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

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

[1] Stanford University, Department of Psychology

Author Note

The authors made the following contributions. Aaron Chuey: Conceptualization, Methodology, Formal analysis, Data Curation, Visualization, Writing - Original Draft; Veronica Boyce: Conceptualization, Methodology, Formal analysis, Data Curation, Visualization, Writing - Review & Editing; Anjie Cao: Conceptualization, Methodology, Formal analysis, Data Curation, Visualization, Writing - Review & Editing; Michael C. Frank: Conceptualization, Methodology, Formal analysis, Data Curation, Visualization, Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to Aaron Chuey. E-mail: chuey@stanford.edu

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 was slightly larger than the mean effect size of their counterparts conducted online, a mean difference of d=.14, but this difference was not significant (95% CI=[.38, -.08]). Additionally, we found no significant moderating effect of dependent measure, online study method, or age. Overall, the current meta-analysis finds no evidence that developmental data collected in-person yields significantly larger effect sizes than data collected online.

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

Word count: X

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

Introduction

Developmental researchers are interested in studying children's behavior, primarily by measuring their behavioral responses to experimental stimuli. Study sessions typically involve visits with local families in a laboratory setting or partnering with remote sites such as schools and museums. Although these interactions are a routine part of developmental research, they are time-consuming for both researchers and participants. Typical studies with dozens of infants or young children can require weeks or months of scheduling visits to a lab or many visits to testing sites. In-person testing also limits the participant pool to children living relatively close to the research site. Additionally, developmental research has been plagued by small, non-diverse samples even more so than research with adults due to limitations imposed by the demographics of the local population as well as the high costs of collecting data from children (Kidd & Garcia, 2022; Nielsen, Haun, Kärtner, & Legare, 2017).

Prior to the rise of video chat software, there were only limited alternatives to in-person interaction for collecting experimental behavioral data from children. However, with the development of inexpensive and reliable video conferencing technology in the 2010s, new frontiers began to emerge for developmental testing.[1] Researchers soon experimented with conducting developmental studies through video-chat platforms, which in theory broaden the pool of participants to anyone with ~~internet access~~ access to internet and an internet enabled device, at nearly any time and location. What began as a few research teams experimenting with online studies (e.g., Lookit: Scott & Schulz, 2017; The Child Lab: Sheskin & Keil, 2018; Pandas: Rhodes et al., 2020) quickly expanded to much of the field as researchers scrambled to conduct safe research during the Covid-19

[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.

pandemic. This shift in research practices has yielded many empirical publications where some or all of the data were collected online in addition to a growing literature on online methodology and best practices (for a recent review, see Chuey, Asaba, et al., 2021).

Some researchers may be eager to return to in-person testing, but online research is likely here to stay and may increase in frequency as communications technologies improve and become more accessible. Online testing has immense potential to change developmental science (Sheskin et al., 2020), much as crowdsourced testing of adults has changed adult behavioral science (Buhrmester, Kwang, & Gosling, 2016). This potential has yet to be fully realized, however, as researchers have yet to fully understand the strengths and weaknesses of this method, as well as how to recruit diverse populations for online studies. Despite undersampling certain populations (Lourenco & Tasimi, 2020), online studies nonetheless allow researchers to sample from a larger, broader pool of participants than ever before as access to the internet continues to increase worldwide. Large, low cost samples and remote cross-cultural research may even become a reality for developmental researchers in the coming years.

Is conducting developmental studies online an effective substitute for conducting them in-person, or do online studies yield systematically different effects? Direct comparison of effects measured in both modalities is critical to answering this question. Researchers have implemented a number of paradigms online and replicated their in-person findings, but the quality of data yielded from online studies in comparison to those conducted in-person more broadly is still largely unknown. Therefore, the current meta-analysis examines how data collected from children online compares to data collected from closely-matched studies in-person. ~~Importantly~~

On the one hand, there is good reason to suspect that modality has little influence over the strength of a study's effect. Fundamentally, studies conducted online and in-person utilize similar measures (e.g., looking time, verbal report) and use similar kinds of stimuli (e.g., moving objects, narrated vignettes). Additionally, experimenters need to contend with extraneous factors like inattention, environmental distractions, and participants' mood. On the other hand, meaningful differences in online and in-person interactions could yield a meaningful difference in the outcome of online and in-person studies, in either direction. In principle, researchers have more control over a child's environment in-person, and in-person studies are usually less susceptible to technical problems such as lag or auditory or visual fidelity issues. Nonetheless, participants typically complete online studies in a more comfortable, familiar environment - their own home. Any of these factors could tip the scales, yielding larger effect sizes in-person or online; as result, we do not make any predictions regarding the presence or direction of an effect of study modality. Additionally, online studies themselves are not a monolith, and differ in a multitude of ways including the presence of a live experimenter, dependent measure, and the age of the sample being tested.

