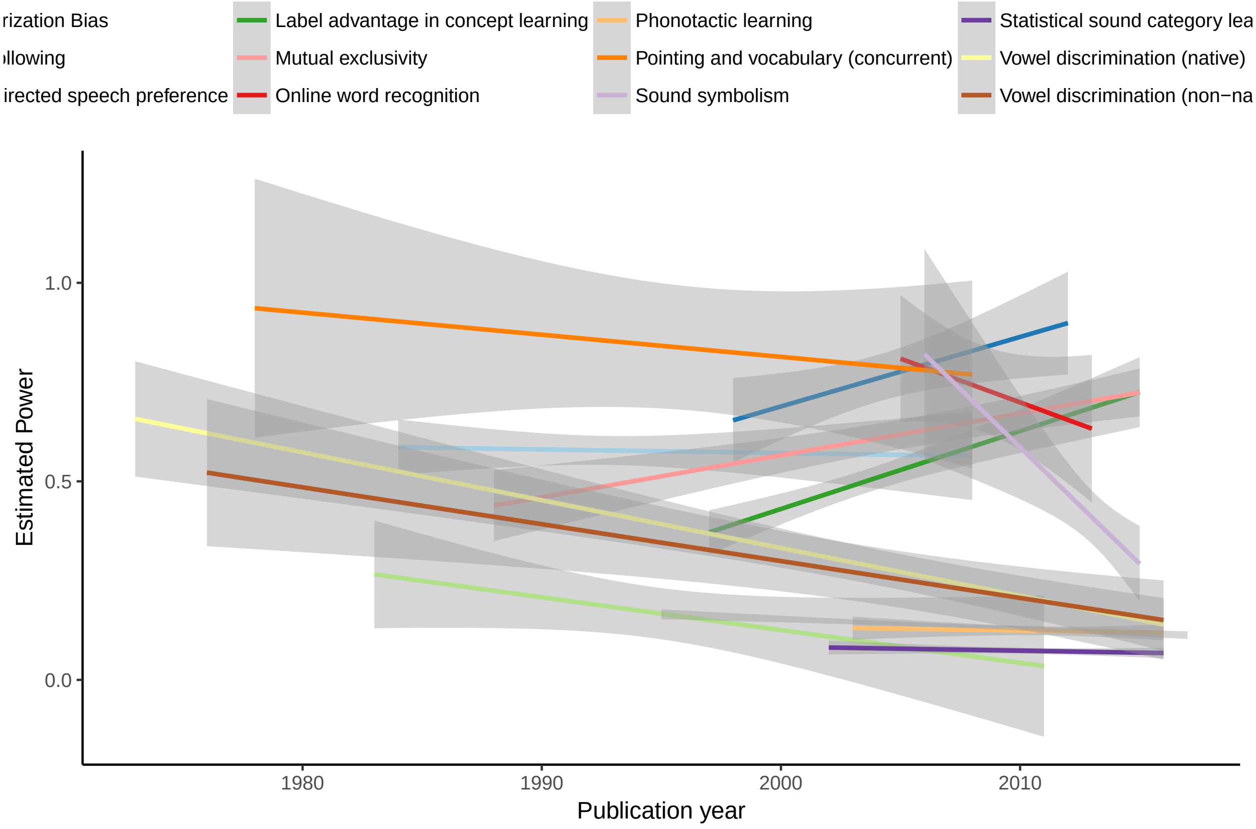


Figure 3: To BE UPDATED: Summary plot of power across years and meta-analyses

Publication biases, p-hacking, etc

It is well established that psychological research has a strong bias against non-significant findings, with an unduly large proportion of published studies containing significant p-values [estimates ranging between .9% and 12% in psychology; @manylabs2; @Greenwald72; @Sterling58; @Sterling95]. Additionally, although it is desirable for results to be surprising, it is equally if not more important that they be interpretable given current theories and previous data. As a result of the latter bias, there may be under-reporting of certain types of results depending on the direction of the result rather than its strengt. We would like to point out that in the context of developmental psychology, both of these bias may be statistically less strong. As to the first, given the field-specific habit of interpreting non-significant results as evidence of absence, it is possible that a higher proportion of results contain non-significant results. As to the second, many infant methods do not specify directions of results that are universally accepted. For instance, in preferential looking time studies infants can show a preference for familiar or for novel stimuli – both are considered interpretable and valid. This may therefore lead to wider acceptance of a range of results.

