A Quantitative Synthesis of Early Language Acquisition Using Meta-Analysis

Supplementary Information

Molly Lewis, Mika Braginsky, Sho Tsuji, Christina Bergmann, Page Piccinini, Alejandrina Cristia, and Michael C. Frank

2017-01-24

Contents

This document was created from an R markdown file. Data from the paper can be interactively explored on the Metalab website, http://metalab.stanford.edu/. The manuscript itself was also produced from an R markdown file, and thus all analyses presented in the paper can be reproduced from that document. Supplementary materials can also be viewed online: http://rpubs.com/mll/synthesisSI.

Search strategies

Meta-analyses were conducted by the authors for all but two phenomena (IDS preference and Pointing and Vocabulary). Data for these two phenomena were obtained by adapting effect size estimates from existing, published meta-analyses (Dunst, Gorman, & Hamby, 2012; Colonnesi et al., 2010). Across phenomena, meta-analyses varied in their degree of systematicty in selecting papers. In the table below, we describe the search strategy for each phenomenon. Quoted descriptions are taken directly from the published source.

```
methods.table = datasets %>% select(name, search_strategy,
    internal_citation) %>% mutate(name = as.factor(name),
    name = plyr::revalue(name, c(`Infant directed speech preference` = "IDS preference",
        `Statistical sound category learning` = "Statistical sound learning",
        `Label advantage in concept learning` = "Concept-label advantage",
        `Vowel discrimination (native)` = "Native vowel discrimination",
        `Vowel discrimination (non-native)` = "Non-native vowel discrimination"))) %>%
    mutate(name = paste(name, internal_citation)) %>%
    select(-internal_citation) %>% .[c(1, 6, 4, 5,
    7, 8, 10, 12, 2, 9, 3, 11), ] %>% rename(Phenomenon = name,
    `Search Strategy` = search_strategy)
```

Phenomenon	Search Strategy
IDS preference (Dunst, Gorman, & Hamby, 2012)	"Studies were located using motherese or parentese or fatherese or infant directed speech or infant-directed speech or infant directed talk or child directed speech or child-directed speech or child directed talk or child-directed talk or baby talk AND infant* or neonate* or toddler* as search terms. Both controlled-vocabulary and natural-language searches were conducted (Lucas & Cutspec, 2007). Psychological Abstracts (PsychInfo), Educational Resource Information Center (ERIC), MEDLINE, Academic Search Premier, CINAHL, Education Resource Complete, and Dissertation Abstracts International were searched. These were supplemented by Google Scholar, Scirus, and Ingenta searches as well as a search of an extensive EndNote Library maintained by our Institute. Hand searches of the reference sections of all retrieved journal articles, book chapters, books, dissertations, and unpublished papers were also examined to locate additional studies. Studies were included if the effects of infant-directed speech on child behavior were compared to the effects of adult-directed speech on child behavior. Studies that intentionally manipulated word boundaries (e.g., Hirsh-Pasek et al., 1987; Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989) or used nonsense words or phrases (e.g., Mattys, Jusczyk, Luce, & Morgan, 1999; Thiessen, Hill, & Saffran, 2005) were excluded."
Phonotactic learning (Cristia, in prep.)	"Studies were considered based on a forward search from the seminal paper (MWG02) on both pubmed and google search, a list of references produced by the author, direct contact of labs having published on the topic, and announcements to several mailing lists."
Native vowel discrimination (Tsuji & Cristia, 2014)	"A full search on scholar.google.com was conducted in September 2012 with the keyword combination "{infant|infancy} & {vowel|speech sound|syllable} & discrimination." Additionally, the search terms were translated into French, German, Japanese, and Spanish for additional searches. We also asked experts in the field to inform us of any published or unpublished studies we had missed. Experts were defined as scientists having participated in at least two studies identified in our intermediate search sample or who were part of a lab where such research had taken place, and who were still active in the field or could be otherwise contacted. Further, articles were added based on a screening of articles cited and articles citing the articles in the remaining search sample. The complete sample is available as a public resource (Tsuji & Cristia, in preparation, https://sites.google.com/site/inphondb/). The search sample was then narrowed down to the final search sample of 19 articles." (See paper for additional details)
Non-native vowel discrimination (Tsuji & Cristia, 2014)	See Native Vowel Discrimination, above.
Statistical sound learning (Cristia, in prep.)	"Studies were considered based on a forward search from the seminal paper (MWG02) on both pubmed and google search, a list of references produced by the author, manual inspection of several editions of two leading conferences (ISIS and IASCL 2004-2012), direct contact of labs having published on the topic, and announcements to several mailing lists."
Word segmentation (Bergmann & Cristia, 2015)	"We first generated a list of potentially relevant items to be included in our meta-analysis using the Google Scholar search engine, with the broad search term 'infant word segmentation' (following Gehanno, Rollin & Darmoni, 2013). This search was carried out on 27 November 2012 and we inspected the first 1000 results. Fifteen additional items were included based on recommendations and by scanning references of included papers. After removing duplicates, we screened the title and abstract of each remaining item and identified 231 items for full-text inspection using the following inclusion criteria: (1) original data were reported; (2) the stimulus material was continuous natural speech spoken in the participants' native language; (3) the dependent measure was looking time (LT) at a neutral visual target (i.e. not a possible referent of one set of stimuli); (4) infants were normally developing."

