Quantifying the syntactic bootstrapping effect in verb learning: A meta-analytic synthesis

Supplementary Information

XXXX and XXXX

2021-07-16

Contents

This document was created from an R markdown file. The repository for the project can be found at XXXXXXXXXX. The data reported in the paper can be explored interactively at the Metalab website.

Main models results using dataset without imputed values

For one paper missing relevant sufficient data to calculate an effect size (Hirsh-Pasek, Golinkoff, & Naigles, 1996), we imputed values from studies with similar design. The tables below report the model results from fitting the exact same models on the dataset excluding the imputed study. The two sets of models did not differ qualitatively.

Predicate Type

Parameter	Estimate	z value	p value
Intercept	0.1 [-0.15, 0.35]	0.76	0.45
Predicate type (Intransitive / Transitive)	0.24 [0.01, 0.46]	2.08	0.04*

Noun phrase type

Parameter	Estimate	z value	p value
Intercept	0.25 [-0.04, 0.54]	1.68	0.09
Noun phrase type (Noun / Pronoun)	-0.04 [-0.4, 0.33]	-0.2	0.84

Character identification phase

Parameter	Estimate	z value	p value
Intercept	0.17 [-0.08, 0.43]	1.34	0.18
Character identification phase (No / Yes)	0.19 [-0.28, 0.66]	0.78	0.43

Practice phase

Parameter	Estimate	z value	p value
Intercept	0.35 [0.07, 0.63]	2.44	0.01*
Practice phase (No / Yes)	-0.21 [-0.51, 0.1]	-1.31	0.19

Synchronicity

Parameter	Estimate	z value	p value
Intercept	0.2 [-0.05, 0.44]	1.54	0.12
Synchronicity (Simultaneous / Asynchronous)	0.09 [-0.24, 0.41]	0.53	0.59

Testing structure

Parameter	Estimate	z value	p value
Intercept	0.11 [-0.12, 0.35]	0.93	0.35
Testing Procedure Structure (Distributed / Mass)	0.39 [-0.04, 0.82]	1.78	0.07

Number of sentence repetitions

Parameter	Estimate	z value	p value
Intercept	0.17 [-0.14, 0.48]	1.05	0.29
Number of sentence repetitions	0.01 [-0.02, 0.03]	0.55	0.58

Details of effect size calculation

Here we demonstrate our method for calculating effect sizes by describing a step-by-step calculation for conditions in an example paper, Yuan & Fisher (2009). The table below shows the original data reported in the source paper (Table 1, pg 622 of Yuan & Fisher, 2009). The values are mean looking time in seconds (and corresponding SE).

Dialogue Type	Sample Size	Two-participant Event	One-participant Event
Transitive	8	4.82 (0.43)	2.87 (0.51)
Intransitive	8	3.33 (0.24)	4.12 (0.40)

To standardize the effect size calculation, we converted the reported raw results to the proportion of correct responses. For looking time studies, when the paper only reported the raw looking time in seconds, we

calculated the proportion of correct response by dividing the mean looking time toward the matching scene by the sum of looking time toward the matching scenes and non-matching scenes (i.e., excluding the look away time from the denominator). For children hearing transitive sentences, the correct scene was the two-participant event; for children hearing intransitive sentences, the correct scene was the one-participant event. Standard errors were converted using a similar method.

Using these standardize we calculated Cohen's d and the variances as follows (the implementation of the script can be found at XXXXXXXXXXXXXXXXXXXXXXXXXXX).

$$
\begin{aligned}
Mean_{transitive} &= \frac{Time_{correct}}{Time_{correct} + Time_{incorrect}} && (1)\\
&= \frac{4.82}{4.82 + 2.87} && (2)\\
&= 0.627 && (3)\\
SD_{transitive} &= \frac{SE_{Raw}}{Time_{correct} + Time_{incorrect}} * \sqrt[2]{N} && (4)\\
&= \frac{0.43}{4.82 + 2.87} * \sqrt[2]{8} && (5)\\
&= 0.158 && (6)\\
& && (7)
\end{aligned}
$$