Funnel plot asymmetry

There are numerous ways to estimate whether the published literature is biased. The most common and straightforward is an assessment of funnel plot asymmetry. A funnel plot displays effect sizes against their variance (with 0 being plotted up). The expectation in the absence of biases is that effect sizes are equally distributed around the meta-analytic mean, and that the are spread out more the larger their variance, creating a triangle-like shape. Biases can lead to distortions in this distribution. The large the asymmetry, the more likely a bias is. We quantify funnel plot asymmetry with a rank correlation test implemented in the metafor package [@metafor].

Across datasets, the difference in distributions and range covered in effect sizes is striking, as is the variance in observed precision (the highest points correspond to high precision and low variance). The indicated relationship between effect sizes and their variance was assessed with a nonparametric test and turnd out to be significant – indicating publication bias in favor of significant results – for five datasets.

Table 1: Non-parametric test for funnel plot assymmetry. A significant value indicates bias.

Meta-analysis	Kendall's Tau	p-value
Phonotactic learning	-0.16	0.124
Statistical sound category learning	-0.17	0.398
Categorization Bias	0.25	0.001
Gaze following	0.17	0.169
Infant directed speech preference	0.17	0.083
Label advantage in concept learning	0.20	0.051
Mutual exclusivity	0.41	< .001
Vowel discrimination (native)	0.42	< .001
Vowel discrimination (non-native)	0.37	< .001
Pointing and vocabulary (concurrent)	0.36	0.116
Sound symbolism	0.13	0.207
Online word recognition	0.47	0.019
Word segmentation	0.10	NA

P-curves

A possible consequences of undisclosed flexibility is a distribution of p values with increased frequency just below the significance threshold, and/or an overall flat distribution of p values indicating that those results that were significant in fact represent the to be expected 5% type-I error. We will use p-curves to assess