Online studies are generally conducted in one of two formats: moderated and unmoderated. In moderated studies, a live experimenter guides participants through a study much like they would in-person, except online, typically via video-chat. Moderated studies are often operationalized as slide share presentations or videos shared with participants while the participants' verbal responses or looking is recorded. In unmoderated studies, conversely, participants complete a study without the guidance of a live experimenter. Instead, researchers create a preprogrammed module that participants or their parents initiate and complete according to instructions. Since no experimenter needs to be present and participants can participate at any time they choose, unmoderated studies offer the potential for fast, inexpensive data collection. However, since they lack an experimenter, participants' experiences also deviate more from in-person studies compared to moderated studies that retain the same core social interaction between experimenter and participant. Therefore, it is possible that data collected via unmoderated sessions is comparatively noisier since an experimenter is unable to focus children's attention or course correct like they can during a live interaction. We consider this possibility in the current meta-analysis.

Like developmental studies more broadly, online studies have also employed a number of dependent measures, including verbal measures and looking measures. Verbal measures are typically straightforward to record, while recording looking measures is more complex. Accurate looking measures require precise camera positioning and coding schemes, and are thus more likely to deviate from their in-person counterparts compared to studies that measure children's verbal responses. To that end, automated gaze annotation is currently being developed and represents an exciting future direction in online methodology (see Erel, Potter, Jaffe-Dax, Lew-Williams, & Bermano, 2022). We examine how the kind of dependent measure employed (looking vs. verbal) might moderate the difference between online and in-person results.

The final moderator we consider is participants' age. Online developmental studies have sampled from a wide age range, including infants (e.g., Dillon, Izard, & Spelke, 2020), toddlers (e.g., Lo, Rosslund, Chai, Mayor, & Kartushina, 2021), preschoolers (e.g., Schidelko, Schünemann, Rakoczy, & Proft, 2021), and elementary schoolers (e.g., Chuey, Lockhart, Sheskin, & Keil, 2020; Chuey, McCarthy, et al., 2021). Because online studies are often conducted in the comfort of their own homes, it is possible that children of all ages might benefit from this aspect of online studies. Conversely, because a child's environment is more difficult to moderate online, infant studies, which often rely on precise environmental setups, may suffer more when conducted online. In addition, as children get older they may gain more experience with on-screen displays, which can contribute to their performance in online studies. We test these competing age moderation hypotheses.

In sum, our meta-analysis addresses the question of whether effect sizes tend to differ across online and in-person experiments with children, and whether these differences are moderated by study format, dependent variable, or participant age.

Methods

~~PRISMA plot detailing our study screening process; numerical values represents the number of papers at each stage of the review process.~~ PRISMA plot detailing our study screening process; numerical values represents the number of papers at each stage of the review process.

We conducted a literature search following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) procedure (Moher et al., 2015); see Figure . For each set of studies determined to be an online replication, we calculated the effect size(s) and associated variance for the main effect of interest. We then conducted a series of random-effects multilevel meta-regressions to estimate the effect of online data collection, as well as three possible moderators (online study method, type of dependent measure, and participant age). Our preregistered data selection, coding, and analysis plan can be found at https://osf.io/hjbxf. The list of papers included in this meta-analysis is shown in Table 1.

Literature Search

Our goal was to find as many published and unpublished online replications of developmental studies as possible. However, because there is no common nomenclature for online replications and the studies themselves cover a wide range of research questions and methodologies, searching via specific terms or keywords was difficult and produced many irrelevant papers; as a result, we could not conduct a completely systematic review. 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 (CDS) and International Congress of Infant Studies (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 codable 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 incomplete or partial samples were not considered.

6. The online and in-person methods must be directly comparable. Some alteration to the study methods is expected when adapting an in-person study to be run online (e.g., ~~changing a preferential reaching measure into a preferential looking measure,~~

having children refer to objects by color instead of pointing~~, etc~~). However, we excluded any studies whose methodologies altered the nature of the task or the conclusions that could be drawn from them (e.g., manipulating the identity of an object instead of its 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 (2022) compared to Stahl and Feigenson (2015)	3	Look	Mod	13
Smith-Flores (2022) compared to Skerry and Spelke (2014)	2	Look	Mod	13
Lo et al. (2021)	1	Verb	Unmod	20
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

Paper	Pairs	Look	Mod	Age
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 and their associated variance (sample size, group means and standard deviation, and t and F statistics when applicable), as well as our preregistered moderators (study modality, data collection method, dependent measure, and participant age).