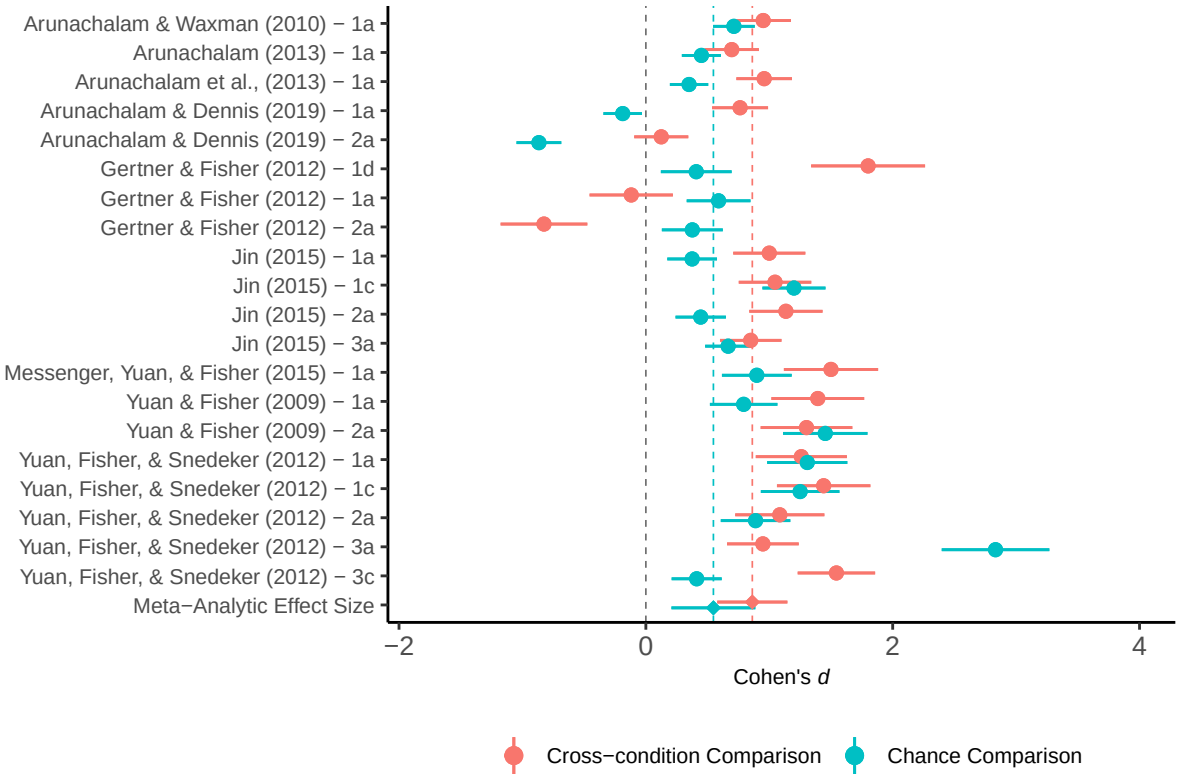
$$
\begin{aligned}
Mean_{intransitive} &= \frac{Time_{correct}}{Time_{correct} + Time_{incorrect}} && (8)\\
&= \frac{4.12}{3.33 + 4.12} && (9)\\
&= 0.553 && (10)\\
SD_{intransitive} &= \frac{SE_{Raw}}{Time_{correct} + Time_{incorrect}} * \sqrt[2]{N} && (11)\\
&= \frac{0.4}{3.33 + 4.12} * \sqrt[2]{8} && (12)\\
&= 0.152 && (13)\\
& && (14)
\end{aligned}
$$

$$
\begin{aligned}
d_{transitive} &= \frac{M_1 - M_2}{\sigma_{pooled}} && (15)\\
&= \frac{M_{correct} - M_{chance}}{\sigma_{correct}} && (16)\\
&= \frac{0.627 - 0.5}{0.158} && (17)\\
&\approx 0.79 && (18)\\
d_{intransitive} &= \frac{M_1 - M_2}{\sigma_{pooled}} && (19)\\
&= \frac{M_{correct} - M_{chance}}{\sigma_{correct}} && (20)\\
&= \frac{0.553 - 0.5}{0.152} && (21)\\
&\approx 0.35 && (22)
\end{aligned}
$$

$$var(d_{transitive}) = \frac{1}{N} + \frac{d^2}{2 * N} \tag{23}$$
$$= \frac{1}{8} + \frac{0.79^2}{2 * 8} \tag{24}$$
$$\approx 0.16 \tag{25}$$
$$var(d_{intransitive}) = \frac{1}{N} + \frac{d^2}{2 * N} \tag{26}$$
$$= \frac{1}{8} + \frac{0.35^2}{2 * 8} \tag{27}$$
$$\approx 0.13 \tag{28}$$
$$\tag{29}$$

Cross-condition Comparison vs. Chance Comparison effect size estimates

The forest plot below compares the two ways of calculating effect sizes for the subset of experimental conditions that use a between-group analysis in the original paper. In other words, the original analyses compared the proportion of looking time at the causative events between transitive conditions and intransitive conditions. Effect sizes calculated using this method are denoted by the pink points. We also present the effect sizes calculated using the against-chance method on the same subset of the experimental conditions. These effect sizes are denoted with green points. The against-chance method is a more conservative way of estimating the effect size. As the forest plot shows, the meta-analytic effect size using the between-group calculation is larger than the meta-analytic effect size using the against-chance method.