Mutual exclusivity (Lewis & Frank, in prep.)	"We conducted a forward search based on citations of Markman and Wachtel (1988) in Google Scholar (September 2013). We also searched from papers using the keyword combination "mutual exclusivity" in both PsychInfo and Google Scholar. We identified additional papers that were cited from this initial list. From these, we identified a relevant subset using the following criteria: (a) monolingual child participants, (b) one familiar object present at test, (c) referents were objects (not facts or object parts), (d) no incongruent cues (e.g. eye gaze at familiar object), and (e) peer-reviewed. We also included a series of studies reported in Frank et al. (2016)."
Sound symbolism (Lammertink et al., 2016)	"We followed the PRISMA statement (Moher, Liberati, Tetzlaff, Altmann, & The PRISMA Group, 2009) for selecting and reporting on the studies to be included in our meta-analysis. We decided to include articles if they were assessing the on-line processing of sound-symbolically matching or mismatching sound-shape correspondences related to the bouba-kiki effect (thus, testing both 'round' and 'spiky' correspondences) in children up to and inclusive of the age of 3 years. 'Matching' responses refer to children's responses to congruent sound-shape associations (round word+round object; spiky word+spiky object) and 'mismatching' responses refer to children's responses to incongruent sound-shape associations (round word+spiky shape; spiky word+round shape), respectively. Since we were already aware of 10 published articles, conference presentations or conference proceedings papers that fit our inclusion criteria,and since we considered our strict inclusion this criteria to lead to a rather small selection of articles, we chose a seed strategy rather than a broad literature search. We began by assembling 4 key articles that fit the inclusion criteria (Asano et al., 2015; Maurer, Pathman & Mondloch, 2006; Ozturk, Krehm & Vouloumanos, 2012; Spector & Maurer, 2013) as well as two recent review papers on sound symbolism including infancy (Imai & Kita, 2014; Lockwood & Dingemanse, 2015). For all of these articles, we screened all potentially relevant references cited in these articles as well as references citing these articles and 'related articles' on their title and abstracts on scholar.google.com. Additionally, we screened titles and abstracts of all articles that cited one of the 4 key articles mentioned above (Asano et al. 2015: 16 citations; Maurer et al. 2006: 241 citations; Ozturk et al. 2012: 54 citations and Spector and Maurer, 2013: 9 citations). This search did not lead to additional eligible articles. In addition, we were aware of 9 conference presentations or conference proceedings papers that fit our inclusion criteria and were not redundant with one of our seed articles, including three by the two first authors of the present article."
Concept-label advantage (Lewis & Long, unpublished)	"We conducted a forward search based on Balaban and Waxman (1997) in Google Scholar and Web of Science (October 2015). This was supplemented with papers identified through citations and publication lists on lab websites. The final sample included only peer-reviewed publications."
Online word recognition (Frank, Lewis, & MacDonald, 2016)	"We conducted a systematic literature review by using Google Scholar to identify peer-reviewed papers citing Fernald et al. (1998). We screened this sample manually to find the subsample of 12 papers that reported both accuracy and reaction time with sufficient detail to permit coding."
Gaze following (Frank, Lewis, & MacDonald, 2016)	"We identified papers using a Google Scholar search for "gaze following" and included those studies that (a) included data from typically-developing children, (b) used a standard face-to-face gaze-following task, and (c) reported percentage accuracy (rather than a score or other composite measure). Although we coded all papers that fit these criteria, we focused on papers with a simple two-alternative forced choice (9 papers); integrating across different numbers of alternatives added additional complexity to our model. In our first iteration of this analysis, we found that very few studies reported reaction times for gaze following, and those that did had no data from children older than 15 months and no data from gaze plus pointing. To remedy this issue we include new analyses of data from Yurovsky, Wade, & Frank (2013) and Yurovsky & Frank (2015)."

Pointing and vocabulary (Colonnesi et al., 2010)	"The search method involved inspection of digital databases (Web of Knowledge, Picarta, PsychInfo) using the following keywords: pointing, gesture, declarative, imperative, precursors, language, words, vocabulary, infancy, intentional communication, and joint attention. Inspection of the reference section of relevant literature was an additional search method (ancestry method). Additionally, also unpublished sources were consulted, such as dissertations and presentations and studies under revision, by using Google Scholar, contacting researchers in the field and consulting digital databases of dissertations (e.g., PROQUEST). Three selection criteria were used to select studies: (a) measurement of infant production and/or comprehension of the pointing gesture; (b) measurement of language by assessing either receptive or expressive language; (c) report of a relation between pointing and language or the presence of data in the article allowing the calculation of a relation between pointing and language development. Exclusion criteria were: (a) subjects with mental or developmental disorders; (b) children older than 60 months; (c) studies in which the pointing gesture was not coded separately from other gestures."

Statistical approach

Effect sizes were computed by a script, compute_es.R, available in the Github repository. We calculated effect sizes from reported means and standard deviations where available, otherwise we relied on reported test-statistics (t and F). Several pre-existing MAs deal with special cases, and these are listed in the script. Except where noted, formulas are from Hedges & Olkin's textbook (2014). All analyses were conducted with the metafor package (Viechtbauer, 2010), using random-effects models. Note that a subset of individual MAs (such as Sound Symbolism) contain effect sizes that are not statistically independent, while the current implementation of random-effect models assumes independence of all individual effect sizes.

For many analyses, the use of a multi-level approach (with grouping by paper) is useful. We do not implement these models in our main analyses because many common statistics are not implemented for these models, e.g. the test for funnel-plot asymmetry. The table below compares the overall ES estimate for the multi-level model to the random effect model (presented in the main text). 95% confidence intervals are given in brackets. These models differ only slightly in their estimates of overall effect size, and in no case do they affect whether the ES differs from zero.

```
overall_es <- function(ma_data, multilevel) {
    if (multilevel) {
        model = metafor::rma.mv(d_calc, V = d_var_calc,
            random = ~1 | short_cite, data = ma_data)
    } else {
        model = metafor::rma(d_calc, d_var_calc, data = ma_data)
    }
    data.frame(dataset = ma_data$short_name[1], overall.d = model$b,
        ci_lower = model$ci.lb, ci_upper = model$ci.ub)
}

all_ds_random = all_data %>% split(.$short_name) %>%
    map(function(ma_data) overall_es(ma_data, 0)) %>%
    bind_rows() %>% mutate_each_(funs(round(., digits = 2)),
    vars = c("overall.d", "ci_lower", "ci_upper")) %>%
    mutate(d_string_random = paste0(overall.d, " [",
        ci_lower, ", ", ci_upper, "]")) %>% mutate(short_name = dataset) %>%
    select(short_name, d_string_random)

all_ds_multi = all_data %>% split(.$short_name) %>%
    map(function(ma_data) overall_es(ma_data, 1)) %>%
```

```
    bind_rows() %>% mutate_each_(funs(round(., digits = 2)),
    vars = c("overall.d", "ci_lower", "ci_upper")) %>%
    mutate(d_string_multi = paste0(overall.d, " [",
        ci_lower, ", ", ci_upper, "]")) %>% mutate(short_name = dataset) %>%
    select(short_name, d_string_multi)

left_join(all_ds_random, all_ds_multi) %>% left_join(select(datasets,
    name, short_name)) %>% select(name, d_string_random,
    d_string_multi) %>% mutate(name = as.factor(name),
    name = plyr::revalue(name, c(`Infant directed speech preference` = "IDS preference",
        `Statistical sound category learning` = "Statistical sound learning",
        `Label advantage in concept learning` = "Concept-label advantage",
        `Vowel discrimination (native)` = "Native vowel discrimination",
        `Vowel discrimination (non-native)` = "Non-native vowel discrimination"))) %>%
    .[c(2, 8, 3, 4, 10, 5, 7, 11, 6, 12, 1, 9), ] %>%
    kable(col.names = c("Phenomenon", "Random-effect model ES",
        "Mixed-effect model ES"), row.names = F, align = c("l",
        "r", "r"))
```