If a paper reported an effect size as cohen's d (referred to below as standardized mean difference, SMD), we coded it directly. Otherwise, we calculated the individual effect sizes for each main effect and each study (online and in-person) ~~using~~ via reported means and standard deviations, t statistic, or directly from the data if it was available using analysis scripts adapted from Metalab (e.g., Bergmann et al., 2018). If the main comparison was to chance performance, we first calculated log odds and then converted the effect size to cohen's d via the compute.es package in R (Del Re & Del Re, 2012). If a given study had multiple dependent measures or central hypotheses, we calculated an effect size and associated variance for each.

Analytic Approach

To determine whether study modality (online or in-person) moderated the size of the main effect of interest for each set of studies, we performed a random-effects multilevel meta-regression using the metafor package (Viechtbauer, 2010). The regression predicts effect size (SMD) with study modality as a fixed effect~~.~~, modeling individual experimental effect sizes with the coefficient of interest being the study modality predictor (online vs. in-person).

~~Our analysis reflects a key design choice for our meta-analysis. Naively, it might appear to be~~ Although it might seem possible to predict the size of the online-offline difference for ~~a particular study. But on examination~~each particular study, many papers are heterogeneous and contain multiple online studies for a single given offline study, or multiple measures ~~for~~ within the same study. In these cases, the appropriate difference was not always clear. ~~Further, many pairs of studies differed on some value of our chosen moderators.~~

~~To deal with these issues, we instead modeled individual experimental effect sizes, with the coefficient of interest being the study modality predictor (online vs. in-person). We~~ To ensure these incongruencies did not bias the analysis, we included two random intercepts in our models. The first random intercept controlled for variation between particular experiments (e.g., modeling the dependency between multiple measurements reported from a single experiment). The second controlled for variation between groups of participants (e.g., modeling the dependency between effect sizes from participants who completed a battery of tasks with multiple effects of interest).

To determine the effect of additional moderators ~~(-~~ online study method (moderated vs unmoderated), dependent measure (looking vs verbal), and participant age ~~), -~~ we conducted three additional multilevel meta-regressions each with an additional fixed effect plus the corresponding interaction with study modality. All analysis scripts were preregistered, and the code is available at https://osf.io/up6qn/?view_only=91ba54134dc24787b04dd8f3b3b70e1e.

