DOI: 10.1111/infa.12353

RESEARCH ARTICLE

INFANCY THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES WILEY

Preregistration in infant research—A primer

Naomi Havron[1] | **Christina Bergmann**[2] | **Sho Tsuji**[3]

[1]Département d'Etudes Cognitives, École Normale Supérieure, École des Hautes Études en Sciences Sociales, Centre Nationale de la Recherche Scientifique, PSL University, Paris, France

[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[3]International Research Center for Neurointelligence (IRCN), The University of Tokyo, Tokyo, Japan

Correspondence
Naomi Havron, Département d'Etudes Cognitives, École Normale Supérieure, École des Hautes Études en Sciences Sociales, Centre Nationale de la Recherche Scientifique, PSL University, 29 rue d'Ulm, Paris 75005, France.
Email: naomi.havron@mail.huji.ac.il

Sho Tsuji, International Research Center for Neurointelligence (IRCN), The University of Tokyo, Med. 1 Bldg., Room N114, 7-3-1 Hongo Bunkyo-ku, Tokyo 113-0033 Japan.
Email: tsujish@gmail.com

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-11-0001-02, ANR-12-DSSA-0005-01 and ANR-17-EURE-0017

Abstract

Preregistration, the act of specifying a research plan in advance, is becoming more common in scientific research. Infant researchers contend with unique problems that might make preregistration particularly challenging. Infants are a hard-to-reach population, usually yielding small sample sizes, they can only complete a limited number of trials, and they can be excluded based on hard-to-predict complications (e.g., parental interference, fussiness). In addition, as effects themselves potentially change with age and population, it is hard to calculate an a priori effect size. At the same time, these very factors make preregistration in infant studies a valuable tool. A priori examination of the planned study, including the hypotheses, sample size, and resulting statistical power, increases the credibility of single studies and adds value to the field. Preregistration might also improve explicit decision making to create better studies. We present an in-depth discussion of the issues uniquely relevant to infant researchers, and ways to contend with them in preregistration and study planning. We provide recommendations to researchers interested in following current best practices.

1 | INTRODUCTION

Amidst increased interest in research transparency and reproducibility in the field of psychology (e.g., Munafò et al., 2017), preregistration is becoming more common for studies testing a priori hypotheses. This can, for instance, be observed on one of the major preregistration platforms in psychology, the Open Science Framework, where preregistrations have been doubling each year from 38 in 2012

to over 12,000 in 2017 (Nosek & Lindsay, 2018). In a preregistration, a researcher describes their hypotheses, methods, and analyses before they conduct their study, and (optionally) posts these decisions to a repository so that they are transparent to others reading the final work (Nosek, Ebersole, DeHaven, & Mellor, 2018).

1.1 | Preregistration in the context of questionable research practices

Psychology and related fields have recently faced a crisis of confidence (e.g., Ioannides, 2005; Pashler & Wagenmakers, 2012), brought about by reports of low replicability of central findings in large-scale replication attempts (e.g., Open Science Collaboration, 2015). Preregistration has been proposed as one way of improving this state of affairs (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). By preregistering the variables to be analyzed, sample size and stopping rule, exclusion criteria, and so on, researchers can avoid unintentionally engaging in questionable research practices. For example, researchers who are interested in testing the hypothesis that young infants can distinguish between surprising and unsurprising events may collect various measures believed to be markers of surprise: infant looking times, event-related potential (ERP) signals, pupil dilation, and social looks. The researchers may conduct *t* tests for the difference between surprising and unsurprising events for each measure, and find a significant difference between the two event types for looking times, but not for the other measures. When the researchers write their paper, they decide that only the looking-time data are interesting, and do not mention the other measures. Their paper, thus, concludes that infants can indeed distinguish between surprising and unsurprising events. The problem with this choice is that the readers of the paper believe only one statistical test was performed, and—with a standard significance level of $\alpha < 0.05$—would conclude that the probability of receiving this result by chance if there was in fact no difference between these two event types is only 5%. However, when the researcher has four different dependent variables, the chances of finding a significant effect on just one of these measures, even when the effect does not really exist, increase. Failing to control for, and report, such multiple comparisons is a questionable research practice. More than 60% of researchers in psychology admitted to have done this (John, Loewenstein, & Prelec, 2012). In a recent survey of infant researchers, however, only 7% of participants gave such an answer (Eason, Hamlin, & Sommerville, 2017)—though note that with negative attention such practices received since John et al. published their paper in 2012, social desirability may play a role in their responses. This and other questionable research practices are problematic because they tend to inflate the rate of false-positive results and undermine the reliability of results in science.

The practice of collecting more data after looking at the results (also known as N-hacking) has been common in psychology (John et al., 2012), and is considered a questionable research practice, as well.[1] Consider a researcher who was planning to test 30 infants. After testing 16 infants, the researcher decides that recruitment has been hard, and 16 infants might be enough. They look at the results for all four measures and discover an encouraging $p = .06$ on the looking-time measure. Given this "nearly" significant result, they complete their study as planned. Now they find their desired result of $p = .04$ for the looking-time measure. However, they used a subset of the data set twice for two dependent statistical tests, so their chance of finding a false positive has risen. This happens because inferential statistics such as the *t* test assume that the observed data are based on a single random sample of a prespecified sample size (for a more in-depth explanation, see Schott, Rhemtulla, & Byers-Heinlein, 2019). Here too, Eason et al. (2017) found a lower

[1]Though see Reinagel (2019), stating that as long as deviations are disclosed and alpha is corrected, they might be beneficial.

