1   Conducting developmental research online vs. in-person: A meta-analysis

2   Aaron Chuey[1], Veronica Boyce[1], Anjie Cao[1], & Michael C. Frank[1]

3   [1] Stanford University, Department of Psychology

4   Author Note

Abstract

An increasing number of psychological experiments with children are being conducted using online platforms, in part due to the COVID-19 pandemic. Individual replications have compared the findings of particular experiments online and in-person, but the general effect of online data collection on data collected from children is still unknown. Therefore, the current meta-analysis examines how the effect sizes of developmental studies conducted online compare to the same studies conducted in-person. Our pre-registered analysis includes 145 effect sizes calculated from 24 papers with 2440 children, ranging in age from four months to six years. We examined several moderators of the effect of online testing, including the role of dependent measure (looking vs verbal), online study method (moderated vs unmoderated), and age. The mean effect size of studies conducted in-person (d = .68) was slightly larger than the mean effect size of their counterparts conducted online (d = .54), but this difference was not significant. Additionally, we found no significant moderating effect of dependent measure, online study method, or age. Overall, the results of the current meta-analysis suggest developmental data collected online are generally comparable to data collected in-person.

Keywords: Methodology, Meta-analysis, Development, Online studies

Word count: X

₃₂   Conducting developmental research online vs. in-person: A meta-analysis

₃₃   **Introduction**

₃₄   Developmental researchers are interested in studying children's behavior, primarily by

₃₅ measuring their behavioral responses to experimental stimuli. Study sessions typically

₃₆ involve visits with local families in a laboratory setting or partnering with remote sites such

₃₇ as schools and museums. Although these interactions are a routine part of developmental

₃₈ research, they are time-consuming for both researchers and participants. Typical studies

₃₉ with dozens of infants or young children can require weeks or months of scheduling visits to a

₄₀ lab or many visits to testing sites. In-person testing also limits the participant pool to

₄₁ children living relatively close to the research site. Additionally, developmental research has

₄₂ been plagued by small, non-diverse samples even more so than research with adults due to

₄₃ limitations imposed by the demographics of the local population as well as the high costs of

₄₄ collecting data from children (Kidd & Garcia, 2022; Nielsen, Haun, Kärtner, & Legare, 2017).

₄₅   Prior to the rise of video chat software, there were only limited alternatives to in-person

₄₆ interaction for collecting experimental behavioral data from children. However, with the

₄₇ development of inexpensive and reliable video conferencing technology in the 2010s, new

₄₈ frontiers began to emerge for developmental testing.[1] Researchers soon experimented with

₄₉ conducting developmental studies through video-chat platforms, which in theory broaden the

₅₀ pool of participants to anyone with internet access at nearly any time and location. What

₅₁ began as a few research teams experimenting with online studies (e.g., Lookit: Scott &

₅₂ Schulz, 2017; The Child Lab: Sheskin & Keil, 2018; Pandas: Rhodes et al., 2020) quickly

₅₃ expanded to much of the field as researchers scrambled to conduct safe research during the

₅₄ Covid-19 pandemic. This shift in research practices has yielded many empirical publications

₅₅ where some or all of the data were collected online in addition to a growing literature on

---

[1] Observational and survey research has long been conducted through the phone or by mail (e.g., Fenson et al., 1994); here we focus primarily on behavioral observation and experimental methods.

56   online methodology and best practices (for a recent review, see Chuey, Asaba, et al., 2021).

57   Some researchers may be eager to return to in-person testing, but online research is

58   likely here to stay and may increase in frequency as communications technologies improve

59   and become more accessible. Online testing has immense potential to change developmental

60   science (Sheskin et al., 2020), much as crowdsourced testing of adults has changed adult

61   behavioral science (Buhrmester, Kwang, & Gosling, 2016). This potential has yet to be fully

62   realized, however, as researchers have yet to fully understand the strengths and weaknesses

63   of this method, as well as how to recruit diverse populations for online studies. Despite

64   undersampling certain populations (Lourenco & Tasimi, 2020), online studies nonetheless

65   allow researchers to sample from a larger, broader pool of participants than ever before as

66   access to the internet continues to increase worldwide. Large, low cost samples and remote

67   cross-cultural research may even become a reality for developmental researchers in the

68   coming years.

69   Is conducting developmental studies online an effective substitute for conducting them

70   in-person, or do online studies yield systematically different effects? Direct comparison of

71   effects measured in both modalities is critical to answering this question. Researchers have

72   implemented a number of paradigms online and replicated their in-person findings, but the

73   quality of data yielded from online studies in comparison to those conducted in-person more

74   broadly is still largely unknown. Therefore, the current meta-analysis examines how data

75   collected from children online compares to data collected from closely-matched studies

76   in-person. Importantly, online studies themselves are not a monolith, and differ in a

77   multitude of ways including the presence of a live experimenter, dependent measure, and the

78   age of the sample being tested.