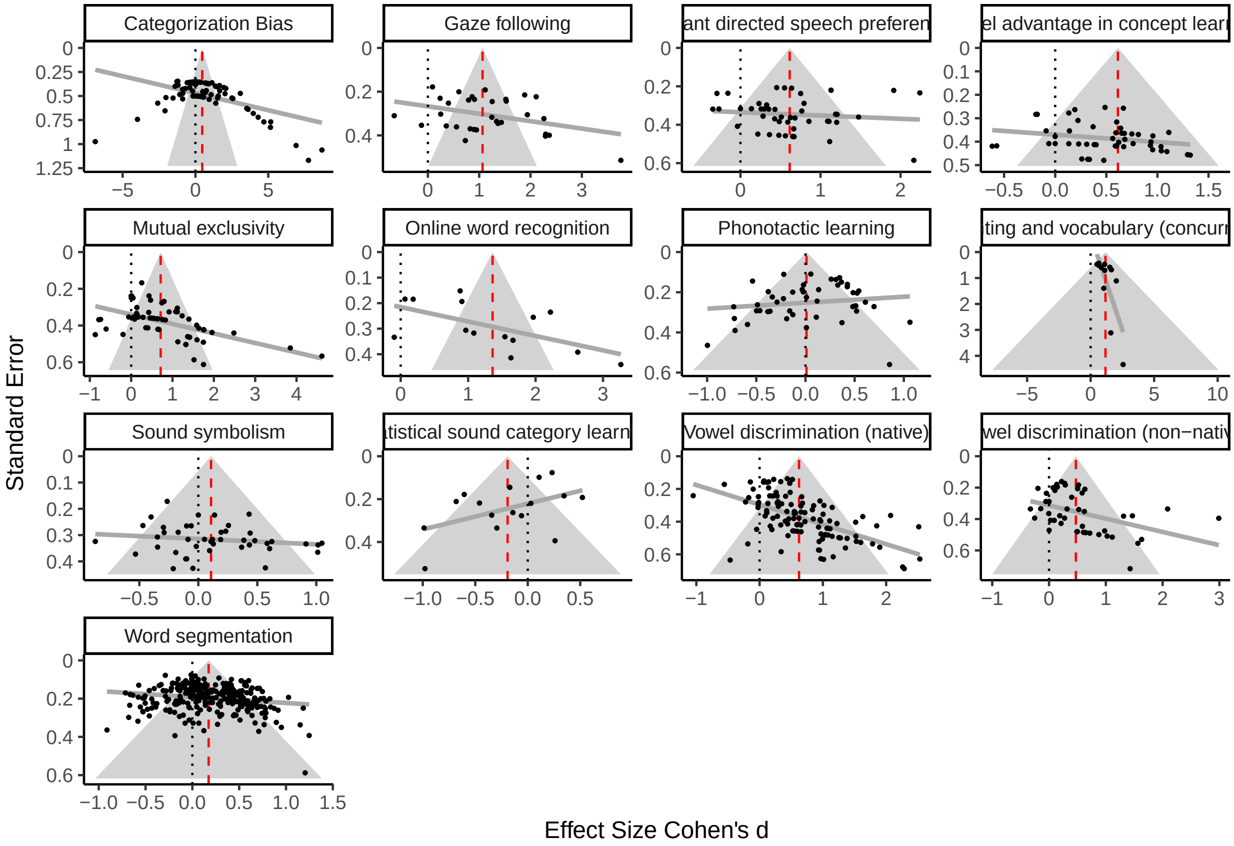


Figure 4: Funnel plots of all datasets with linear regression line indicated in grey. The dashed red line indicates the median effect size, the dotted black line zero. The grey triangle denotes the expected range area of effect sizes in the absence of bias or heterogeneity.

both whether there is an excess of p values just below the significance threshold and whether p values are distributed in a way that is or is not consistent with a true phenomenon being tested [@pcurve].

For analyses involving p values, we re-computed p values from our effect-size estimates. This is due to the following reasons: First, we did not have the same information available for all data points, even within the same meta-analysis. Second, exact p values are often not reported but rather as p<.05 and similar notations. In addition, two datasets only contain effect sizes, because they are based on extant meta-analyses. Finally, p values are not always computed, or, when they were computed, p values might not be reported correctly and/or consistently [@statcheck]. The recalculation pipeline is as follows: For papers where t values were not available, we transform Cohen's d into Pearson's r, from which it is possible to calculate a t value. p values were then computed accordingly from t values (either reported or re-calculated), taking into account the degrees of freedom of a specific experiment.

Figure 3 shows the distribution of p values below the significance threshold of .05. The bars indicate counts and the line plot represents density. Note that p values were recalculated based on effect sizes to ensure a consistent basis for our conclusions. For reliability purposes, we only discuss datasets with more than 10 p values between 0 and .05. In the absence of questionable research practices and the presence of an effect, we expect a distribution biased towards small values. In the absence of both p-hacking and an effect, the distribution should be flat, as all p values are equally likely to occur. Unexpected "bumps" towards higher p values in contrast can indicate severe p-hacking, including adding and removing samples and/or predictors, and conducting multiple statistical analyses [@Ioannidis2005; @Simmons2011].