Phenomenon	Random-effect model ES	Mixed-effect model ES
IDS preference	0.7 [0.52, 0.88]	0.74 [0.47, 1.01]
Phonotactic learning	0.04 [-0.09, 0.16]	0.12 [-0.01, 0.25]
Native vowel discrimination	0.68 [0.56, 0.81]	0.7 [0.51, 0.89]
Non-native vowel discrimination	0.66 [0.42, 0.9]	1 [0.41, 1.59]
Statistical sound learning	-0.19 [-0.42, 0.03]	-0.26 [-0.58, 0.05]
Word segmentation	0.19 [0.14, 0.23]	0.16 [0.11, 0.21]
Mutual exclusivity	1.01 [0.68, 1.33]	0.82 [0.54, 1.1]
Sound symbolism	0.12 [-0.02, 0.25]	0.22 [0, 0.44]
Concept-label advantage	0.47 [0.33, 0.61]	0.41 [0.26, 0.57]
Online word recognition	1.36 [0.84, 1.88]	1.24 [0.74, 1.74]
Gaze following	1.27 [0.93, 1.61]	1.17 [0.8, 1.55]
Pointing and vocabulary	0.98 [0.62, 1.34]	0.98 [0.62, 1.34]

Funnel plots

```
funnel.es.with.outliers = all_data %>% mutate(dataset = as.factor(dataset),
    dataset = gdata::reorder.factor(dataset, new.order = c(2,
        6, 10, 11, 9, 12, 4, 8, 3, 5, 1, 7)), dataset = plyr::revalue(dataset,
        c(`Infant directed speech preference` = "IDS preference",
            `Statistical sound category learning` = "Statistical sound learning",
            `Label advantage in concept learning` = "Concept-label advantage",
            `Vowel discrimination (native)` = "Vowel discrimination\n(native)",
            `Vowel discrimination (non-native)` = "Vowel discrimination\n(non-native)"))) %>%
    group_by(dataset) %>% mutate(outlier = ifelse(d_calc >
    mean(d_calc) + (3 * sd(d_calc)) | d_calc < mean(d_calc) -
    (3 * sd(d_calc)), 1, 0), outlier = as.factor(outlier))
```

If an effect size is an extreme outlier from the overall mean, this may indicate that the effect size estimates a different psychological phenomenon than the one estimated by others in the sample. One approach for dealing with this type of heterogenity is to exclude outliers from analyses. Here we present funnel plots that

exclude effect sizes that lie 3 standard deviations above or below the mean effect size for each meta-analysis. Of the 772 effect sizes in the dataset, 7 were outliers (0.9%).

```
CRIT_95 = 1.96

funnel.es.data = funnel.es.with.outliers %>% filter(outlier ==
    0) %>% mutate(se = sqrt(d_var_calc), es = d_calc,
    center = mean(d_calc), lower_lim = max(se) + 0.05 *
        max(se))

# separate df for 95 CI funnel shape
funnel95.data.wide <- funnel.es.data %>% select(center,
    lower_lim, dataset) %>% group_by(dataset) %>% summarise(x1 = (center -
    lower_lim * CRIT_95)[1], x2 = center[1], x3 = center[1] +
    lower_lim[1] * CRIT_95, y1 = -lower_lim[1], y2 = 0,
    y3 = -lower_lim[1])

funnel95.data.x = funnel95.data.wide %>% select(dataset,
    dplyr::contains("x")) %>% gather("coordx", "x",
    2:4) %>% arrange(dataset, coordx) %>% select(-coordx)

funnel95.data.y = funnel95.data.wide %>% select(dataset,
    dplyr::contains("y")) %>% gather("coordy", "y",
    2:4) %>% arrange(dataset, coordy) %>% select(-coordy)

funnel95.data = bind_cols(funnel95.data.x, funnel95.data.y)

ggplot(funnel.es.data, aes(x = es, y = -se)) + facet_wrap(~dataset,
    scales = "free") + xlab("Effect Size") + ylab("Standard Error\n") +
    scale_colour_solarized(name = "") + geom_polygon(aes(x = x,
    y = y), data = funnel95.data, fill = "white") +
    geom_vline(aes(xintercept = x2), linetype = "dashed",
        color = "red", size = 0.8, data = funnel95.data.wide) +
    geom_vline(xintercept = 0, linetype = "dashed",
        color = "grey44", size = 0.8) + scale_y_continuous(labels = function(x) {
    abs(x)
}) + geom_point(size = 0.5) + theme(panel.grid.major = element_line(colour = "grey",
    size = 0.2), panel.grid.minor = element_line(colour = "grey",
    size = 0.5), strip.text.x = element_text(size = 9),
    strip.background = element_rect(fill = "grey"))
```

We next compare the results of funnel skew (Egger's test) for the dataset with outliers excluded to the full dataset (which is reported in the main text). There is a difference in significance for only Statistical Sound Learning: With outliers excluded, these meta-analyses no longer show evidence for skew.

```
eggers_tests <- function(ma_data){
    model = metafor::rma(d_calc, d_var_calc, data = ma_data) # model
    egg.random = metafor::regtest(model) # Egger's test
    data.frame(dataset = ma_data$short_name[1],
               egg.random.z = egg.random$zval,
               egg.random.p = egg.random$pval)
}

eggers.data.f = all_data %>%
```