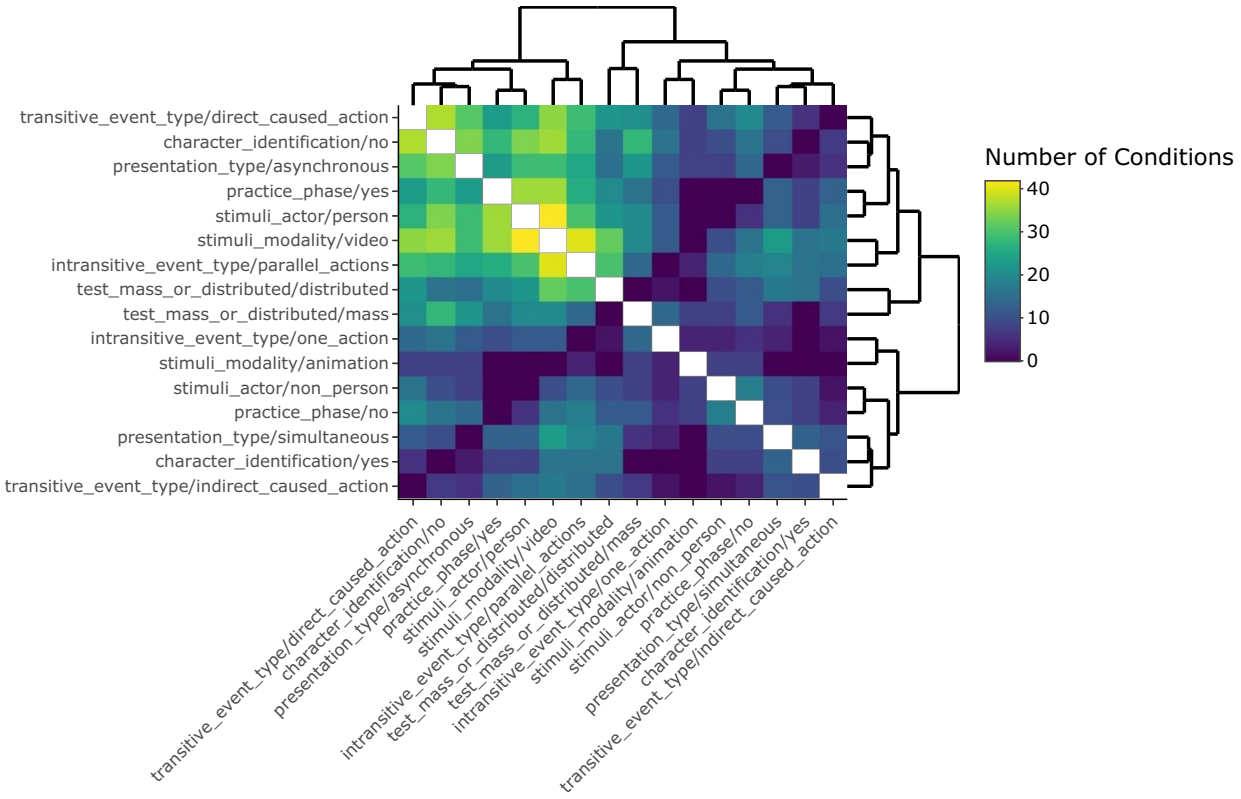
Relationship between additional methodological moderators

We coded a number of additional methodological variables that substantially overlap with those in the paper. We report them here for completeness.