79   Online studies are generally conducted in one of two formats: moderated and

80   unmoderated. In moderated studies, a live experimenter guides participants through a study

81   much like they would in-person, except online, typically via video-chat. Moderated studies

82 are often operationalized as slide share presentations or videos shared with participants while

83 the participants' verbal responses or looking is recorded. In unmoderated studies, conversely,

84 participants complete a study without the guidance of a live experimenter. Instead,

85 researchers create a preprogrammed module that participants or their parents initiate and

86 complete according to instructions. Since no experimenter needs to be present and

87 participants can participate at any time they choose, unmoderated studies offer the potential

88 for fast, inexpensive data collection. However, since they lack an experimenter, participants'

89 experiences also deviate more from in-person studies compared to moderated studies that

90 retain the same core social interaction between experimenter and participant. Therefore, it is

91 possible that data collected via unmoderated sessions is comparatively noisier since an

92 experimenter is unable to focus children's attention or course correct like they can during a

93 live interaction. We consider this possibility in the current meta-analysis.

94   Like developmental studies more broadly, online studies have also employed a number

95 of dependent measures, including verbal measures and looking measures. Verbal measures

96 are typically straightforward to record, while recording looking measures is more complex.

97 Accurate looking measures require precise camera positioning and coding schemes, and are

98 thus more likely to deviate from their in-person counterparts compared to studies that

99 measure children's verbal responses. To that end, automated gaze annotation is currently

100 being developed and represents an exciting future direction in online methodology (see Erel,

101 Potter, Jaffe-Dax, Lew-Williams, & Bermano, 2022). We examine how the kind of dependent

102 measure employed (looking vs. verbal) might moderate the difference between online and

103 in-person results.

104   The final moderator we consider is participants' age. Online developmental studies

105 have sampled from a wide age range, including infants (e.g., Dillon, Izard, & Spelke, 2020),

106 toddlers (e.g., Lo, Rosslund, Chai, Mayor, & Kartushina, 2021), preschoolers (e.g., Schidelko,

107 Schünemann, Rakoczy, & Proft, 2021), and elementary schoolers (e.g., Chuey, Lockhart,

108 Sheskin, & Keil, 2020; Chuey, McCarthy, et al., 2021). Because online studies are often

109 conducted in the comfort of their own homes, it is possible that children of all ages might

110 benefit from this aspect of online studies. Conversely, because a child's environment is more

111 difficult to moderate online, infant studies, which often rely on precise environmental setups,

112 may suffer more when conducted online. In addition, as children get older they may gain

113 more experience with on-screen displays, which can contribute to their performance in online

114 studies. We test these competing age moderation hypotheses.

115 In sum, our meta-analysis addresses the question of whether effect sizes tend to differ

116 across online and in-person experiments with children, and whether these differences are

117 moderated by study format, dependent variable, or participant age.

Methods

119 We conducted a literature search following the Preferred Reporting Items for

120 Systematic Reviews and Meta-Analyses (PRISMA) procedure (Moher et al., 2015); see

121 Figure 1. For each set of studies determined to be an online replication, we calculated the

122 effect size(s) and associated variance for the main effect of interest. We then conducted a

123 series of random-effects multilevel meta-regressions to estimate the effect of online data

124 collection, as well as three possible moderators (online study method, type of dependent

125 measure, and participant age). Our preregistered data selection, coding, and analysis plan

126 can be found at https://osf.io/hjbxf. The list of papers included in this meta-analysis is

127 shown in Table 1.

Literature Search

129 Our goal was to find as many published and unpublished online replications of

130 developmental studies as possible. However, because there is no common nomenclature for

131 online replications and the studies themselves cover a wide range of research questions and