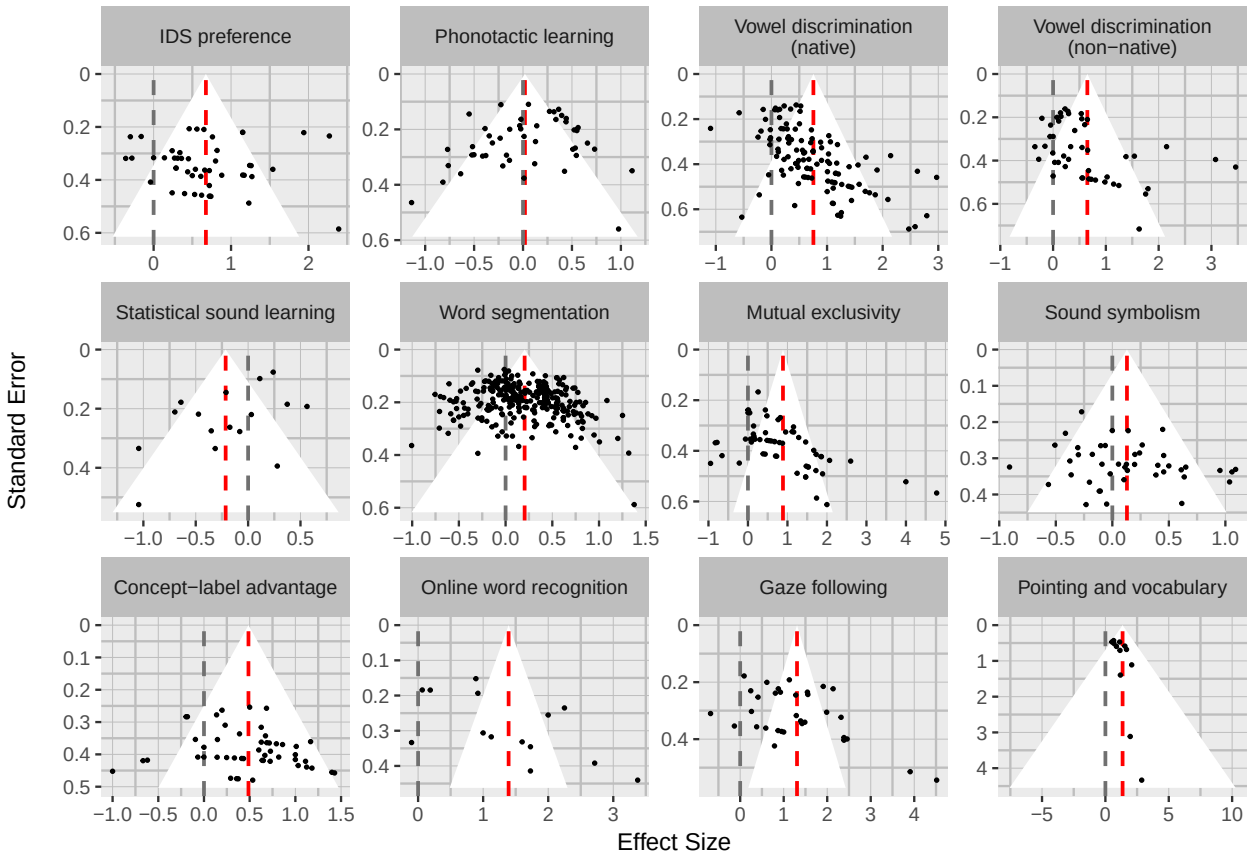


Figure S1: Funnel plots for each meta-analysis with outliers excluded. Each effect size estimate is represented by a point, and the mean effect size is shown as a red dashed line. The grey dashed line shows an effect size of zero. The funnel corresponds to a 95% CI around this mean.

```
  group_by(dataset) %>%
  mutate(outlier = ifelse(d_calc > mean(d_calc)  + (3 * sd(d_calc)) |
                            d_calc < mean(d_calc)  - (3 * sd(d_calc)), 1, 0),
         outlier = as.factor(outlier)) %>%
  filter(outlier == 0) %>%
  ungroup() %>%
  split(.$short_name) %>%
  map(function(ma_data) eggers_tests(ma_data)) %>%
  bind_rows() %>%
  mutate(egg.random.z = round(egg.random.z, digits = 2)) %>%
  mutate(egg.random.p = round(egg.random.p, digits = 2)) %>%
  mutate(egg_string.f = paste0(egg.random.z, " (", egg.random.p, ")")) %>%
  select(dataset, egg_string.f)

eggers.data.all = all_data %>%
  group_by(dataset) %>%
  ungroup() %>%
  split(.$short_name) %>%
  map(function(ma_data) eggers_tests(ma_data)) %>%
  bind_rows() %>%
  mutate(egg.random.z = round(egg.random.z, digits = 2)) %>%
  mutate(egg.random.p = round(egg.random.p, digits = 2)) %>%
  mutate(egg_string.all = paste0(egg.random.z, " (", egg.random.p, ")")) %>%
  select(dataset, egg_string.all)

left_join(eggers.data.all, eggers.data.f) %>%
              ungroup() %>%
             .[c(2,8,3,4,10,5,7,11,6,12,1,9),]  %>% # reorder rows
              left_join(select(datasets, name, short_name),
                        by=c("dataset" = "short_name" )) %>%
              select(-dataset) %>%
              mutate(dataset = as.factor(name),
                     dataset = plyr::revalue(dataset,
                     c("Infant directed speech preference" ="IDS preference",
                     "Statistical sound category learning"= "Statistical sound learning",
                     "Label advantage in concept learning" =  "Concept-label advantage"))) %>%
              mutate(egg_string.f = sub("(0)", "(< .01)", egg_string.f, fixed = T),
                     egg_string.all = sub("(0)", "(< .01)", egg_string.all, fixed = T)) %>%
  select(dataset, egg_string.all, egg_string.f) %>%
  kable(col.names = c("Phenomenon", "funnel skew (all conditions)",
                      "funnel skew (excluding outliers)"),
        align = c("l", "r", "r"))
```