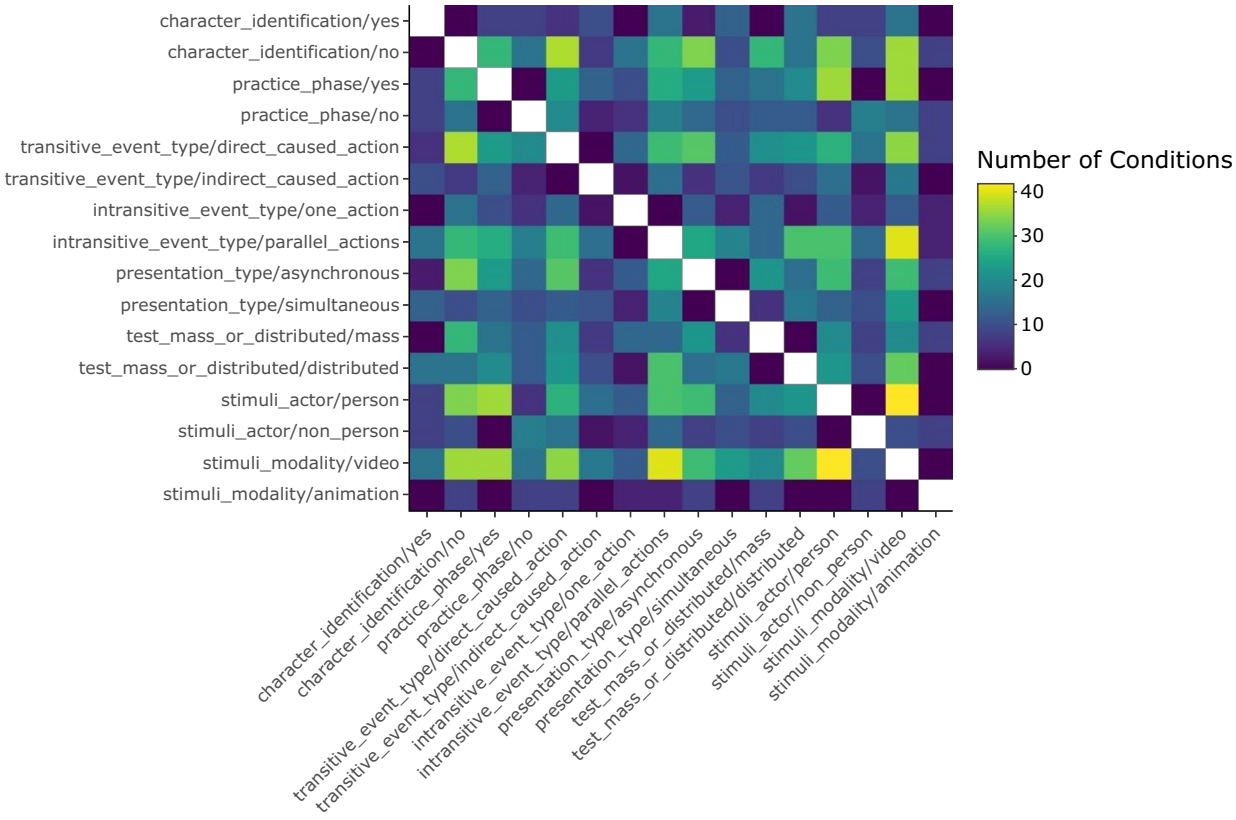
The additional coded moderators were as follows. First, we coded the modality of the visual stimuli. Stimuli modality has two levels: videos and animations. We coded this moderator following the details provided in the method sections of the papers. Stimuli actors have two levels, human actors and non-human actors. Studies using visual stimuli with human actors wearing animal suits were coded as using non-human actors. Second, we coded the types of the events presented in the visual stimuli. To capture the event types with details, we coded the transitive action stimuli and the intransitive action stimuli separately. For transitive action stimuli, we coded two levels: direct caused action and indirect caused action. The event was coded as using direct caused action if the agent in the action directly acted upon the patient. It was coded as using indirect caused action if the agent caused the patient to move via another medium. For example, the agent may pull a band on the patient's waist causing her to move. Likewise, the intransitive event also has two levels: one action versus parallel actions. Here we coded the levels by number of participants presented on the screen. An intransitive event was coded as "one action" if and only if there was only one agent presented on the screen. If an event involves more than one actor in the intransitive event (e.g. two actors doing parallel actions or one actor with one stander-by), then the event was coded as parallel-actions.

These additional moderators were not included in the main analyses because of their close relationships between each other and with the main moderators. The heatmaps below showed the overlap between moderators. Each cell corresponds to the co-occurrence between two moderator levels. Brighter colors indicate a higher frequency of co-occurrence, and darker colors indicate lower frequency. You can hover your mouse on the heatmap to see the corresponding value and combination of each cell.