TABLE 1 Questionable research practices (QRPs), why they are problematic and what to do. Note that this list is not exhaustive

QRP	Definition	Time	Problem	Solution
Flexible stopping (N-hacking)	Deciding to add samples or stop early depending on intermediate test results	During data collection	Inflates false-positive rates by repeatedly testing the same data	Determine sample size a priori (including sequential sampling solutions) and/or report when peeking at data while still testing.
HARKing (hypothesizing after results are known)	Presenting exploratory results as confirmatory	When writing the paper	Results cannot be correctly interpreted	State all a priori hypotheses in advance and clearly label exploratory analyses
p-hacking	Conducting multiple statistical tests, applying different exclusion criteria, transforming variables, to obtain a significant result	During data analysis	Inflates false-positive rates	State all planned statistical tests, exclusion criteria, etc., in advance
Look elsewhere or Cherry picking	Reporting only one or some of the dependent measures	When writing the paper	Inflates false-positive rates	State all measures and variables collected or manipulated in advance
File-drawering	Deciding to not publish a study (author) or rejecting to publish a study (editor/reviewer) that is methodologically sound but does not show a significant result in the expected direction	When deciding whether to write a paper (author), or during the review process (editor/reviewer)	Inflates false-positive rates; distorts the publicly accessible record of research findings	Write up results and try to publish/archive in a publicly accessible repository (author); judge a study on soundness of research question and methods, not results (editor/reviewer)

percentage of undisclosed N-hacking in infant researchers (6%–17%, depending on interpretation of respondents' answer).

The effects of different questionable practices multiply (see Table 1 for some more examples of such practices). For example, Simmons, Nelson, and Simonsohn (2011) found, using simulations, that when a researcher uses two dependent measures (but reports one), continues testing a second batch after looking at the data, controls for gender or interaction of gender with treatment (after looking at the results), and optionally drops one of three conditions, the chances of finding a $p < .05$ become 60.7% (see Figure 1 for an illustration of possible post hoc decisions). Moreover, since this simulation only included four out of a large number of questionable research practices (see, e.g., John et al., 2012; Munafò et al., 2017), it can then be easy to see how making such decisions post hoc and after knowing the results can potentially raise the false-positive rate to as much as 100% (Gelman & Loken, 2013; Simmons et al., 2011).

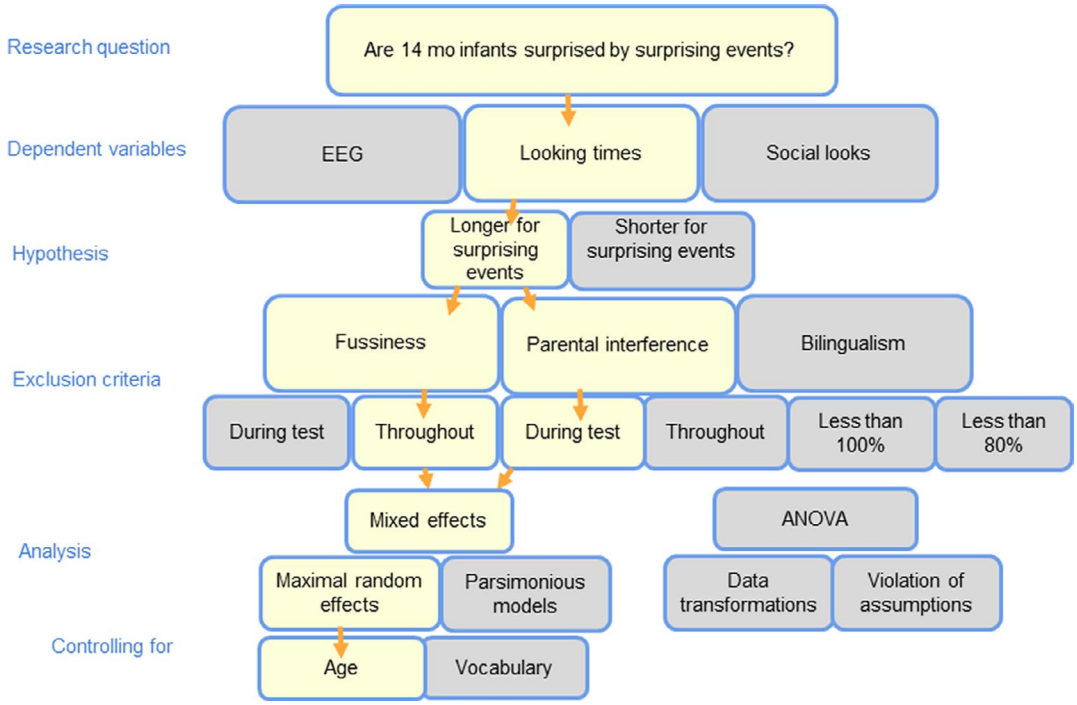

FIGURE 1 Some of the myriad of possibilities open to a researcher doing a study. In yellow, the path taken, and in gray, other paths not taken

Questionable research practices are not necessarily intentional, but may stem from post hoc rationalization or unawareness, either regarding the statistical consequences of such practices, regarding one's own plans in the past, or from reviewers' requests—most commonly additional analyses. Even if these analyses are more adequate than those the researcher originally planned, researchers *should report in the manuscript* when analyses were conducted only after seeing the results, and whether they were suggested by a reviewer or determined by the author to be more appropriate. Conversely, researchers who are not aware of what research practices are questionable might also promote the same questionable practices when reviewing papers—by suggesting additional analyses or by suggesting that the authors eliminate certain findings that were not significant.