132 methodologies, searching via specific terms or keywords was difficult and produced many
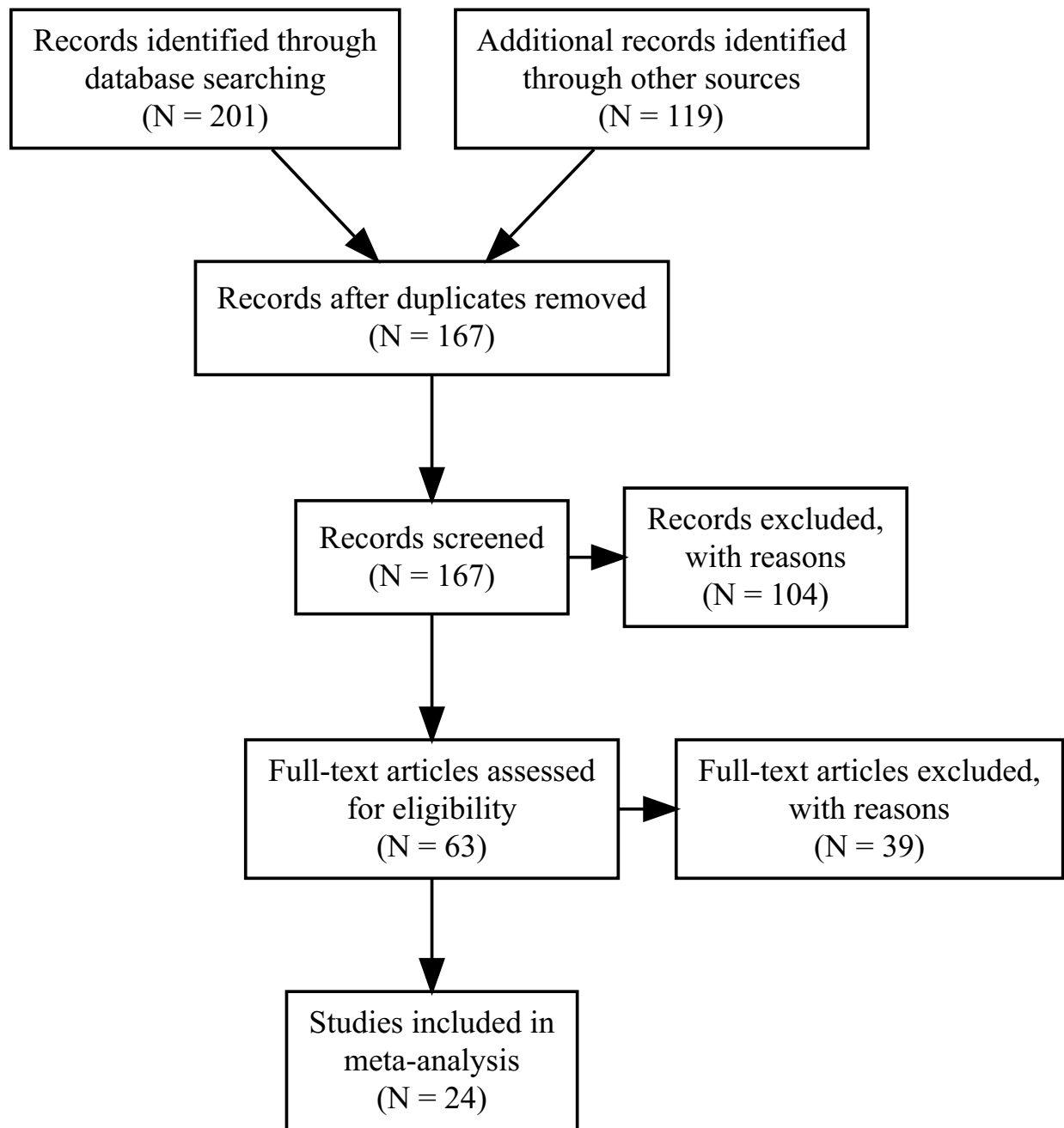
Figure 1. PRISMA plot detailing our study screening process.

irrelevant papers. Instead, we preregistered a forward citation search strategy based on key papers on online developmental research. We used the papers that conducted initial validation of popular online testing platforms as our seeds, including Lookit (Scott, Chu, & Schulz, 2017; Scott & Schulz, 2017), The Child Lab (Sheskin & Keil, 2018), and Pandas (Rhodes et al., 2020). We also included all papers published in the Frontiers in Psychology Special Issue: Empirical Research at a Distance: New Methods for Developmental Science, which largely focused on online developmental studies and replications. Finally, we posted a call for contributions to the Cognitive Development Society and ICIS listservs, two popular emailing lists frequented by developmental researchers. This call yielded several publications our initial search strategy missed, as well as six unpublished but complete online replications.

We preregistered several eligibility criteria to filter articles from our search:

1. The study must be experimental, where participants complete a task with a stimulus. This criterion precludes surveys or purely observational measures.

2. The studies must report two groups of children, one tested online and another tested in-person. Although the online sample must be collected by the researchers reporting the results, the in-person sample could either be collected at the same time or referenced from an existing publication.

3. The mean age of the sample should be under six years. This criterion limits the studies to those conducted on relatively younger children for whom online data collection methods have not been traditionally employed.

4. All data reported or referred to must contain codeable effect sizes. Verbal comparison alone between an online or in-person study or a qualitative description of results is not enough to determine the precise effect size of interest.

5. Data collection for both the in-person and online sample must be complete; any

157   incomplete or partial samples were not considered.

158   6. The online and in-person methods must be directly comparable. Some alteration to the

159   study methods is expected when adapting an in-person study to be run online (e.g.,

160   changing a preferential reaching measure into a preferential looking measure, having

161   children refer to objects by color instead of pointing, etc). However, we excluded any

162   studies whose methodologies altered the nature of the task or the conclusions that

163   could be drawn from them (e.g., manipulating the identity of an object instead of its

164   location).

Table 1

Papers used in this meta-analysis. Some papers contained both online and in-person results, others contained online replications compared to previous in-person papers. Pairs is number of online – in-person pairs contributed by each paper (set). Look is whether the studies are use looking, verbal, or both types of dependent measures. Mod is whether the online studies were moderated, unmoderated, or both. Age is the average age of the participants in months.