Ordered by Row Average

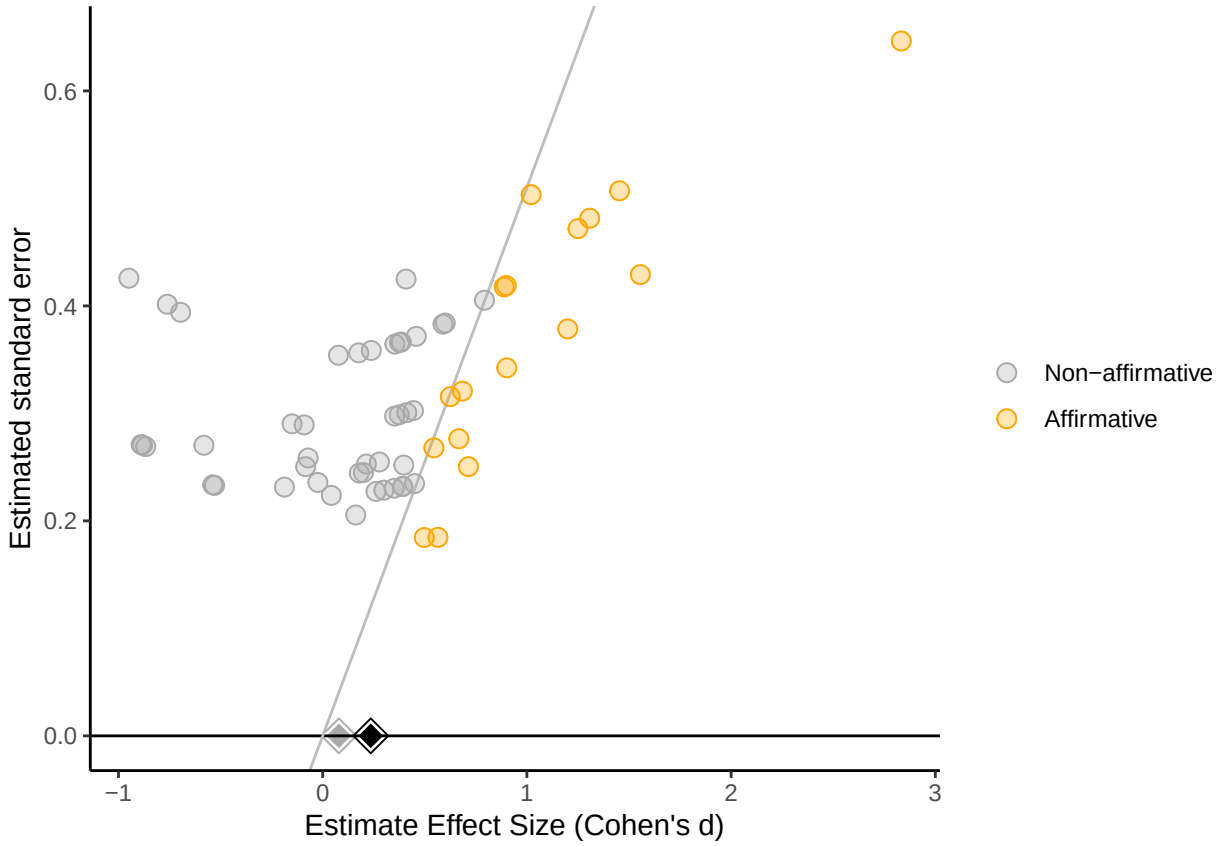


Ordered by groups



Sensitivity analysis

The plot below shows a modified funnel plot, or "significance funnel" where significant studies are shown in orange and non-significant studies are shown in grey (Marthur & VanderWeele, 2020). The x-axis shows effect size estimates, and the y-axis shows estimated standard error for each estimate. Studies lying on the grey line have a p-value of .05. The black diamond shows the meta-analytic effect size estimate for all studies; the grey diamond shows the meta-analytic effect size estimate for significant studies only (the "worst-case" publication scenario). Note that the worst case scenario appreciable attenuates the effect size estimate, but does not attenuate the point estimate to 0 (worst case estimate: 0.08 [-0.1, 0.25]).



Main model results

The tables below show the estimates for all single-moderator models reported in the main text. The tables present point estimates for the model parameters and their 95% confidence intervals (i.e., [lower bound, upper bound]). Asterisks (*) indicate significance at the .05 level. For categorical variables, the base levels are represented as the first ones appeared in the parentheses.

Mean age

Parameter	Estimate	z value	p value
Intercept	0.6 [0.07, 1.13]	2.23	0.03*
Mean Age (months)	-0.01 [-0.03, <.001]	-1.47	0.14

Median productive vocabulary size

Parameter	Estimate	z value	p value
Intercept	0.66 [0.04, 1.28]	2.08	0.04*
Median productive vocabulary size	-0.01 [-0.02, <.001]	-1.74	0.08

Predicate Type

Parameter	Estimate	z value	p value
Intercept	0.1 [-0.14, 0.34]	0.8	0.42
Predicate type (Intransitive / Transitive)	0.24 [0.02, 0.46]	2.1	0.04*

Noun phrase type

Parameter	Estimate	z value	p value
Intercept	0.26 [-0.02, 0.53]	1.83	0.07
Noun phrase type (Noun / Pronoun)	-0.04 [-0.4, 0.31]	-0.24	0.81