A related key concept is *researcher degrees of freedom*, meaning that there exist many different possible choices for analysis, exclusion, etc. Researcher degrees of freedom inevitably inflate the false-positive rate, even without considering possible alternative choices explicitly. Thus, in Figure 1, we present some of the options open to a researcher who is planning or analyzing a study. In yellow, we mark the final decisions reported in the manuscript. In gray, we mark other options for hypothesis, variables, exclusion criteria, and analyses. The gray represents the researcher's degrees of freedom. If the researcher has preregistered the final analysis, then these degrees of freedom are not problematic. However, if the researcher chose one of these alternatives after the results were known, then they are inflating their false-positive rate. Preregistering studies reduces researcher degrees of freedom during confirmatory analyses, while never excluding exploratory analysis. We will not further discuss the value of exploratory research, because all remarks here concern exclusively confirmatory analyses, i.e., those that test hypotheses formulated before data were collected and/or inspected.

1.2 | The importance of preregistration in infant studies

Some of the problems with transparency and reproducibility that psychology faces as a whole might be even more inflated in infant studies. Particularly, the prevalence of small sample sizes due to difficulties in recruiting and testing participants (Bergmann et al., 2018), and more noisy measurements partly due to the low number trials infants can endure and partly to noisy paradigms, is likely to make it harder to detect an effect. Likewise, if an effect is significant, these reasons make it more likely be an overestimation or false positive (Ioannidis, 2005). Those studies that have nonsignificant results might remain, in turn, often unpublished (*file-drawer studies*). There could be different reasons why researchers choose to not publish a paper. Not publishing a sound result because it is inconsistent with one's perspective, theory, or hypothesis is a questionable research practice. Not publishing a sound result because after the fact the researchers realize that the study was uninformative does not have to be problematic. Additionally, researchers may decide that their study design or implementation was not sound after all, or is outdated, and choose not to submit it for publication. In the latter two cases, the researcher might still consider uploading a short summary of the study to a searchable location, such as figshare or PsyArXiv, since such work might be informative for researchers conducting a literature search or meta-analysis for future study planning.

Interestingly, the difficulty in testing infants may in itself reduce the field's overall rate of file-drawer studies. When experiments are cheap and easy to run, the cost of trying to publish null results can outweigh the cost of testing. For example, fields that commonly use online testing can test 200 participants a day (Stewart, Chandler, & Paolacci, 2017), thus potentially inflating the file-drawer problem. In infant studies, running a small experiment with 24 infants usually takes months, which might push researchers to publish regardless of outcomes. This might partially explain the large number of null results and little evidence for publication bias observed in a collection of meta-analyses on infant data (Bergmann et al., 2018; Tsuji, Cristia, Frank, & Bergmann, 2019).

Preregistration offers particular advantages to infant researchers who sample a population that is both difficult to recruit and to test: Preregistering sample size and rationale reduces the chances that sample size decisions are made on the fly in ways that raise the false-positive rate (e.g., continuing to test only if results are not significant without disclosing or correcting alpha). Even when sample size decisions were made a priori, a preregistration proves this for readers and reviewers. Moreover, preregistration allows to apply a technique especially relevant for researchers for whom recruiting large numbers of participants is challenging, namely planned data peeking. This technique can potentially reduce the necessary sample size without inflating type I error (see Schott et al., 2019, for a guide in infant studies).

In addition to helping avoid questionable research practices, preregistration also requires thorough thinking about the rationale, design, and analysis before the research is underway, thus potentially improving the quality of the work (see also Wagenmakers & Dulith, 2016). Given the difficulties of infant research, preregistration may increase the proportion of studies conducted that provide useful information for the community. By carefully deciding a priori on the design, analyses, exclusion, etc., researchers are less likely to conduct a study and discover some fatal flaw only after data collection is complete. We will discuss the necessary features of a preregistration that leads to a carefully designed confirmatory statistical analysis throughout this paper.

1.3 | Advantages of preregistration for the individual researcher

For the single researcher, preregistration can have practical advantages in addition to increasing the credibility of findings and reducing unconscious biases. Consistent, laboratory-level preregistration

for all studies can avoid duplicating work and can contribute to more efficient workflows. Concretely, preregistrations are a natural home for study-related laboratory policies, such as typical exclusion criteria. The survey mentioned above, by Eason et al. (2017), showed that particularly junior members of the same laboratory were unaware of existing policies. Along the same lines, preregistrations can efficiently document piloting policies, and typical experiment procedures (e.g., whether there are warm-up phases). When analysis scripts are shared along with preregistration documents, it is not only much easier to understand the purpose of these scripts, but they also become reusable for others. This way, re-creating similar code for comparable purposes is avoided, potentially increasing productivity. Detailed documentation, as instantiated by preregistration, can also work toward avoiding mistakes, which might be costly to correct later on (Rouder, Haaf, & Snyder, 2019).

1.4 | Challenges for preregistration

Preregistration helps reduce researchers' own biases and increase the trustworthiness and credibility of the completed work (Munafò et al., 2017; Nosek et al., 2018; Roettger, 2019). While it is not a guarantee that researchers will avoid faulty practice, it may contribute to reducing selective reporting of results and p-hacking; for instance, the number of positive results reported dropped from 57% to 8% after new regulations for obligatory preregistration in one area of clinical research (Kaplan & Irvin, 2015).

Despite this, preregistration in practice can suffer from some of the very problems it was meant to address: When comparing the content of authors' preregistrations and the final papers, numerous deviations have been found, some of which were undisclosed in the published paper (Claesen, Gomes, Tuerlinckx, & Vanpaemel, 2019). Deviating from a preregistration might be well justified. Undisclosed deviations, in contrast, are questionable scientific practice and do not increase research transparency. Moreover, authors engaging in them might falsely profit from the credibility that preregistration enjoys, thereby potentially undermining the credibility of preregistration. Note that disclosed deviations can also add researcher degrees of freedom (for details, see FAQ in the Supporting Information).