Paper	Pairs	Look	Mod	Age
Gasparini et al. (2022)	5	Verb	Mod	4
Bánki, Eccher, Falschlehner, Hoehl, and Markova (2022)	4	Look	Mod	5
DeJesus, Venkatesh, and Kinzler (2021)	3	Verb	Mod	5
Bochynska and Dillon (2021) compared to Dillon et al. (2020)	2	Look	Unmod	7
Bulgarelli and Bergelson (2022)	3	Look	Mod	8
Yuen and Hamlin (2022) compared to Hamlin (2015)	2	Both	Mod	9
Smith-Flores, Perez, Zhang, and Feigenson (2022a) compared to Stahl and Feigenson (2015)	3	Look	Mod	13
Smith-Flores, Perez, Zhang, and Feigenson (2022b) compared to Skerry and Spelke (2014)	2	Look	Mod	13
Lo et al. (2021)	1	Verb	Unmod	20

Paper	Pairs	Look	Mod	Age
Margoni, Baillargeon, and Surian (2018)	2	Look	Mod	21
Chuey, Asaba, et al. (2021)	3	Both	Mod	24
Man (2022)	1	Look	Mod	24
Morini and Blair (2021)	1	Verb	Mod	30
Silver et al. (2021)	1	Verb	Mod	33
Schidelko et al. (2021)	4	Verb	Mod	44
Lapidow, Tandon, Goddu, and Walker (2021)	4	Verb	Both	44
Scott et al. (2017) compared to Téglás, Girotto, Gonzalez, and Bonatti (2007) and Pasquini, Corriveau, Koenig, and Harris (2007)	17	Both	Unmod	45
Yoon and Frank (2019)	2	Verb	Unmod	48
Kominsky, Shafto, and Bonawitz (2021)	1	Verb	Mod	55
Escudero, Pino Escobar, Casey, and Sommer (2021)	2	Verb	Mod	57
Vales et al. (2021)	3	Verb	Mod	58
Nelson, Scheiber, Laughlin, and Demir-Lira (2021)	8	Verb	Mod	59
Gerard (2022)	1	Verb	Unmod	60
Aboody, Yousif, Sheskin, and Keil (2022)	1	Verb	Mod	72

Data Entry

All papers (320) yielded by our search procedure went through three rounds of evaluation to determine if they met our inclusion criteria. First, we screened the titles of the papers to determine whether they might include an online experiment. Those that clearly did not meet one or more of our inclusion criteria were excluded from further evaluation. Next, we performed a similar evaluation based on the papers' abstracts, before a final round based on the article as a whole. All remaining papers were entered into a spreadsheet that coded the necessary information for us to calculate the size of the main effect(s) of interest

173 and their associated variance (sample size, group means and standard deviation, and t and F

174 statistics when applicable), as well as our preregistered moderators (study modality, data

175 collection method, dependent measure, and participant age).

176 If a paper reported an effect size as cohen's d (referred to below as standardized mean

177 difference, SMD), we coded it directly. Otherwise, we calculated the individual effect sizes

178 for each main effect and each study (online and in-person) using reported means and

179 standard deviations, t statistic, or directly from the data if it was available. If the main

180 comparison was to chance performance, we first calculated log odds and then converted the

181 effect size to cohen's d via the compute.es package in R (Del Re & Del Re, 2012). If a given

182 study had multiple dependent measures or central hypotheses, we calculated an effect size

183 and associated variance for each.

184 **Analytic Approach**

185 To determine whether study modality (online or in-person) moderated the size of the

186 main effect of interest for each set of studies, we performed a random-effects multilevel

187 meta-regression using the metafor package (Viechtbauer, 2010). The regression predicts

188 effect size (SMD) with study modality as a fixed effect.

189 Our analysis reflects a key design choice for our meta-analysis. Naively, it might

190 appear to be possible to predict the size of the online-offline difference for a particular study.

191 But on examination, many papers are heterogeneous and contain multiple online studies for

192 a given offline study, or multiple measures for the same study. In these cases, the appropriate

193 difference was not always clear. Further, many pairs of studies differed on some value of our

194 chosen moderators.

195 To deal with these issues, we instead modeled individual experimental effect sizes, with

196 the coefficient of interest being the study modality predictor (online vs. in-person). We

197 included two random intercepts in our models. The first random intercept controlled for

198 variation between particular experiments (e.g., modeling the dependency between multiple

199 measurements reported from a single experiment). The second controlled for variation

200 between groups of participants (e.g., modeling the dependency between effect sizes from

201 participants who completed a battery of tasks with multiple effects of interest).

202 To determine the effect of additional moderators (online study method, dependent

203 measure, and participant age), we conducted three additional multilevel meta-regressions

204 each with an additional fixed effect plus the corresponding interaction with study modality.