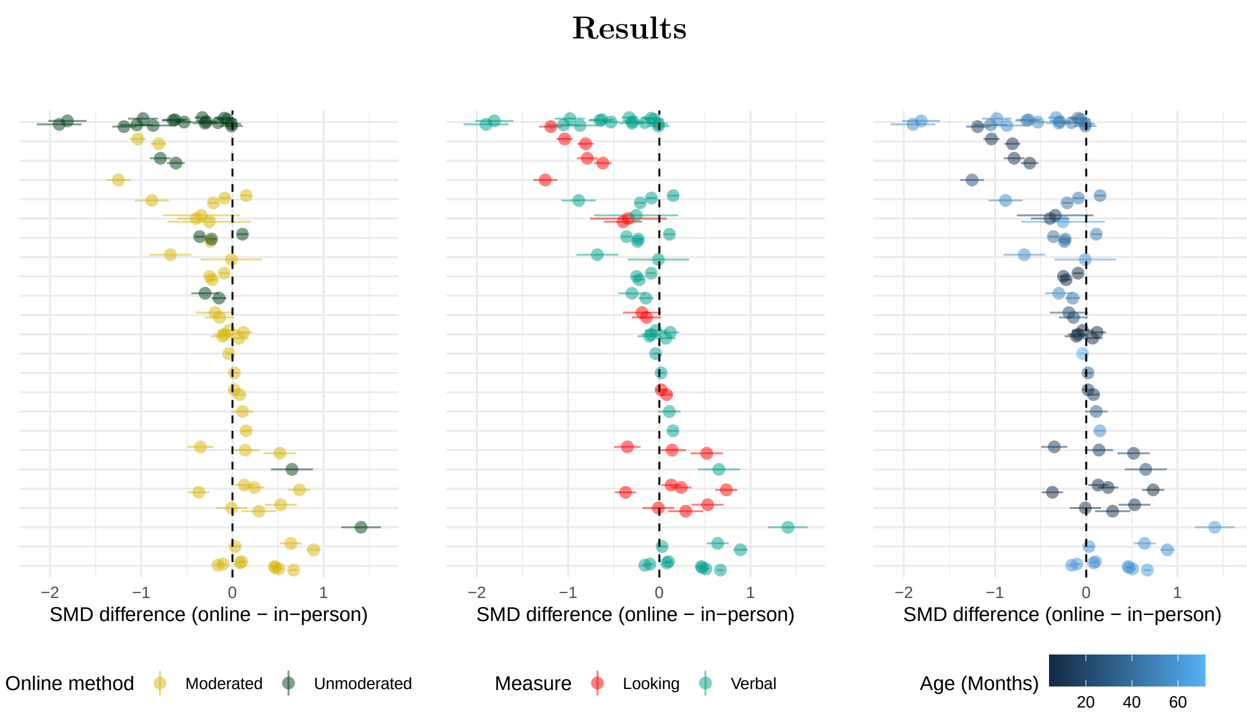

Figure 1. Forest plots of studies, sorted by difference in SMD. Each dot is the difference between and in-person measure and a corresponding online measure. Each row is one study (paper or pair of papers).

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

Planned 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 moderators and study modality. See Table 2 for coefficient values. Figure 1 shows the effect size differences of experiments by moderators.

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

Exploratory Analysis

In addition to our multi-level meta-analysis, we conducted an exploratory equivalence test to determine whether the effect sizes of studies conducted online and in-person meaningfully differed from one another, defined as a difference of d = .02 or greater (Hedges & Schauer, 2019). To aggregate the effect sizes by modality, we first averaged the effect sizes by study then across studies by modality. The equivalence test was inconclusive, finding that the effect sizes of online and in-person studies did not significantly differ from one another, 95% CI=[-.04, .37], but that they were also not statistically equivalent either, 90% CI [-.003, .34].

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

Modality	Method	Measure	SMD
Online	Moderated	Verbal	0.511
Online	Unmoderated	Looking	0.177
Online	Unmoderated	Verbal	0.570

~~As an exploratory analysis~~To investigate why the effect sizes of online and in-person studies in our sample might not be statically equivalent, we examined which combinations of methods and measures tended to yield the strongest and weakest effect sizes relative to their in-person counterparts. We fit a meta-analytic model containing method, response mode, and modality as well as 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.

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 2. According to Egger's regression test for funnel plot asymmetry, this plot is asymmetric (p=.005) and the ~~limit estimate of the~~

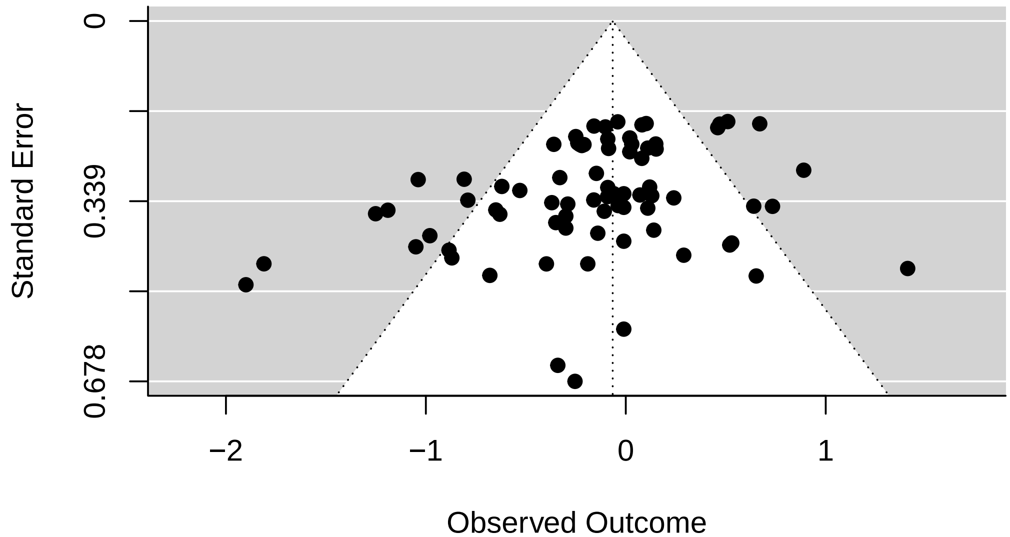

Figure 2. Funnel plot of the differences in effect size between pairs of in-person and online studies. A positive observed outcome means the online study had a large effect.