Character identification phase

Parameter	Estimate	z value	p value
Intercept	0.19 [-0.06, 0.43]	1.51	0.13
Character identification phase (No / Yes)	0.18 [-0.28, 0.64]	0.75	0.45

Practice phase

Parameter	Estimate	z value	p value
Intercept	0.35 [0.09, 0.62]	2.59	0.01*
Practice phase (No / Yes)	-0.21 [-0.5, 0.09]	-1.35	0.18

Synchronicity

Parameter	Estimate	z value	p value
Intercept	0.2 [-0.05, 0.44]	1.59	0.11
Synchronicity (Asynchronous / Simultaneous)	0.1 [-0.22, 0.41]	0.59	0.55

Testing structure

Parameter	Estimate	z value	p value
Intercept	0.13 [-0.1, 0.35]	1.1	0.27
Testing Procedure Structure (Distributed / Mass)	0.37 [-0.05, 0.79]	1.75	0.08

Number of sentence repetitions

Parameter	Estimate	z value	p value
Intercept	0.18 [-0.12, 0.48]	1.18	0.24
Number of sentence repetitions	0.01 [-0.02, 0.03]	0.52	0.6

Models with methodological moderators and theoretical moderators

Syntactic bootstrapping studies differ in their implementational details. Here we examine the extent to which influences of the theoretical moderators can be accounted for by the methodological factors. The tables below present the results of models that include a single theoretical moderator along with all the methodological moderators in additive models. The row corresponding to the theoretical moderator is highlighted in yellow. The effect of the theoretical moderators is qualitatively identical to the models without the methodological moderators included.

With age

Parameter	Estimates	z value	p value
Intercept	-0.03 [-0.93, 0.88]	-0.06	0.95
Character identification phase (No / Yes)	0.24 [-0.3, 0.78]	0.88	0.38
Practice phase (No / Yes)	-0.11 [-0.46, 0.24]	-0.62	0.54
Stimuli synchronicity (Asynchronous / Simultaneous)	0.29 [-0.26, 0.83]	1.02	0.31
Testing structure (Distributed / Mass)	0.47 [0.02, 0.92]	2.07	0.04*
Number of sentence repetitions	0.02 [-0.02, 0.06]	0.93	0.35
Mean age (months)	**-0.01 [-0.03, 0.02]**	**-0.53**	**0.59**

With productive vocabulary size

Parameter	Estimates	z value	p value
Intercept	-0.24 [-2.19, 1.72]	-0.24	0.81
Character identification phase (No / Yes)	0.33 [-0.76, 1.43]	0.60	0.55
Practice phase (No / Yes)	0.01 [-1.05, 1.07]	0.02	0.98
Stimuli synchronicity (Asynchronous / Simultaneous)	0.17 [-1.14, 1.48]	0.25	0.8
Testing structure (Distributed / Mass)	0.95 [0.23, 1.67]	2.59	0.01*
Number of sentence repetitions	0.01 [-0.1, 0.12]	0.14	0.89
Median productive vocabulary size	**-0.01 [-0.03, 0.01]**	**-0.74**	**0.46**

With predicate type

Parameter	Estimates	z value	p value
Intercept	-0.4 [-1.01, 0.21]	-1.29	0.2
Character identification phase (No / Yes)	0.27 [-0.25, 0.79]	1.03	0.3
Practice phase (No / Yes)	-0.08 [-0.38, 0.22]	-0.51	0.61
Stimuli synchronicity (Asynchronous / Simultaneous)	0.29 [-0.22, 0.8]	1.11	0.27
Testing structure (Distributed / Mass)	0.54 [0.09, 0.99]	2.34	0.02*
Number of sentence repetitions	0.02 [-0.01, 0.06]	1.19	0.23
Predicate type (Intransitive / Transitive)	**0.24 [>-.001, 0.47]**	**1.98**	**0.05***