A related problem is imprecisely formulated preregistrations, which increase degrees of freedom by underspecifying their implementation (Uri, 2017). For example, in a looking-while-listening experiment, planning an analysis of looks to target without specifying the analysis time window leaves room to change that window after the results are known. There is no standard for preregistration, and therefore, it can be difficult to know the appropriate level of specificity. Some researchers would argue that a vague preregistration does not serve as protection against questionable research practices, and all steps of an analysis should be declared (for another example, see FAQ in Supporting Information). Others may argue that there is value even in an underspecified preregistration, although underspecification can do little to constrain researcher degrees of freedom. These researchers would claim that there is value in documenting any decision ahead of time, even if the research team is not (or not yet) ready to preregister all the variables, analysis window, electrode sites, etc. Thus, even preregistering the alternative ways of looking at the data—and providing some information about how the team will decide to use one or the other—can be beneficial. Unless or until journals, universities, or funding agencies decide on a standard definition of what preregistration should (minimally) contain, researchers should make the decision of how specific they can or should be in their preregistration depending on their own considerations and knowledge.

Questions about determining the exact analysis and responsibly deviating from it are especially relevant for the field of infant studies, where it can be difficult to predict an effect precisely enough to

decide on the exact analysis parameters in advance. Protocol deviations might therefore be frequent, and underspecified preregistrations, common. Consider the case of online word recognition, where infants are presented with two pictures on a screen while one of them is named, and their above-chance-level fixation onto the correct picture is taken as an indication of word recognition. Published studies vary in the exact time window they analyze. Indeed, factors such as native language, infant age, and word familiarity might affect the timing of this effect (e.g., Fernald, Zangle, Portillo, & Marchman, 2008). However, since word recognition has only been tested across a select number of languages, ages, and word types, it is difficult to predict the appropriate time window for a novel study (see Von Holzen & Bergmann, 2019, for analysis of the effect of different choices on study conclusions; Figure 3 and an in-depth discussion of different solutions for preregistration below).

Infant studies will probably continue to be costly to conduct, and the possibility of getting clean measures will remain limited. Therefore, the problem of measurement noise, data scarcity, and formulating precise predictions will prevail (though see some new developments, e.g., https://lookit.mit.edu/, an online platform for testing infants; Scott & Schulz, 2017). While we cannot offer a one-size-fits-all solution to dealing with protocol deviations or uncertainties in protocol planning, one way to reduce these is to include, where possible, decision tree type preregistrations, where potential points of deviation are anticipated and steps to decide on the form of deviations are described (though see Williams & Albers, 2019, for counterarguments). We also recommend establishing laboratory-level or community-level best practices that in the long run facilitate preregistrations because a large portion of possible decisions will have been made (c.f. ManyBabies Consortium, 2020, for a model in infant preferential looking research).

Finally, keep in mind that a preregistered study is not necessarily a high-quality study: Preregistration does not automatically improve a flawed research question or proposed methodology (e.g., Szollosi et al., 2019).

1.5 | Registered reports

Registered Reports (Chambers, 2013; Nosek & Lakens, 2014) refer to submitting a preregistration, in the form of the Introduction and Methods section of a journal article, to a journal for peer review before testing has begun and/or data have been inspected. Like a regular journal article, this manuscript can undergo several rounds of peer review and is accepted or rejected based on editorial decision. Registered reports have obvious advantages: Acceptance means (near) guaranteed publication, which is based on the evaluation of the research question and methodology, not significant outcomes. Because of expert peer review, design and analysis plan have likely improved throughout the review process, and many errors the researcher might otherwise have committed are avoided. In fact, the bar for registered reports can often be higher than for papers submitted after the fact. Reviewers must decide whether the study should be published no matter how the results turn out. That is why many journals ask authors for outcome-neutral criteria, which will enable researchers, readers, and reviewers to judge after the fact that any null results reported are not the result of some methodological or technical error.

One disadvantage is the time investment needed for peer review (though the overall process would not necessarily take longer time compared with a regular paper). Therefore, submitting a registered report might not always be realistic for researchers on a temporary contract. Moreover, like any format, registered reports are certainly not a cure-all solution, with prestudy expert review not being a guarantee for a sound study design. However, we recommend infant researchers to consider this form of publication when feasible, because improvements made via peer review are highly beneficial to infant researchers given the resource-intensive nature of their research.

2 | BEST PRACTICES IN PREREGISTRATION FOR INFANT STUDIES

At minimum, a preregistration should include the research hypothesis, the design (including the planned sample size), the exclusion criteria, and the analysis plan. Current platforms that allow to preregister online include the Open Science Framework and aspredicted.org. Other options for preregistration include repositories such as GitHub, personal websites (with time-stamped documents, such as created by tools in Google Drive), or creation of a time-stamped document where changes are transparent (e.g., a pdf that will not be altered).

2.1 | The research hypothesis

Stating and recording hypotheses before conducting the study can help avoid HARKing and make findings more credible. It is important to state the hypothesis clearly. It is possible to state openly that several patterns of results are possible, and what theory each would support. For example, a recent study testing 14- to 15-month-old infants' understanding of sentences preregistered an if-then hypothesis (Maillot et al., 2019): If infants of this age understand the sentences, they should behave one way, but if they do not understand these sentences, then they are expected to behave in another way. In this case, the authors did not hypothesize whether they will be able to understand these sentences.