Phenomenon	funnel skew (all conditions)	funnel skew (excluding outliers)
IDS preference	1.5 (0.13)	0.46 (0.65)
Phonotactic learning	-1.43 (0.15)	-1.43 (0.15)
Vowel discrimination (native)	8.55 (< .01)	8.18 (< .01)
Vowel discrimination (non-native)	3.86 (< .01)	3.24 (< .01)
Statistical sound learning	-2.99 (< .01)	-1.89 (0.06)
Word segmentation	2.59 (0.01)	2.8 (0.01)
Mutual exclusivity	8.26 (< .01)	5.2 (< .01)
Sound symbolism	1.42 (0.16)	1.42 (0.16)
Concept-label advantage	1.37 (0.17)	1.37 (0.17)

Phenomenon	funnel skew (all conditions)	funnel skew (excluding outliers)
Online word recognition	2.61 (0.01)	2.61 (0.01)
Gaze following	3.3 (< .01)	3.3 (< .01)
Pointing and vocabulary	1.25 (0.21)	1.25 (0.21)

P-curves

When available, we calculated p-values based on test statistics reported in the paper. However, when unavailable, we calculated p-values based on raw descriptive statistics (means and standard deviations) or reported effect sizes (the method used for IDS Preference). The main text shows the results of the p-curve analysis based on p-values derived by both approaches. Here, we compare these results to the same analysis on the subset of p-values derived only from reported test statistics. Presented below are p-curves calculated from this subset.

```
ALPHA = 0.05
P_INCREMENT = 0.01

pc.data <- get_all_pc_data(all_data, ALPHA, P_INCREMENT,
    transform = FALSE)

p.source = pc.data %>% select(f.transform, f.value,
    dataset, study_ID, p_round) %>% group_by(dataset) %>%
    summarise(n.total = n(), n.transform = length(which(!is.na(f.transform))),
        sig.p = length(which(p_round < ALPHA))) %>%
    mutate(dataset = plyr::revalue(dataset, c(`Infant directed speech preference` = "IDS preference",
        `Statistical sound category learning` = "Statistical sound learning",
        `Label advantage in concept learning` = "Concept-label advantage",
        `Vowel discrimination (native)` = "Vowel discrimination\n(native)",
        `Vowel discrimination (non-native)` = "Vowel discrimination\n(non-native)")),
        dataset = as.factor(dataset), dataset = gdata::reorder.factor(dataset,
            new.order = c(2, 6, 10, 11, 9, 12, 4, 8,
                3, 5, 1, 7))) %>% mutate(stat_only = ifelse(n.total >
    n.transform, 1, 0)) %>% arrange(-stat_only) %>%
    mutate(prop.ts = 1 - n.transform/n.total, prop.ts.string = as.character(round(prop.ts,
        digits = 2))) %>% as.data.frame()

get.all.CIS.multi <- function(df) {
    ps <- seq(P_INCREMENT, ALPHA, P_INCREMENT)
    props = ps %>% map(function(p, d) {
        sum(d == p)
    }, df$p_round) %>% unlist()
    cis = MultinomialCI::multinomialCI(props, alpha = ALPHA)
    data.frame(dataset = df$dataset[1], p = ps, ci.lower = cis[,
        1], ci.upper = cis[, 2])
}

ci.data = pc.data %>% split(.$dataset) %>% map(function(data) get.all.CIS.multi(data)) %>%
    bind_rows() %>% mutate(dataset = as.factor(dataset),
    dataset = plyr::revalue(dataset, c(`Infant directed speech preference` = "IDS preference",
        `Statistical sound category learning` = "Statistical sound learning",
        `Label advantage in concept learning` = "Concept-label advantage",
        `Vowel discrimination (native)` = "Vowel discrimination\n(native)",
```

```
        `Vowel discrimination (non-native)` = "Vowel discrimination\n(non-native)")))

ci.data[ci.data$dataset == "Sound symbolism" & ci.data$p ==
    0.01, "ci.lower"] = 0  # there's only one datapoint for this dataset

pc.data %>% group_by(dataset) %>% do(get_p_curve_df(.,
    ALPHA, P_INCREMENT)) %>% ungroup() %>% mutate(dataset = as.factor(dataset),
    dataset = plyr::revalue(dataset, c(`Statistical sound category learning` = "Statistical sound learni
        `Label advantage in concept learning` = "Concept-label advantage",
        `Vowel discrimination (native)` = "Vowel discrimination\n(native)",
        `Vowel discrimination (non-native)` = "Vowel discrimination\n(non-native)")),
    dataset = gdata::reorder.factor(dataset, new.order = c(3,
        7, 8, 6, 9, 2, 5, 1, 4))) %>% ggplot() + facet_wrap(~dataset,
    nrow = 3) + geom_ribbon(aes(ymin = ci.lower, ymax = ci.upper,
    x = p), fill = "grey87", data = ci.data) + geom_line(size = 1,
    aes(x = p, y = value, linetype = measure, color = measure)) +
    scale_colour_manual(name = "", values = c("red",
        "green", "blue"), labels = c("Null of no effect",
        "Null of 33% power", "Observed")) + scale_linetype_manual(values = c("dashed",
    "dashed", "solid"), guide = FALSE) + ylab("Proportion p-values\n") +
    xlab("p-value") + geom_text(aes(label = paste("prop. test stat. = ",
    prop.ts.string, "\nnum. sig. ps = ", sig.p), x = 0.028,
    y = 0.8), data = p.source, colour = "black", size = 2,
    hjust = 0) + theme_bw() + theme(legend.position = "top",
    legend.key = element_blank(), legend.background = element_rect(fill = "transparent"),
    strip.text.x = element_text(size = 9), axis.title = element_text(colour = "black",
        size = 12), panel.margin = unit(0.65, "lines"),
    strip.background = element_rect(fill = "grey"))
```