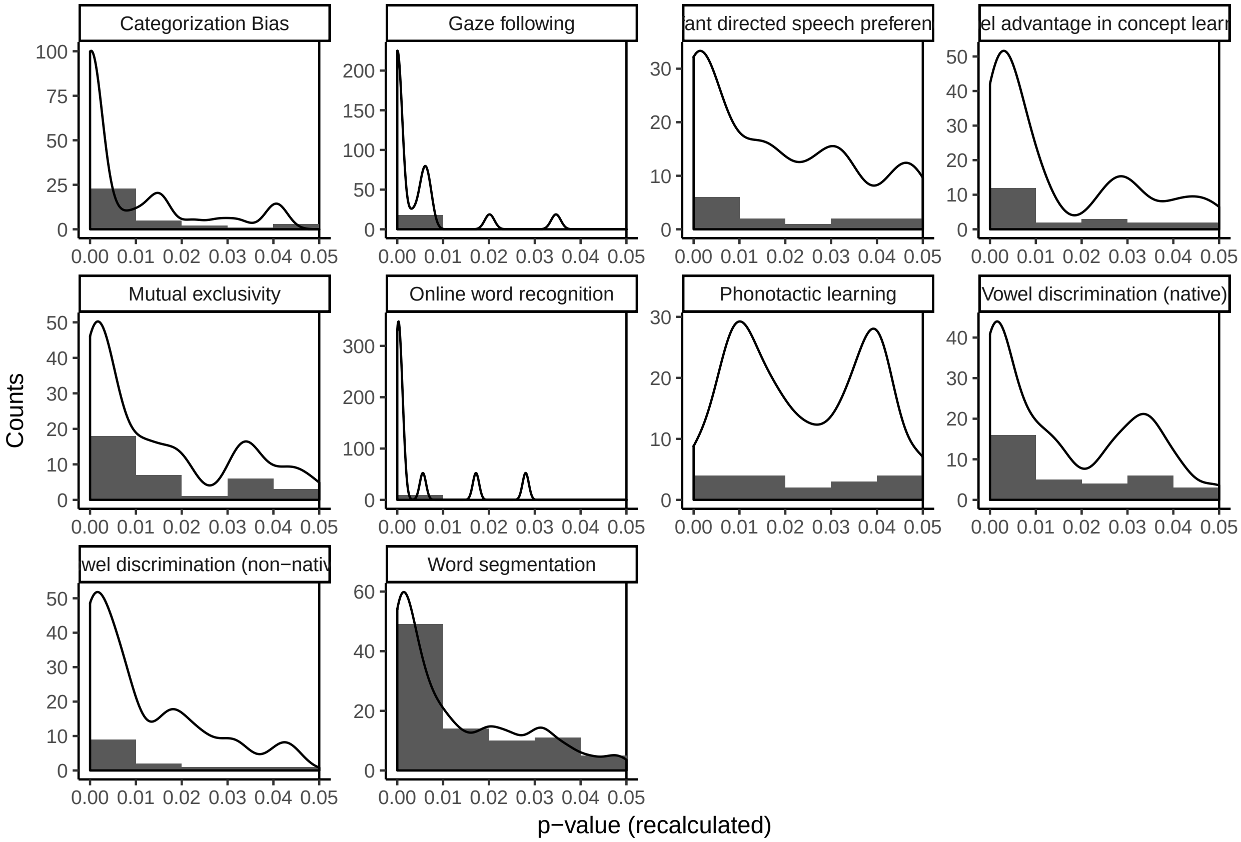


Figure 5: P-curves of all datasets with more than 10 significant (re-calculated) p values, the bars denote frequency, the lines density.

All of the datasets that could be included in this analysis display the expected right skew, but for some, p values just below .05 are more frequent than smaller ones between .02 and .03. For one dataset, phonotactic learning, this shape is particularly concerning. Further, the meta-analytic effect size points to an absence of an effect. Both observations have been made in the paper describing this meta-analysis in depth and are discussed there in more detail [@MetaPhon]. In all remaining cases the most frequent *p* values were the smallest, this is in line with the expected distribution assuming there is evidential value – an observation confirmed by the according statistical tests [see also @SynthesisPaper]. We thus conclude that the present meta-analyses largely reflect phenomena that are real and are based on reports that show no tangible symptoms of questionable research practices.