Stating the directionality of the hypothesis is also important, if warranted based on theoretical or empirical considerations. Consider again the example of online word recognition: If infants recognized a word, looks to the named image should increase. There would be little reason to predict a decrease in looks to the named image in this task. Such a result is, however, statistically possible even if it does not reflect a true effect, and would constitute a so-called sign error (see Gelman & Carlin, 2014). If the researcher came across such a sign error and had not preregistered the expected direction, this would leave open the possibility of the post hoc rationalization of this direction of results. Given the rather strong theoretical and practical grounds to expect one particular direction of results in the present example, such post hoc rationalization might, in practice, not be very likely, but it could be in other cases (see, for instance, discussion on potential sign errors in preferential looking studies in Bergmann, Rabagliati, & Tsuji, 2019). Further, a directional hypothesis allows for more high-powered one-sided significance tests. Preregistering a one-tailed test shows that it was not selected post hoc as a way to p-hack.

2.2 | The design

A preregistration is a great opportunity to think through the rationale and soundness of a study, potentially resulting in an improved design and analysis plan.

2.2.1 | Sample size rationale

One crucial element of a preregistration is the inclusion of a sample size rationale. The probability to detect a true effect—the statistical power of a test—is a function of effect size and sample size. The smaller the size of the underlying effect, the more the participants are needed for high-powered studies (which are typically defined as 80% power or higher). Power analysis helps to estimate the necessary

number of participants for a given effect size and a desired level of power. A review of meta-analyses on infant language acquisition found that in this domain, studies are often underpowered, with a mean effect size of Cohen's $d = 0.45$ (Bergmann et al., 2018), but a median sample size of only 18 infants, amounting to a median power of only 44%. Although the awareness of the importance of power analysis has increased, researchers still often base their sample size decisions on suboptimal decision strategies, for instance, on one or a small number of previous studies (Anderson, Kelley, & Maxwell, 2017). This is also true for infant studies, where chosen sample sizes are on average closer to the sample sizes of seminal studies than to meta-analytic effect sizes (Bergmann et al., 2018).

In infant studies, sample size planning is particularly important. Moreover, infant studies often trace the emergence of an ability or contrast conditions, which means that often at least one study or condition is expected to show a null effect. Considering this special status of null effects, underpowered studies and the resulting high probability to miss true effects might be especially costly, since a false negative might be interpreted as evidence that an ability has not yet emerged. While it is common that researchers aim for 80% power—that is, the probability to detect a true effect in 8 of 10 studies—the more difficult it is to recruit and test participants, the higher the power a researcher should strive for (because the cost of not detecting an effect can be larger for them). In fact, when submitting a registered report for review, journals might ask for 90% power or a justification why this would not be feasible (see, e.g., *Developmental Science* submission guidelines and guidelines for registered reports on the OSF website: https://osf.io/pukzy/).

Calculating the necessary sample size should rely on effect sizes estimated based on existing literature. However, one should consider a few limitations. First, the literature might itself be misleading due to publication bias and/or underpowered studies. The calculated effect size based on a meta-analysis for the ManyBabies1 project (The ManyBabies Consortium, 2020) was $d = 0.72$ (from Dunst, Gorman, & Hamby, 2012), the actually achieved effect size in the study was about half (as is also common in large-scale replication efforts in other fields in psychology; e.g., Hagger et al., 2016; Open Science Collaboration, 2015). This is why it is recommended to conservatively assume that the effect is smaller than published effect sizes for an age range, method, and field.

Another way to calculate desired sample size is to run informed simulations. Simulations sample from a given population multiple times, and examine power with various-sized samples or with various effect sizes. Simulations can be run on an existing data set, such as raw results from a previous similar study. If no raw data are available, it is possible to simulate a full data set given the mean and standard deviation from a previous study, and the shape of the distribution of that sample (e.g., a normal distribution) using different tools (e.g., Goldfeld, 2019; see Havron, 2019 for Bayesian *t* test and ANOVA simulations). Using such simulations, Oakes (2017) found that with effect sizes of $d = 0.6$, testing 20–24 infants per cell could be enough—stressing the importance of knowing the target effect size.

The above examples imply that the planned study is very close to previous studies. However, this might not be the case. In order to still conduct a useful power analysis under uncertainty, it is important to lean on studies that are as close as possible, and consider them as a reference for the magnitude of the targeted sample size—for instance, it will be useful to know whether the sample should be around 20 or 200 infants. Across different phenomena studied in language acquisition research alone, required sample sizes would vary between around 10 and 300 infants (Bergmann et al., 2018). This estimate can then be adjusted based on additional considerations. For instance, if an effect size is based on several studies that used different methods, it is advisable to investigate whether differences in method might affect the effect size, and adjust expectations accordingly. The ManyBabies1 study found that the head-turn preference procedure yielded a larger effect than both a single-screen central fixation and eye-tracking (The ManyBabies Consortium, 2020).

While sample size decisions would ideally be based on a carefully conducted power analysis or simulations, in infant studies it is sometimes hard to have full control over sample size. An alternative to stating a target sample size is to specify other limiting factors, such as preregistering recruitment from a daycare, with an estimation that about 30 children will have parental consent. In such a preregistration, stating that all infants in a daycare who had parental consent will be tested, the final sample might include 26 infants, with five more dropouts. Similarly, since infant recruitment can be slow and irregular, it might be impossible to preregister a set sample size for a project with a strict deadline, for instance, for submitting a master's thesis. In this case, one could preregister, e.g., testing until a target of 40 infants or until July 30.

What if the effect is small, and thus requires a large number of infants to achieve desired statistical power, but it is unfeasible to test that many infants? One possibility is to recruit one or several additional laboratories and run the same study jointly. While this solution can be technically complex, the benefits of running a well-powered study, combined with the advantage of testing a more diverse pool of participants and learning from other laboratories' practices, can outweigh the challenges involved. A recent large-scale collaboration involved 67 laboratories worldwide (The ManyBabies Consortium, 2020; Frank et al., 2017). However, for most purposes it might be enough to contact one or two additional researchers for a joint study.