205 All analysis scripts were preregistered, and the code is available at

206 https://osf.io/up6qn/?view_only=91ba54134dc24787b04dd8f3b3b70e1e.
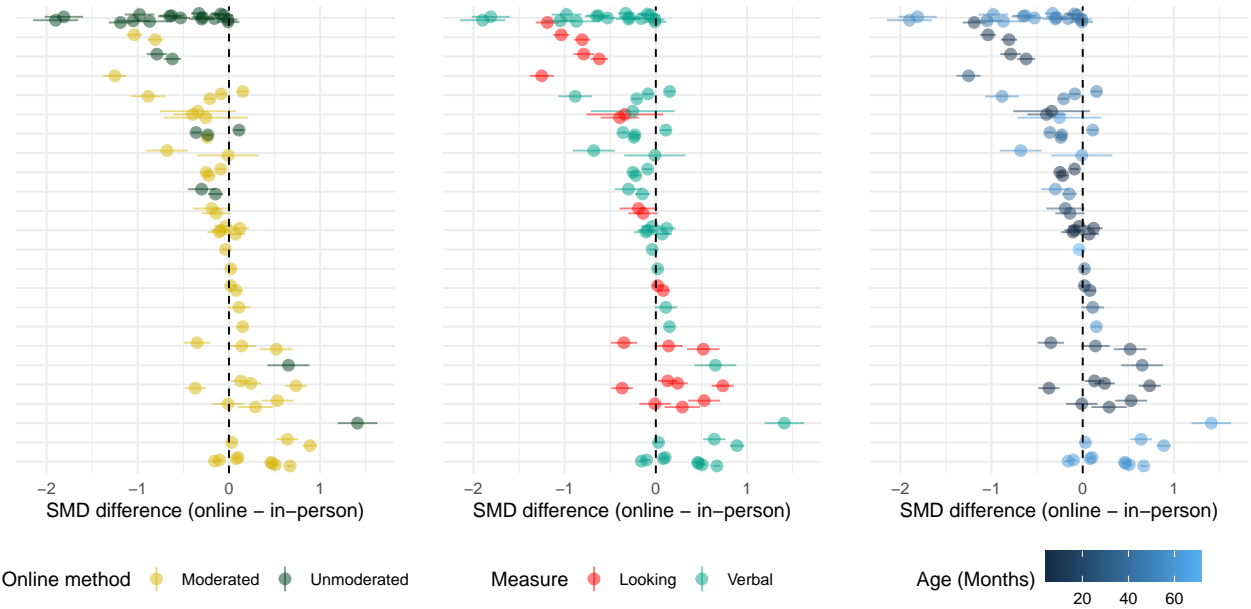
207 **Results**

Figure 2. Forest plots of studies, sorted by difference in SMD. Each dot is the difference between and in-person measure and a corresponding online measure. Each row is one study (paper or pair of papers).

Table 2

Table of coefficients for the pre-registered models. The overall model is shown first, followed by the three models with moderators.

Coefficient	Estimate	95% CI	P-value
Overall			
Intercept	0.84	[0.46, 1.21]	0.000
Online	-0.15	[-0.38, 0.08]	0.210
Looking v Verbal			
Intercept	0.73	[0.42, 1.04]	0.000
Online	-0.29	[-0.7, 0.11]	0.155
Verbal	-0.06	[-0.43, 0.31]	0.745
Online:Verbal	0.23	[-0.27, 0.72]	0.375
Age			
Intercept	0.68	[0.51, 0.86]	0.000
Online	-0.14	[-0.38, 0.1]	0.244
Age	0.00	[-0.01, 0.01]	0.731
Online:Age	0.01	[-0.01, 0.02]	0.342
Moderated v Un-moderated			
Intercept	0.69	[0.52, 0.86]	0.000
Online	-0.19	[-0.45, 0.07]	0.151
Unmoderated	0.13	[-0.22, 0.48]	0.461

Confirmatory Analysis

Overall, the meta-analysis estimated a small negative, non-significant effect of online study modality, Est=-0.15, 95% CI=[-0.38, 0.08], p=0.21. Additionally, we did not find any significant effect of our preregistered moderators or any significant interactions between the

212 moderators and study modality. See Table 2 for coefficient values. Figure 2 shows the effect
213 size differences of experiments by moderators.

214 Because our meta-analysis averaged across effects from very different paradigms (which
215 could yield different effect sizes independent of the effect of testing modality), we expected
216 substantial heterogeneity. Consistent with that expectation, all tests for residual
217 heterogeneity were highly significant (all $ps < .0001$). Values of τ^2 for the models were 0.23
218 (primary model), 0.23 (looking-time model), 0.23 (age model), and 0.23 (moderated
219 vs. unmoderated model), respectively, confirming the impression that these moderators did
220 not reduce heterogeneity.