We next compare the test of right-skew presented in the main text for both the full set of p-values and those only derived from test statistics. In no case does the significance of the test differ between the two analyses.

```
stouffer.data = pc.data %>% group_by(dataset) %>% do(data.frame(stouffer = stouffer_test(.,
    ALPHA))) %>% filter(stouffer.pp.measure == "ppr.full") %>%
    full_join(datasets %>% select(name, short_name),
        by = c(dataset = "name")) %>% select(short_name,
    stouffer.Z.pp, stouffer.p.Z.pp) %>% mutate_each_(funs(round(.,
    digits = 2)), vars = c("stouffer.p.Z.pp", "stouffer.Z.pp")) %>%
    mutate(stouff_string = ifelse(is.na(as.character(stouffer.Z.pp)),
        "", paste0(stouffer.Z.pp, " (", stouffer.p.Z.pp,
            ")"))) %>% mutate(stouff_string = sub("(0)",
    "(< .01)", stouff_string, fixed = T)) %>% select(dataset,
    stouff_string)

stouffer.data_all = get_all_pc_data(all_data, ALPHA,
    P_INCREMENT, transform = TRUE) %>% group_by(dataset) %>%
    do(data.frame(stouffer = stouffer_test(., ALPHA))) %>%
    filter(stouffer.pp.measure == "ppr.full") %>% full_join(datasets %>%
    select(name, short_name), by = c(dataset = "name")) %>%
    select(short_name, stouffer.Z.pp, stouffer.p.Z.pp) %>%
    mutate_each_(funs(round(., digits = 2)), vars = c("stouffer.p.Z.pp",
        "stouffer.Z.pp")) %>% mutate(stouff_string = ifelse(is.na(as.character(stouffer.Z.pp)),
    "", paste0(stouffer.Z.pp, " (", stouffer.p.Z.pp,
```

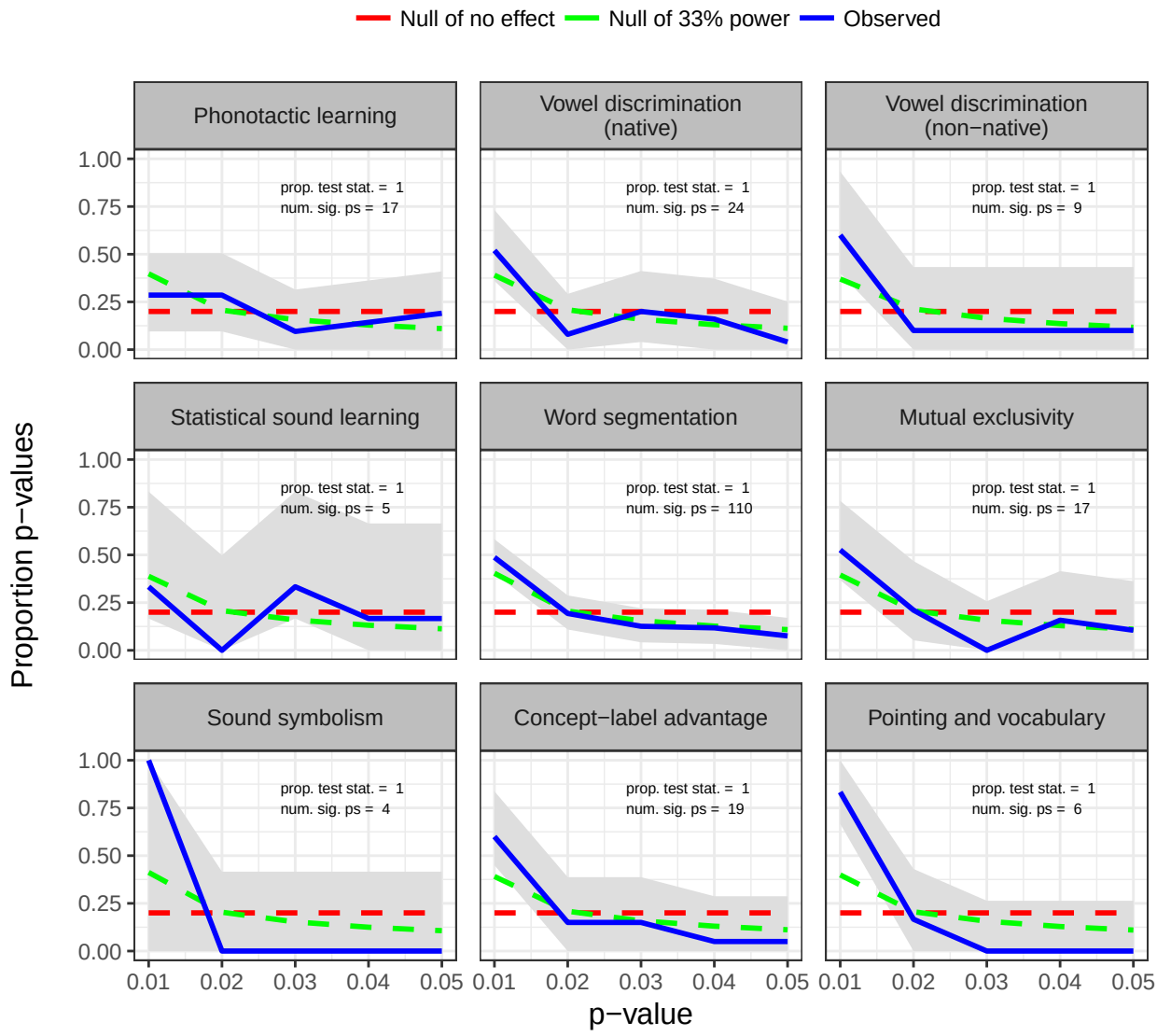


Figure S2: In the main text, we calculate p-curves based on all conditions in the dataset. In cases where a p-value was not directly available from the reported test statistic, we calculated a p-value based on a significance test using the reported means and standard deviations. The table compares the test of right-skew (Stouffer method) for this full dataset, as reported in the main text, to the subset of conditions for which p-values were directly available. Error bars are 95% confidence intervals calculated from a multinomial distribution.

```
        ")"))) %>% mutate(stouff_string = sub("(0)",
    "(< .01)", stouff_string, fixed = T)) %>% select(dataset,
    stouff_string)

# p-curve data using from all conditions (same as
# reported in paper)
pc.data.all <- get_all_pc_data(all_data, ALPHA, P_INCREMENT,
    transform = TRUE)

left_join(stouffer.data_all, stouffer.data, by = "dataset") %>%
    .[c(2, 6, 10, 11, 9, 12, 4, 8, 3, 5, 1, 7), ] %>%

kable(col.names = c("Phenomenon", "p-curve skew  (all conditions)",
    "p-curve skew (p-values only from test-statistics)"),
    align = c("l", "r", "r"))
```