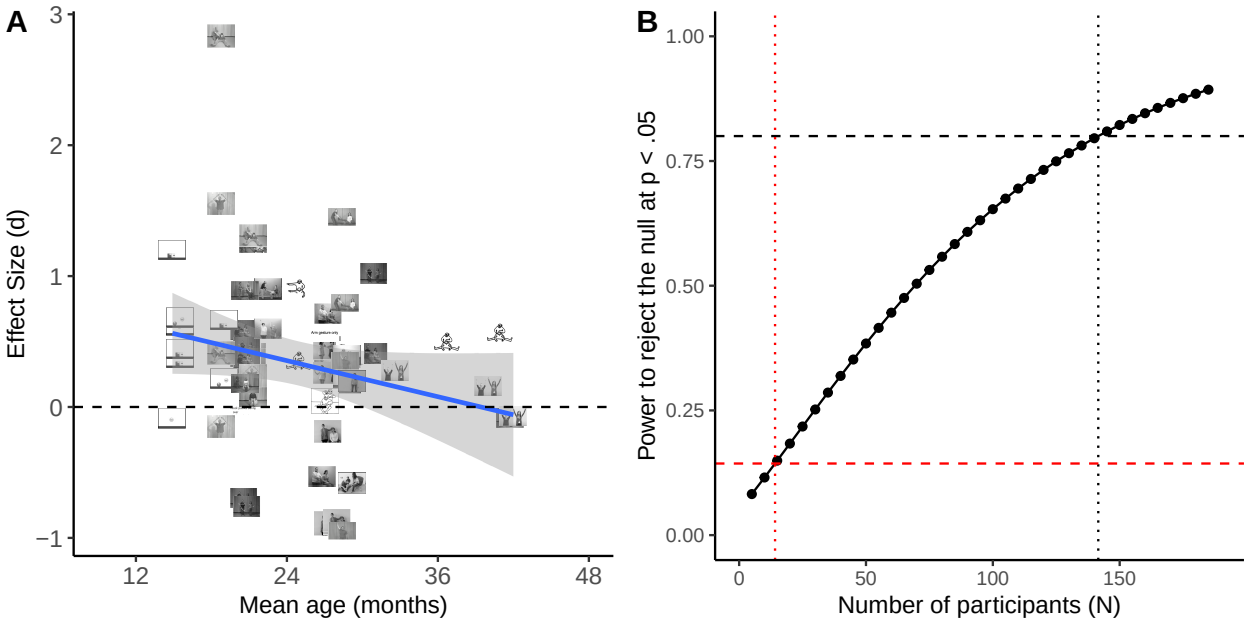
With Noun phrase type

Parameter	Estimates	z value	p value
Intercept	-0.26 [-0.98, 0.45]	-0.73	0.47
Character identification phase (No / Yes)	0.21 [-0.32, 0.73]	0.77	0.44
Practice phase (No / Yes)	-0.17 [-0.5, 0.16]	-1.03	0.31
Stimuli synchronicity (Asynchronous / Simultaneous)	0.37 [-0.28, 1.03]	1.12	0.26
Testing structure (Distributed / Mass)	0.44 [-0.1, 0.98]	1.61	0.11
Number of sentence repetitions	0.02 [-0.01, 0.06]	1.23	0.22
Noun phrase type (Noun / Pronoun)	**0.07 [-0.5, 0.65]**	**0.25**	**0.8**

Variability in visual stimuli as a function of age

To assess the relationship between visual stimuli complexity and participant age, we collected sample visual stimuli for each condition in our sample (Panel A). Schematic illustrations of the visual stimuli were used when the actual screenshots were not provided. Screenshots of the text descriptions of the events were used when the visual stimuli were unavailable. Note that because some papers' publishers converted to the visual stimuli to black-and-white, we decided to grayscale all visual stimuli for easier visual comparison.

With the exception of studies from one paper targeting the youngest participants in our sample (Jin, 2015), there was little systematic variability in the complexity of visual stimuli as a function of age. This suggests that adaptation of visual stimuli to participants' age is unlikely to be the cause of the lack of developmental change of the strength of the effect.



Power analysis for experiment

We conducted a power analysis using the `pwr` package (Panel B; Champely et al., 2018). The x-axis represents the number of participants in each condition, and the y-axis represents the estimated power based on the power of estimated meta-analytic effect size ($d = 0.24$). The horizontal black dotted line represents 80%

power, and the vertical black dotted lines represent the number of participants needed to reach 80% power ($N = 142$). The red lines represents the current power (14.36%) based on the approximate mean sample sizes ($N = 14$) of the conditions included in the meta-analysis.