221 **Exploratory Analysis**

Table 3

*Mean SMD across studies by study modality, data-collection method, and type of dependent
measure*

Modality	Method	Measure	SMD
In-person	Moderated	Looking	0.699
In-person	Moderated	Verbal	0.677
Online	Moderated	Looking	0.597
Online	Moderated	Verbal	0.511
Online	Unmoderated	Looking	0.177
Online	Unmoderated	Verbal	0.570

222 As an exploratory analysis, we examined which combinations of methods and measures
223 tended to yield the strongest and weakest effect sizes relative to their in-person counterparts.
224 We fit a meta-analytic model containing method, response mode, and modality as well as
225 their two- and three-way interactions, with the same random effects structure as our

previous model. We cannot draw any strong conclusions about these noisy estimates due to our relatively small sample size. That said, unmoderated online studies with looking measures were estimated to have the weakest effect sizes compared with their in-person counterparts, an average difference of 0.52 (See Table 3). In contrast, as estimated by this model, moderated online studies with looking and verbal measures as well as unmoderated online studies with verbal measured all differed by less than .2 SMD from their in-person counterparts.
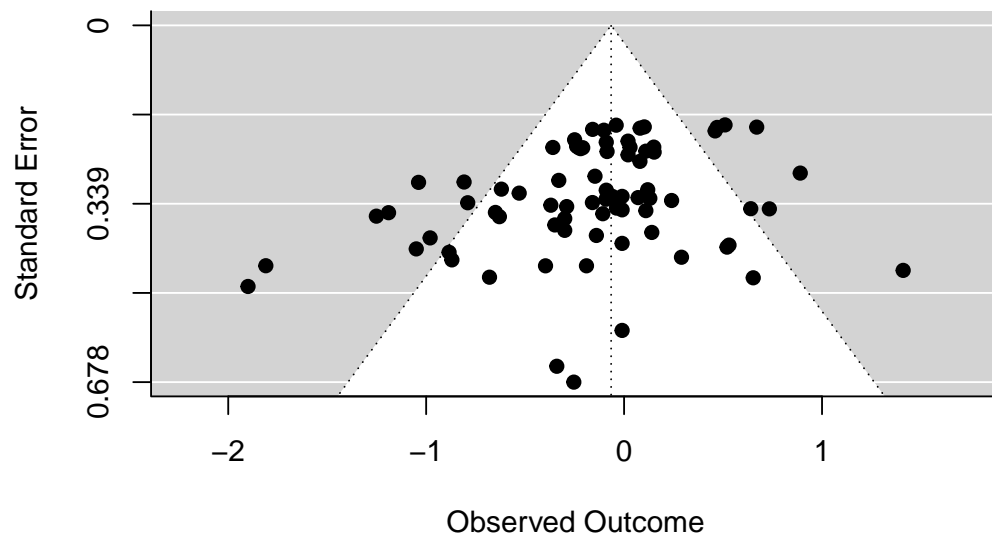

Figure 3. Funnel plot of the differences in effect size between pairs of in-person and online studies. A positive observed outcome means the online study had a large effect.

We additionally conducted an exploratory analysis of potential publication bias. It was unclear *a priori* how we might expect publication biases to manifest themselves, given that there is some possibility of notoriety for either showing *or* failing to show differences between online and in-person testing. In either case our hypothesized selection process operated on the *differences* in effect sizes between each online and in-lab pair of samples.

For each online and in-person pair on the same study, we calculated a standard mean difference in effect size between the two studies as well as the variance of this difference. The resulting funnel plot is shown in Figure 3. According to Egger's regression test for funnel

241 plot asymmetry, this plot is asymmetric (p=.005) and the limit estimate of the effect as

242 standard error goes to zero is 0.37 [0.01, 0.72]. This analysis suggests the possibility of

243 publication bias favoring negative comparisons between online and in-lab modalities,

244 signalling that perhaps the true effect of online testing is zero or perhaps even positive. We

245 interpret this conclusion with caution, however, noting the large width of the estimated CI

246 and the relatively low power of Egger's test (Sterne, Gavaghan, & Egger, 2000).

Discussion

248 By aggregating across a growing literature of online studies, the current meta-analysis

249 provides a birds-eye view of how developmental studies traditionally conducted in-person

250 fare compared to closely matched counterparts conducted online. Our results suggest that

251 overall, the results of online studies are comparable to those conducted in-person. Based on

252 our analysis, the method of online data collection, type of dependent measure, and

253 participant age did not appear to have a significant impact either. Nonetheless, the relatively

254 small sample size limits our ability to make sweeping generalizations about any of our

255 moderators, so future analysis is needed to determine the moderating effect, if any, that

256 these factors exercise on the outcome of developmental studies conducted online.

257 It is also important to consider additional factors that could influence these results or

258 the way we interpret them. Chiefly, the current analysis is quite coarse-grained and considers

259 one particular dichotomy within study modality: in-person vs online. Yet, there are many

260 ways that developmental studies can be further subdivided. For example, studies are

261 conducted both in quiet spaces (e.g., in lab, at home) and loud spaces (e.g., parks, museums).

262 Therefore, online studies might out- or underperform studies conducted in particular

263 in-person locations. Our moderators are also correspondingly course-grained, particularly

264 dependent measure (looking vs verbal). Qualitatively, unmoderated studies with looking

265 measures had the smallest effect sizes relative to their in-person counterparts. However,

266 smaller effect sizes online could reflect true non-replications of the in-person results rather

than a lack of online studies' sensitivity. Because our small sample size renders our analysis underpowered to detect weaker effects of moderators, the current results and their interpretation are subject to change as online methods improve and comparisons to in-person studies are better understood.