A final way to increase power and thus lower the necessary sample size is to reduce measurement noise. In infant research, this can be tricky, as typical strategies such as including more trials (Goulet & Cousineau, 2019; McClelland, 2000) might not be possible. Further, only few systematic investigations into method effects exist (but see Bergmann et al., 2018; The ManyBabies Consortium, 2020). If it is not feasible to reach the desired sample size, a longer-term plan might be to accumulate evidence over time and integrate it with meta-analytic methods (Tsuji, Bergmann, & Cristia, 2014) as new, small scale, and possibly noisy studies are being published (Braver, Thoemmes, & Rosenthal, 2014). This strategy requires single researchers to be cautious in their interpretation of results they know are likely underpowered, and for their consideration of previous work not only in a qualitative review (e.g., in Introduction section), but also when computing their results. This approach is further facilitated by the move toward open (anonymized) data, which makes joint analyses feasible even across studies that were conducted in locations and times.

A single study is never sufficient to establish an underlying effect, but the noise increases with lower power. Although preregistration cannot solve the problem of small sample sizes in infant studies, it is an opportunity to maximize the chances of finding whether an effect is real—without necessarily increasing sample size—through sequential testing. Here, the researcher peeks at the results at prespecified points and continues to test until reaching a final sample size conditioned on the results. Such responsible data peeking is more reliable and credible when preplanned and preregistered (Lakens & Evers, 2014). For a detailed tutorial and theory for ethical data peeking in infant studies using null-hypothesis significance testing, see Schott et al. (2019). In Bayesian statistics, researchers are more flexible to look at the data, among other reasons, because they test for both evidence against *and for* H0. This framework allows for testing until either sufficient evidence in either direction is accumulated, or until a final sample size has been reached (see, e.g., Schönbrodt & Wagenmakers, 2016). Here is an example from a registered report in infant studies (Havron, Babineau, & Christophe, 2020):

> We will employ sequential hypothesis testing with Bayes factors with a pre-specified inference criterion of BF10 > 3 as evidence for an asymmetry between the size or direction of the difference between trial types in the two conditions, or less than 0.3, indicating evidence for H0. The Bayes factors will be obtained from a Bayesian t-test using the JASP software (JASP team, 2018). The first analyses will be done after at least 20 infants have

been tested in each condition, then for every eight new infants (four in each condition). We will use a Cauchy prior of 1, as recommended by Schönbrodt et al. (2015), but will also examine other priors to assess the robustness of our results with different priors.

To conclude, while preregistration in itself will not make it easier for infant researchers to recruit more infants, it will increase the credibility of any result by ruling out N-hacking and can help reduce the overall sample size in ways we describe above, or to make a principled decision to collaborate with other laboratories. See Figure 2 for a diagram on how to plan your sample size.

2.2.2 | Inclusion and exclusion criteria

When specifying the population, the devil is in the details: For instance, "nine-month-old infants" could be either 8.5- to 9.5-month-olds or 9- to 10-month-olds. The more precise the criteria are, the better. It could also be more general, such as "all infants below one year whose parents consented to participate on day x in museum y." The specificity depends on resources and existing infrastructure, as well as the goals and scope of the study.

Exclusion decisions allow many degrees of freedom and can bias results if decided on after looking at the results (see Figure 1). Failure to define criteria in advance leaves the (theoretical) possibility to exclude an infant because they did not show the desired effect. Some exclusion criteria are straightforward, at least at first sight. For example, studies on language development typically exclude non-monolingual infants. However, what does "monolingual" mean? Infants who only hear one language from birth, or infants who hear the target language over 80% of the time (see Byers–Heinlein, 2015, for a discussion of defining who should be considered a bilingual infant)? Other exclusion decisions are even less straightforward. For example, what is a fussy baby? Out of 101 studies recently surveyed, only two mentioned an operational definition for fussiness (Slaughter & Suddendorf, 2007). Although results were not correlated with attrition rate, again, not defining criteria in advance leaves the possibility that researchers would exclude an infant who did not show the desired effect (for suggestions on preregistering criteria of fussiness, see FAQ in the Supporting Information). Parental interference is another criterion that leaves room for interpretation. For example, a parent pointing to the screen during an attention getter might not justify exclusion, but pointing during the actual trials might. Even with a list of well-defined criteria for parental interference, parents can interfere in unforeseen ways. Therefore, we recommend stating the motivation behind removing trials or infants where interference occurs, in addition to the criteria listed. This allows exclusion, when parents interfere in unforeseen ways that potentially affect infants' responses. Additional common exclusion criteria include experimenter error or technical problems (see FAQ in the Supporting Information). Another group of exclusion criteria are general participant-level criteria, such as premature birth, diagnosed developmental disorders, and repeated ear/eye infections. Finally, there are study-specific criteria, for instance, failure to habituate in habituation studies or gaze loss in eye-tracking studies. Data loss, within a trial or participant, is relevant because of increased noise. When deciding on exclusion criteria, on the one hand, noisy data reduce power (which may lead to stringent criteria, such as >25% gaze loss in eye-tracking studies), but on the other hand, stringent criteria may lead to more excluded participants (thus leading a researcher to decide on a laxer criterion such as >50% gaze loss). One way to take this into account in a preregistration is to use an objective, predefined criterion based on which one would adjust the data loss criterion after results were collected. See this example from the preregistration for a follow-up study of a published conference paper (Tsuji et al., 2019):