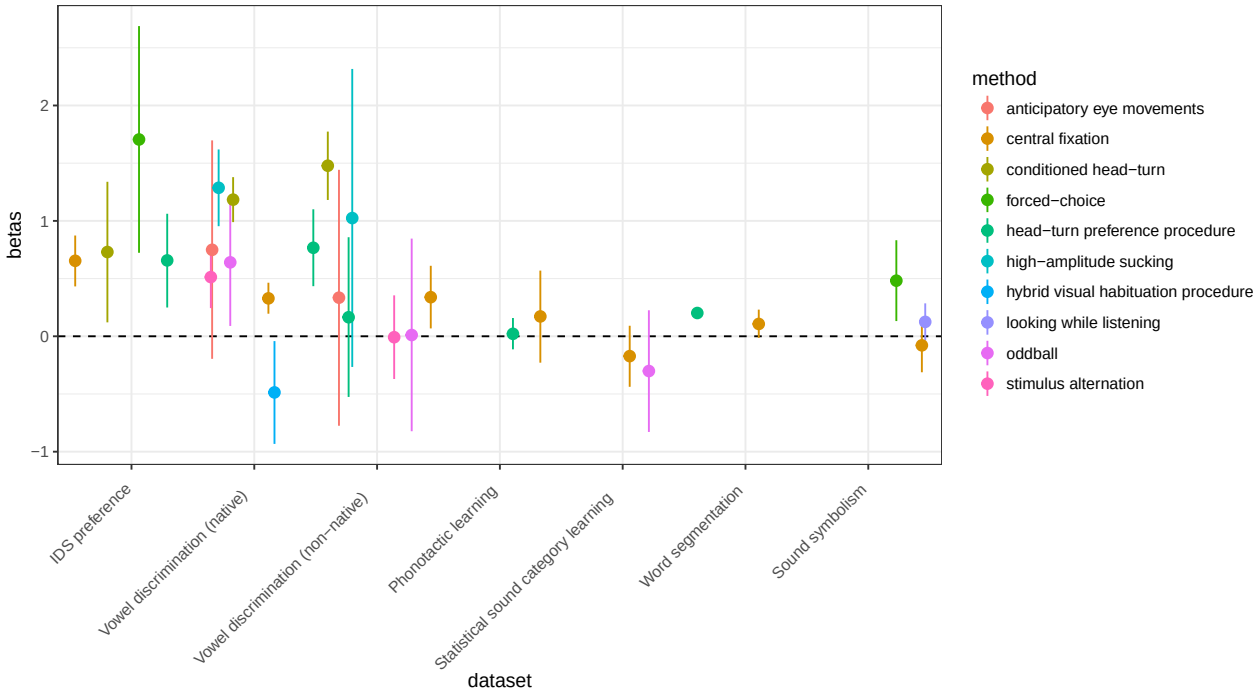
Phenomenon	p-curve skew (all conditions)	p-curve skew (p-values only from test-statistics)
Infant directed speech preference	-10.4 (< .01)	
Phonotactic learning	-1.52 (0.06)	-1.52 (0.06)
Vowel discrimination (native)	-9.76 (< .01)	-5.14 (< .01)
Vowel discrimination (non-native)	-8.89 (< .01)	-3.24 (< .01)
Statistical sound category learning	-1.03 (0.15)	-0.65 (0.26)
Word segmentation	-9.4 (< .01)	-9.82 (< .01)
Mutual exclusivity	-12.87 (< .01)	-5 (< .01)
Sound symbolism	-5.56 (< .01)	-5.1 (< .01)
Label advantage in concept learning	-4.79 (< .01)	-4.54 (< .01)
Online word recognition	-14.51 (< .01)	
Gaze following	-18.66 (< .01)	
Pointing and vocabulary	-6.33 (< .01)	-6.33 (< .01)

Method heterogeneity

The plot below presents model coefficients for method for datasets with more than one method. Coefficients are estimated from random-effects meta-analytic models. For the most part, we see that method only has a small influence on the effect size within a given phenomenon. There are exceptions, however: For example, for Sound Symbolism the forced choice method has an overall larger effect size than other methods.

```
single_method_datasets = all_data %>% group_by(dataset) %>%
    summarise(n_methods = length(levels(as.factor(method)))) %>%
    filter(n_methods == 1) %>% .[["dataset"]]

method.betas = data.frame()
for (i in 1:length(datasets$name)) {

    if (!(datasets$name[i] %in% single_method_datasets)) {
        d = filter(all_data, dataset == datasets$name[i])
        model = metafor::rma(d_calc ~ method - 1, vi = d_var_calc,
            data = d, method = "REML")

        d = data.frame(dataset = datasets$name[i],
            method = row.names(model$b), betas = model$b,
```