~~effect as standard error goes to zero~~ estimated effect assuming no variance is 0.37 [0.01, 0.72]. This analysis suggests the possibility of publication bias favoring ~~negative comparisons between online and in-lab modalities~~studies that have smaller effect sizes online compared to in-person, signaling that perhaps ~~the true effect of online testing is zero or perhaps even positive~~online studies may have relatively larger effect sizes on average compared to what has been reported. We interpret this conclusion with caution, however, noting the large width of the estimated CI and the relatively low power of Egger's test (Sterne, Gavaghan, & Egger, 2000).

Discussion

By aggregating across a growing literature of online studies, the current meta-analysis provides a birds-eye view of how developmental studies traditionally conducted in-person fare compared to closely matched counterparts conducted online. Our results suggest that overall, ~~the results of online studies are comparable to those~~ developmental studies conducted in-person do not yield significantly larger effect sizes compared to similar studies conducted online. Based on our analysis, the method of online data collection, type of dependent measure, and participant age did not ~~appear to~~ have a significant impact either. Nonetheless, ~~the relatively small sample size~~ our lack of statistical precision, indicated by relatively wide confidence intervals, limits our ability to make ~~sweeping generalizations about~~ strong conclusions about the effect of any of our moderators~~, so future~~. Future analysis is needed to determine the moderating effect, if any, that these factors exercise on the outcome of developmental studies conducted online.

It is also important to consider additional factors that could influence these results or the way we interpret them. Chiefly, the current analysis is quite coarse-grained and considers one particular dichotomy within study modality: in-person vs online. Yet, there are many ways that developmental studies can be further subdivided. For example, studies are conducted both in quiet spaces (e.g., in lab, at home) and loud spaces (e.g., parks, museums). Therefore, online studies might out- or underperform studies conducted in particular in-person locations. Our moderators are also correspondingly course-grained, particularly dependent measure (looking vs verbal). ~~Qualitatively~~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.

Nonetheless, unmoderated studies with looking measures noticeably had the smallest effect sizes relative to their in-person counterparts. This could reflect the relative difficulty of both collecting and coding looking data online using participants' own webcams. However, smaller effect sizes online could ~~reflect true non-replications~~ instead reflect genuinely smaller effect sizes 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~~Developmental research has suffered from many failures to replicate in the past, especially studies with infants (e.g., Davis-Kean & Ellis, 2019), and many of the online studies in our sample were conducted after their in-person counterparts, sometimes years later. Therefore, it is possible that smaller online effect sizes simply represent a more accurate estimation of the true (smaller) effect rather than an effect of study modaility per se.

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~~ also underestimate the potential 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~~ online studies might become more accurate at capturing cognitive constructs for any number of reasons, including better experimenter-participant interactions, better stimulus design (see Chuey, Asaba, et al., 2021), 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~~ that if publication bias exists, it likely favors stronger in-person 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 onlinereplications that were available to include in our sample~~

Conclusion

Our meta-analysis found that, across closely matched developmental studies conducted in-person and online, study modality did not have a significant effect on the size of the main effect of interest. Although we also found no significant effect of how the online studies were conducted (moderated vs unmoderated), dependent measures (looking vs verbal), or participant age, our lack of statistical precision limit our ability to make strong conclusions about their relationship to online replicability. In particular, we found ~~comparable effect sizes online (compared to~~ that unmoderated looking studies conducted online had noticeably smaller effect sizes relative to their in-person ~~)~~counterparts; however, ~~researchers should exercise caution as this sample may not be representative for their particular questions of interest.~~

~~Although online data collection precludes certain research methodologies or measures (e.g. , exploration of a physical environment)~~it is unclear whether this is reflective of the limits of online methodology in-person or a lack of replicability for these kinds of studies more generally. Therefore, as additional online replications are published, experimenters should also take their findings into account when considering the appropriate modality to conduct their particular flavor of developmental research. Nonetheless, the general similarity in outcomes for in-person and online studies with children paint an optimistic picture for online developmental research more broadly 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.

* Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009.

* Bochynska, A., & Dillon, M. R. (2021). Bringing home baby euclid: Testing infants' basic pre 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.

Davis-Kean, P. E., & Ellis, A. (2019). An overview of issues in infant and developmental research for the creation of robust and replicable science. *Infant Behavior and Development*, *57*, 101339.