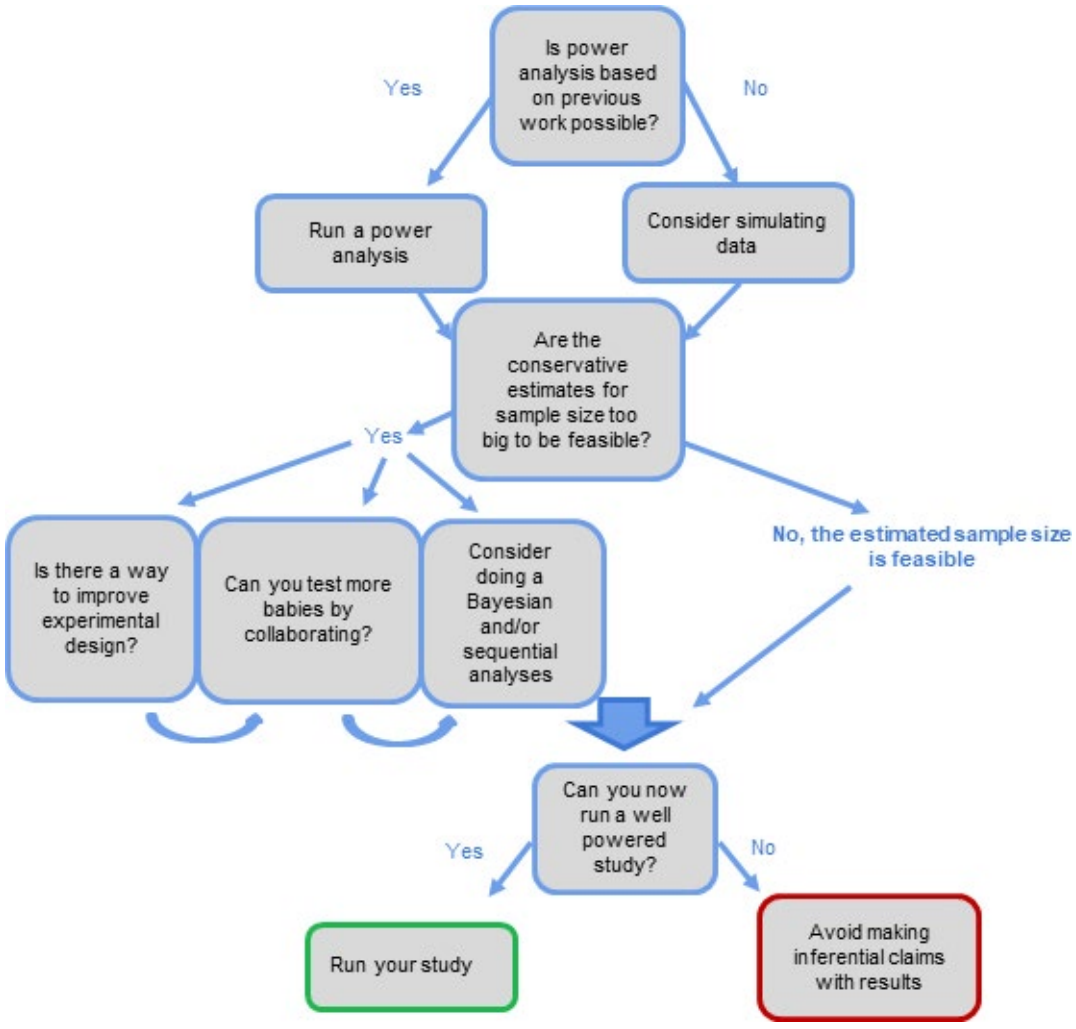

FIGURE 2 A diagram on how to plan your sample size

The above exclusion criteria rely on a cut-off of 50%. If, however, these criteria turn out to exclude an unexpectedly large number of toddlers (>25% of toddlers that are not excluded before data examination), we will adjust them to a criterion that allows us to include more toddlers. We will make available the original analysis and the rationale for adjustment when publishing the results.

2.2.3 | The study variables

Independent variables

Independent variables can be continuous or categorical. For example, when age is an independent variable, 14- and 16-month-old toddlers might be tested as two distinct age groups, so age is categorical. Alternatively, 14- to 16-month-olds might be tested, in which case age is a continuous variable. In the latter case, it is possible to dichotomize age for ease of analysis and presentation into "younger" and "older" infants. One common way to divide the sample is a *median split*. However, splitting a sample post hoc after seeing the results is again a questionable research practice, because splitting