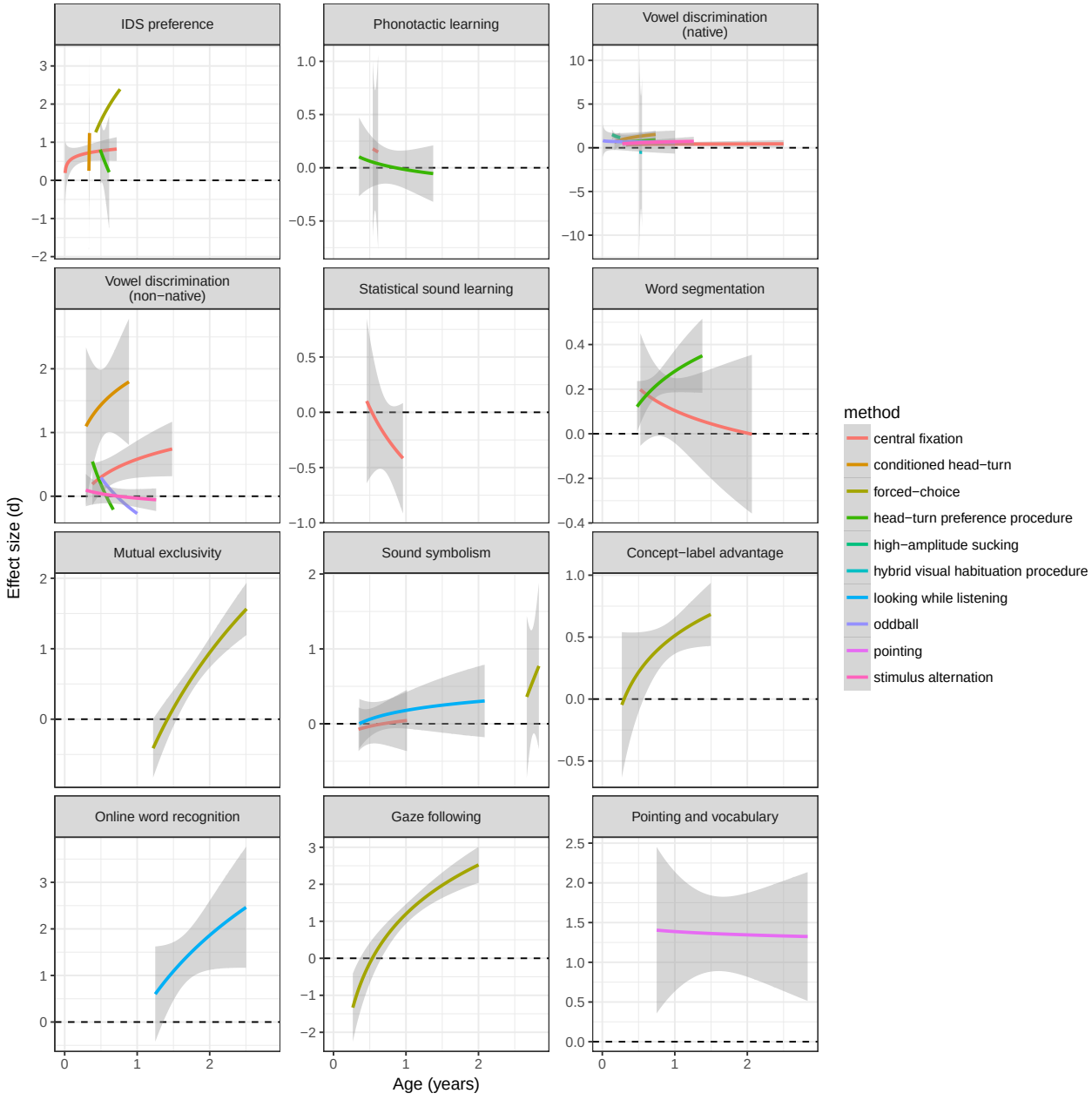
```
            ci.lb = model$ci.lb, ci.ub = model$ci.ub,
            row.names = NULL)

        method.betas = rbind(method.betas, d)
    }
}

method.betas = method.betas %>% mutate(dataset = as.factor(dataset),
    dataset = plyr::revalue(dataset, c(`Infant directed speech preference` = "IDS preference")),
    method = gsub("method", "", method))

ggplot(method.betas, aes(x = dataset, y = betas, ymin = ci.lb,
    ymax = ci.ub, color = method)) + geom_hline(aes(yintercept = 0),
    linetype = "dashed") + geom_pointrange(position = position_jitter(0.5)) +
    theme_bw() + theme(axis.text.x = element_text(angle = 45,
    hjust = 1.1))
```



The plot below presents the developmental trajectory of each phenomenon, with a separate color for each method. Lines show log-linear model fits. Word Segmentation shows the most notable interaction between age and method: Effect sizes increase with age for head-turn preference procedure, but decrease for central fixation.

```
all_data %>% filter(mean_age_1/365 < 3) %>% mutate(dataset = as.factor(dataset),
    dataset = gdata::reorder.factor(dataset, new.order = c(2,
        6, 10, 11, 9, 12, 4, 8, 3, 5, 1, 7)), dataset = plyr::revalue(dataset,
        c(`Infant directed speech preference` = "IDS preference",
            `Statistical sound category learning` = "Statistical sound learning",
            `Label advantage in concept learning` = "Concept-label advantage",
            `Vowel discrimination (native)` = "Vowel discrimination\n(native)",
            `Vowel discrimination (non-native)` = "Vowel discrimination\n(non-native)"))) %>%
    ggplot(aes(x = mean_age_1/365, y = d_calc, color = method)) +
```

```
    geom_hline(yintercept = 0, linetype = "dashed",
        color = "black") + facet_wrap(~dataset, scales = "free_y",
    ncol = 3) + geom_smooth(method = "lm", formula = y ~
    log(x)) + xlab("Age (years)") + ylab("Effect size (d)") +
    theme_bw() + theme(legend.position = "right", legend.key = element_blank(),
    legend.background = element_rect(fill = "transparent"))
```



References

Bergmann, C., & Cristia, A. (in press). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*.

Colonnesi, C., Stamsa, G. J., Kostera, I., & Noomb, M. J. (2010). The relation between pointing and language development: A meta-analysis. *Developmental Review*, 30, 352–366.

Cristia, A. (in prep.). Infants' phonology learning in the lab.

Dunst, C. J., Gorman, E., & Hamby, D. W. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, *5*(1).

Frank, M. C., Lewis, M., & MacDonald, K. (2016). A performance model for early wordlearning. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society.*

Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis.* Academic Press.

Lammertink, I., Fort, M., Peperkamp, S., Fikkert, P., Guevara-Rukoz, A., & Tsuji, S. (2016). SymBuki: A meta-analysis on the sound-symbolic bouba-kiki effect in infants and toddlers. Poster presented at the XXI Biennial International Congress of Infant Studies, New Orleans, USA.

Lewis, M. & Frank, M. (in prep.). Mutual exclusivity: A meta-analysis.

Lewis, M. & Long, B. (unpublished). A meta-analysis of the concept-label advantage.

Tsuji, S. & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*(2), 179-191.

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