Simulations of meta-analytic moderator power

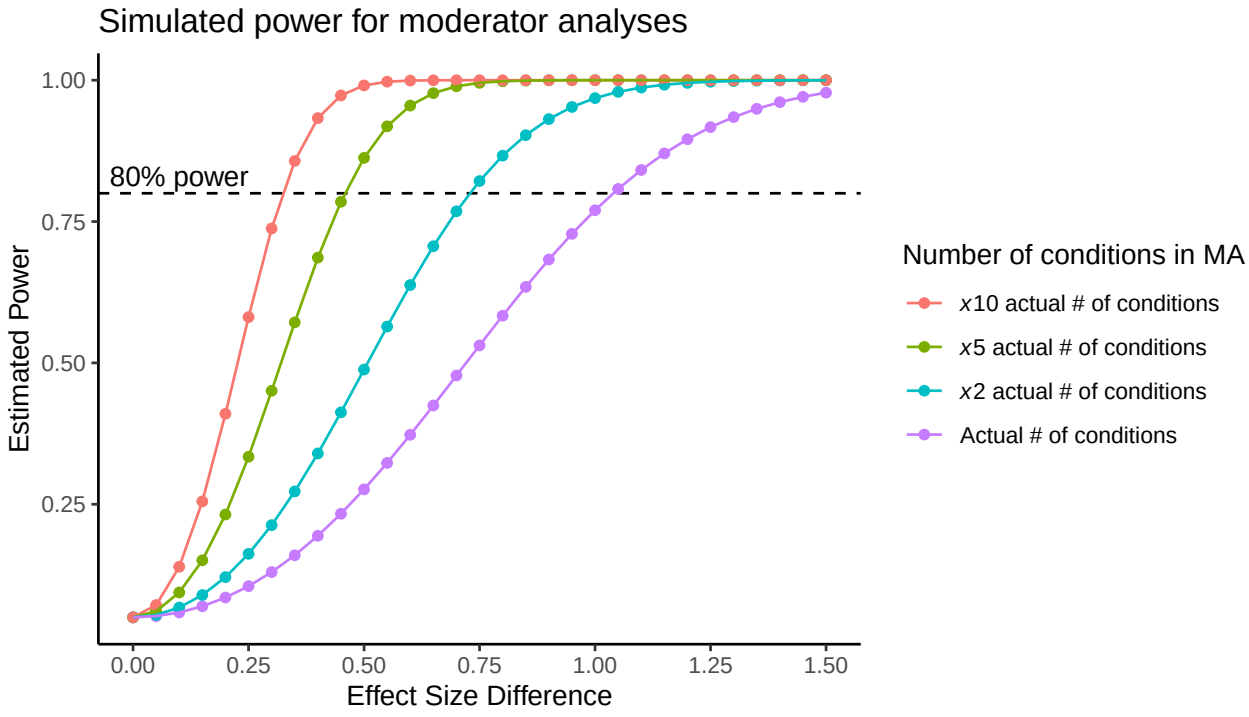
For a number of the moderators we examined in our meta-analysis (e.g., noun phrase type), the average effect size did not statistically differ from zero. As such, in these cases, we failed to reject the null hypothesis. This failure could be for two reasons: there is no true effect or there is an effect but we do not have the statistical power to detect it. We cannot determine which of these two reasons lead to the null moderating effects that we report, but we can examine our power for moderator effects in the current meta-analysis (i.e. probability of rejecting a false null hypothesis).

There are not agreed-upon methods for calculating power for the multilevel mixed effect meta-analytic models we report in the paper, but methods do exist for mixed effect models (Hedges & Pigott, 2004). We conducted simulations of power for moderating effects using an implementation of these methods in R (metapower; Griffin, 2020).

To estimate power for a categorical moderating effect, we varied the effect size difference between the two levels of the moderator (e.g., "pronoun" vs. "noun"), and the number of conditions (studies) present for each level in the meta-analysis. We estimated the between study variance empirically ($I^2 = 71.69$).

The figure below shows the estimated power if our meta-analytic models for moderator effects. Shown in purple is the estimated power, given the actual number of conditions present in our meta-analysis; other lines show estimated power if more conditions were present in our meta-analysis. The dashed line shows 80% power.

This analysis suggests that we have reasonably high power to detect large effects size differences (at least d = ~1), but low power to detect smaller effect sizes. Thus, the null effects for moderating effects reported in the paper should be interpreted with caution: while these analyses provide strong evidence that moderator effects are not large, it remains possible that some of the null effects we report are actually non-zero but small in magnitude.



References

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., ... & De Rosario, M. H. (2018). Package 'pwr'. R package version, 1(2).

Griffin JW (2020). metapoweR: an R package for computing meta-analytic statistical power. R package version 0.2.1, https://CRAN.R-project.org/package=metapower.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. Psychological methods, 9(4), 426.

Hirsh-Pasek, K., Golinkoff, R. M., & Naigles, L. (1996). Young children's use of syntactic frames to derive meaning. The origins of grammar: Evidence from early language comprehension, 123-158.

Jin, K. S. (2015). The role of syntactic and discourse information in verb learning (Doctoral dissertation, University of Illinois at Urbana-Champaign).

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. Journal of the Royal Statistical Society. Series C, Applied Statistics, 69(5), 1091.

Yuan, S., & Fisher, C. (2009). "Really? She blicked the baby?" Two-year-olds learn combinatorial facts about verbs by listening. Psychological science, 20(5), 619-626.