* DeJesus, J. M., Venkatesh, S., & Kinzler, K. D. (2021). Young children's ability to make pre 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. *Infancy*, *27*(4), 765–779.

* Escudero, P., Pino Escobar, G., Casey, C. G., & Sommer, K. (2021). Four-year-old's online versus face-to-face word learning via eBooks. *Frontiers in Psychology*, 450.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., … Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i–185.

* Gasparini, C., Caravale, B., Focaroli, V., Paoletti, M., Pecora, G., Bellagamba, F., … Addessi, E. (2022). Online assessment of motor, cognitive, and communicative achievements in 4-month-old infants. *Children*, *9*(3), 424.

* Gerard, J. (2022). The extragrammaticality of the acquisition of adjunct control. *Language Acquisition*, *29*(2), 107–134.

* Hamlin, J. (2015). The case for social evaluation in preverbal infants: Gazing toward one's goal drives infants' preferences for helpers over hinderers in the hill paradigm. *Frontiers in Psychology*, *5*, 1563.

* Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses for studying replication: pre Meta-analytic perspectives. *Psychological Methods*, *25*(5), 557.

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, 01427237211066405.

* Kominsky, J. F., Shafto, P., & Bonawitz, E. (2021). "There's something inside": Children's intuitions about animate agents. *PloS One*, *16*(5), e0251081.

* Lapidow, E., Tandon, T., Goddu, M., & Walker, C. M. (2021). A tale of three platforms: Investigating preschoolers' second-order inferences using in-person, zoom, and lookit methodologies. *Frontiers in Psychology*, *12*, 731404.

* Lo, C. H., Rosslund, A., Chai, J. H., Mayor, J., & Kartushina, N. (2021). Tablet assessment of word comprehension reveals coarse word representations in 18–20-month-old toddlers. *Infancy*, *26*(4), 596–616.

Lourenco, S. F., & Tasimi, A. (2020). No participant left behind: Conducting science during COVID-19. *Trends in Cognitive Sciences*, *24*(8), 583–584.

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

* Margoni, F., Baillargeon, R., & Surian, L. (2018). Infants distinguish between leaders and bullies. *Proceedings of the National Academy of Sciences*, *115*(38), E8835–E8843.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., . . . Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-p) 2015 statement. *Systematic Reviews*, *4*(1), 1–9.

* Morini, G., & Blair, M. (2021). Webcams, songs, and vocabulary learning: A comparison of in-person and remote data collection as a way of moving forward with child-language research. *Frontiers in Psychology*, 3347.

* Nelson, P. M., Scheiber, F., Laughlin, H. M., & Demir-Lira, Ö. (2021). Comparing face-to-face and online data collection methods in preterm and full-term children: An exploratory study. *Frontiers in Psychology*, 5025.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38.

* Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216.

Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., … Ocampo, J. D. (2020). Advancing developmental science via unmoderated remote research with children. *Journal of Cognition and Development*, *21*(4), 477–493.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638.

* Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online testing yields the same results as lab testing: A validation study with the false belief task. *Frontiers in Psychology*, 4573.

* Scott, K., Chu, J., & Schulz, L. (2017). Lookit (part 2): Assessing the viability of online developmental research, results from three case studies. *Open Mind*, *1*(1), 15–29.

Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14.

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

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al.others. (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences*, *24*(9), 675–678.

* Silver, A. M., Elliott, L., Braham, E. J., Bachman, H. J., Votruba-Drzal, E., Tamis-LeMonda, C. S., … Libertus, M. E. (2021). Measuring emerging number knowledge in toddlers. *Frontiers in Psychology*, 3057.

* Skerry, A. E., & Spelke, E. S. (2014). Preverbal infants identify emotional reactions that are incongruent with goal outcomes. *Cognition*, *130*(2), 204–216.

* Smith-Flores, A. S. (2022). *Replication of skerry & spelke (2014).*

* Smith-Flores, A. S., Perez, J., Zhang, M. H., & Feigenson, L. (2022). Online measures of looking and learning in infancy. *Infancy*, *27*(1), 4–24.

* Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53*(11), 1119–1129.

* Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences*, *104*(48), 19156–19159.

* Vales, C., Wu, C., Torrance, J., Shannon, H., States, S. L., & Fisher, A. V. (2021). Research at a distance: Replicating semantic differentiation effects using remote data collection with children participants. *Frontiers in Psychology*, *12*, 697550.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

* Yoon, E. J., & Frank, M. C. (2019). Preschool children's understanding of polite requests. *CogSci*, 3179–3185.

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