Although developmental researchers have had decades of experience designing and running experiments in-person, most have only had a few years or less of experience developing online studies. Thus, our meta-analysis might underestimate the effectiveness of online studies due to researcher and experimenter inexperience. Over the next several years, as developmental researchers develop expertise and experience with online studies, effect sizes might increase for any number of reasons, including better experimenter-participant interactions, better stimulus design, and more accurate methods of measurements (i.e., automatic looking time measures, see Erel et al., 2022). Relatedly, as new methods are developed and adapted for online experiments, researchers should not take the current findings as a blanket declaration that all online studies produce comparable results to their in-person counterparts; some might underperform, while others might outperform. Nonetheless, the current results suggest that across currently employed developmental methodologies, studies conducted with children online are generally comparable to those conducted in-person.

The composition of our sample might also bias our results. To match online and in-person methods as closely as possible, we only considered direct online replications for the current meta-analysis. While this approach ensures that data were collected online and in-person using similar methods and procedures, it limits our sample size and may bias our sample. For example, perhaps researchers disproportionately choose to conduct online replications of strong or well-established effects rather than replicate more subtle, weaker effects. Nonetheless, our analysis found no significant publication bias in terms of favoring stronger online effect sizes or non-replications among the studies we sampled. We also

included an open call for unpublished data in an attempt to limit the file drawer problem (see Rosenthal, 1979). Of the published and unpublished online replications that were available to include in our sample, we found comparable effect sizes online (compared to in-person); however, researchers should exercise caution as this sample may not be representative for their particular questions of interest.

Conclusion

Although online data collection precludes certain research methodologies or measures (e.g., exploration of a physical environment), the general similarity in outcomes for in-person and online studies with children paint an optimistic picture for online developmental research going forward. However, beyond enabling the collection of high quality, low cost data, online research also stands to benefit the broader scientific community as a whole. Conducting studies online allows researchers to sample beyond the local community surrounding their home institution. And importantly, for many online participants, an online study with a developmental researcher is their first interaction with a scientist. As online research expands among developmental researchers, we are presented with an unprecedented outreach opportunity to directly interact more closely with those we hope our research will allow us to better understand and help – parents and children.

References

* Aboody, R., Yousif, S. R., Sheskin, M., & Keil, F. C. (2022). Says who? Children consider informants' sources when deciding whom to believe. *Journal of Experimental Psychology: General.*

* Bánki, A., Eccher, M. de, Falschlehner, L., Hoehl, S., & Markova, G. (2022). Comparing online webcam-and laboratory-based eye-tracking for the assessment of infants' audio-visual synchrony perception. *Frontiers in Psychology*, 6162.

* Bochynska, A., & Dillon, M. R. (2021). Bringing home baby euclid: Testing infants' basic shape discrimination online. *Frontiers in Psychology*, 6002.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). *Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?*

* Bulgarelli, F., & Bergelson, E. (2022). Talker variability shapes early word representations in english-learning 8-month-olds. *Infancy*, *27*(2), 341–368.

* Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., et al.others. (2021). Moderated online data-collection for developmental research: Methods and replications. *Frontiers in Psychology*, 4968.

Chuey, A., Lockhart, K., Sheskin, M., & Keil, F. (2020). Children and adults selectively generalize mechanistic knowledge. *Cognition*, *199*, 104231.

Chuey, A., McCarthy, A., Lockhart, K., Trouche, E., Sheskin, M., & Keil, F. (2021). No guts, no glory: Underestimating the benefits of providing children with mechanistic details. *Npj Science of Learning*, *6*(1), 1–7.

* DeJesus, J. M., Venkatesh, S., & Kinzler, K. D. (2021). Young children's ability to make predictions about novel illnesses. *Child Development*, *92*(5), e817–e831.

* Dillon, M. R., Izard, V., & Spelke, E. S. (2020). Infants' sensitivity to shape changes in 2D visual forms. *Infancy*, *25*(5), 618–639.

Erel, Y., Potter, C. E., Jaffe-Dax, S., Lew-Williams, C., & Bermano, A. H. (2022). iCatcher: A neural network approach for automated coding of young children's eye movements.

337   *Infancy*, *27*(4), 765–779.

338   * Escudero, P., Pino Escobar, G., Casey, C. G., & Sommer, K. (2021). Four-year-old's online

339   versus face-to-face word learning via eBooks. *Frontiers in Psychology*, 450.

340   Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., . . . Stiles, J.

341   (1994). Variability in early communicative development. *Monographs of the Society for*

342   *Research in Child Development*, i–185.

343   * Gasparini, C., Caravale, B., Focaroli, V., Paoletti, M., Pecora, G., Bellagamba, F., . . .

344   Addessi, E. (2022). Online assessment of motor, cognitive, and communicative

345   achievements in 4-month-old infants. *Children*, *9*(3), 424.

346   * Gerard, J. (2022). The extragrammaticality of the acquisition of adjunct control. *Language*

347   *Acquisition*, *29*(2), 107–134.

348   * Hamlin, J. (2015). The case for social evaluation in preverbal infants: Gazing toward one's

349   goal drives infants' preferences for helpers over hinderers in the hill paradigm. *Frontiers*

350   *in Psychology*, *5*, 1563.

351   Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First*

352   *Language*, 01427237211066405.

353   * Kominsky, J. F., Shafto, P., & Bonawitz, E. (2021). "There's something inside": Children's

354   intuitions about animate agents. *PloS One*, *16*(5), e0251081.

355   * Lapidow, E., Tandon, T., Goddu, M., & Walker, C. M. (2021). A tale of three platforms:

356   Investigating preschoolers' second-order inferences using in-person, zoom, and lookit

357   methodologies. *Frontiers in Psychology*, *12*, 731404.

358   * Lo, C. H., Rosslund, A., Chai, J. H., Mayor, J., & Kartushina, N. (2021). Tablet

359   assessment of word comprehension reveals coarse word representations in

360   18–20-month-old toddlers. *Infancy*, *26*(4), 596–616.

361   Lourenco, S. F., & Tasimi, A. (2020). No participant left behind: Conducting science during

362   COVID-19. *Trends in Cognitive Sciences*, *24*(8), 583–584.

363   * Man, N. (2022). *Thematic priming*.

364 * Margoni, F., Baillargeon, R., & Surian, L. (2018). Infants distinguish between leaders and

365   bullies. *Proceedings of the National Academy of Sciences*, *115*(38), E8835–E8843.

366 Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., . . . Stewart, L.

367   A. (2015). Preferred reporting items for systematic review and meta-analysis protocols

368   (PRISMA-p) 2015 statement. *Systematic Reviews*, *4*(1), 1–9.

369 * Morini, G., & Blair, M. (2021). Webcams, songs, and vocabulary learning: A comparison

370   of in-person and remote data collection as a way of moving forward with child-language

371   research. *Frontiers in Psychology*, 3347.

372 * Nelson, P. M., Scheiber, F., Laughlin, H. M., & Demir-Lira, Ö. (2021). Comparing

373   face-to-face and online data collection methods in preterm and full-term children: An

374   exploratory study. *Frontiers in Psychology*, 5025.

375 Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in

376   developmental psychology: A call to action. *Journal of Experimental Child Psychology*,

377   *162*, 31–38.

378 * Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor

379   the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216.

380 Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., . . .

381   Ocampo, J. D. (2020). Advancing developmental science via unmoderated remote

382   research with children. *Journal of Cognition and Development*, *21*(4), 477–493.

383 Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological

384   Bulletin*, *86*(3), 638.

385 * Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online testing yields

386   the same results as lab testing: A validation study with the false belief task. *Frontiers in

387   Psychology*, 4573.

388 * Scott, K., Chu, J., & Schulz, L. (2017). Lookit (part 2): Assessing the viability of online

389   developmental research, results from three case studies. *Open Mind*, *1*(1), 15–29.

390 Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental

391   research. *Open Mind*, *1*(1), 4–14.

392   Sheskin, M., & Keil, F. (2018). *TheChildLab. Com a video chat platform for developmental*

393   *research*.

394   Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al.others.

395   (2020). Online developmental science to foster innovation, access, and impact. *Trends in*

396   *Cognitive Sciences*, *24*(9), 675–678.

397   * Silver, A. M., Elliott, L., Braham, E. J., Bachman, H. J., Votruba-Drzal, E.,

398   Tamis-LeMonda, C. S., . . . Libertus, M. E. (2021). Measuring emerging number

399   knowledge in toddlers. *Frontiers in Psychology*, 3057.

400   * Skerry, A. E., & Spelke, E. S. (2014). Preverbal infants identify emotional reactions that

401   are incongruent with goal outcomes. *Cognition*, *130*(2), 204–216.

402   * Smith-Flores, A. S., Perez, J., Zhang, M. H., & Feigenson, L. (2022b). Online measures of

403   looking and learning in infancy. *Infancy*, *27*(1), 4–24.

404   * Smith-Flores, A. S., Perez, J., Zhang, M. H., & Feigenson, L. (2022a). Online measures of

405   looking and learning in infancy. *Infancy*, *27*(1), 4–24.

406   * Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning

407   and exploration. *Science*, *348*(6230), 91–94.

408   Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in

409   meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of*

410   *Clinical Epidemiology*, *53*(11), 1119–1129.

411   * Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities

412   shape expectations about the future at 12 months and beyond. *Proceedings of the*

413   *National Academy of Sciences*, *104*(48), 19156–19159.

414   * Vales, C., Wu, C., Torrance, J., Shannon, H., States, S. L., & Fisher, A. V. (2021).

415   Research at a distance: Replicating semantic differentiation effects using remote data

416   collection with children participants. *Frontiers in Psychology*, *12*, 697550.

417   Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal*

418   *of Statistical Software, 36*(3), 1–48.

419   * Yoon, E. J., & Frank, M. C. (2019). Preschool children's understanding of polite requests.

420   *CogSci*, 3179–3185.

421   * Yuen, F., & Hamlin, K. (2022). *Replication of hamlin (2015)*.