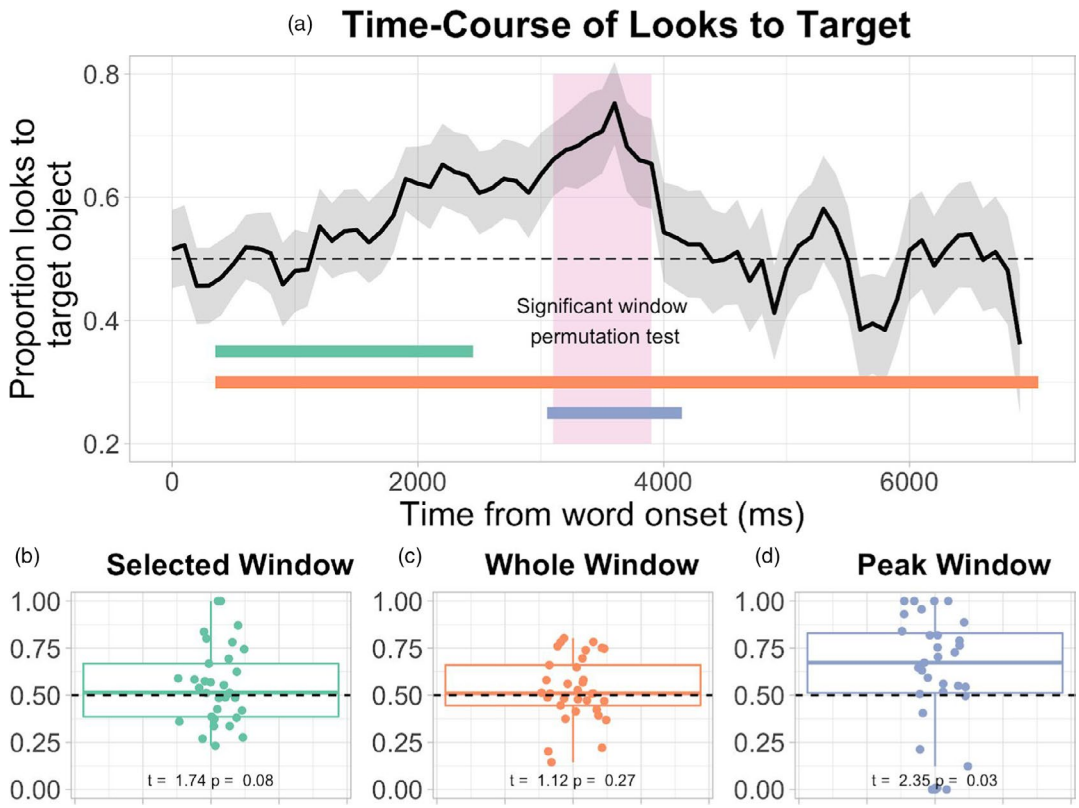

FIGURE 3 (a) Time-course of looks to named object in a looking-while-listening experiment. The dotted line represents chance level, and the solid black line represents the time-course of proportion of looks to named object. The green, orange, and blue lines represent the time window chosen for analysis under a strategy where a window is selected based on previous studies, the whole window is chosen, or a window is determined based on data structure, respectively (see text for more details). The red shaded area represents the time window in which a permutation analysis has shown a significant cluster of results. (b–d) Boxplot and aggregated by-subject point over the different time windows of analysis. Data from Crimon (2019)

allows many degrees of freedom (see FAQ in Supporting Information). Splitting a variable might also have other undesirable outcomes such as loss of information about individual differences; loss of effect size; spurious statistical significance; and loss of measurement reliability (MacCallum, Zhang, Preacher, & Rucker, 2002).

Experimental manipulation also requires detailed description. The preregistration should include as much detail as possible about how the two conditions will differ so that an independent researcher could construct a comparable study without further information. If possible, a description of the experimental manipulation should include links to the actual stimuli (videos, sound files, images, but note that not all stimuli can be shared freely). Additional control variables should also be mentioned in the preregistration. These may include gender, socioeconomic status, or any other factor of interest. In short, preregister all variables that will be collected.

Dependent variables

In infant studies, there are three key major types of dependent variables: parental reports/questionnaires, behavioral measures, and physiological measures/neural correlates. For all of these groups of variables, precision about the planned measurement is important.

The first group of measures, parental reports/ questionnaires, often relies on standardized tools or on lists of items developed for the study. In case there are several options, be mindful of the different kinds of tools, taking into consideration factors such as length and reliability (see also FAQ in the Supporting Information). Behavioral measures can be body movements such as head turns, or gaze measures such as looking times. One should mention how exactly these are measured—for instance, based on online observation or post hoc manual video coding? Another group of key dependent variables in infant studies are physiological measures such as heart rate, or neural correlates such as those assessed with an electroencephalogram (EEG). They are, in principle, comparable to behavioral measures recorded with a physical sensor like an eye-tracker, because they both sample infants' responses at a high frequency over the course of an experiment. However, since physiological measures are often multidimensional, the associated degrees of freedom are higher. Consider EEG data, which provide tens of channels that all can be analyzed in different ways (e.g., by summing over responses to events or subsetting to channels over target areas). Depending on the hardware, the maximum number of channels might also differ. In sum, we again recommend a level of specificity that allows an independent researcher to replicate the study without additional information (for a detailed discussion of preregistration in EEG research in general, see Paul, Govaart, Craddock, Schreiner, Schettino, 2020).

2.2.4 | The study procedure

This section concerns detailed aspects of the study beyond the hypothesis and variables—such as the order of study elements, the number of conditions, blocks or trials, how infants are assigned to conditions, and whether any randomization takes place. Specifying these aspects, again, reduces the researcher's degrees of freedom and prevents researchers from changing aspects of the study protocol without declaring it. We will elaborate on two crucial aspects of the study procedure that serve the additional purpose of reducing researcher bias: blinding and randomization.

Blinding

There is ample evidence that blinded and unblinded experiments systematically produce different results (e.g., Hróbjartsson et al., 2014). Will experimenters be blind to the experimental conditions and how will blinding be achieved? If the experimenter will not be blind, explain why this is either not an issue for your study (e.g., this is an eye-tracking study that proceeds without experimenter control and the experimenter is outside the testing booth), or why it is unavoidable (e.g., it is a museum study where the experimenter cannot be in a different room, thus cannot avoid monitoring the screen). Also mention whether parents will be blind to the experimental condition (e.g., wear headphones or opaque glasses). If they will not be, why is this unavoidable or not an issue?

Randomization

Specifying how to allocate participants to different experimental conditions helps avoid biased decision making. In a between-participant study, with no prespecified allocation of participants to condition, the experimenter might (subconsciously) allocate an infant that seems in a good mood to the experimental rather than the control condition, which might bias the outcomes. A preregistration should answer the following questions: Will participants be allocated to different experimental conditions randomly? How will randomization be achieved? Will the order of blocks or trials be randomized or counterbalanced and how?

2.3 | The analysis plan

Committing to an analysis plan not only reduces the chances of p-hacking, but also helps think through a research design critically and ask whether it fits the planned analyses. As a result, it becomes clearer whether the experiment neatly addresses the key question, and whether all necessary data are collected. It is important to write the analysis plan in a way that avoids ambiguities. Adding the actual formula or script can be an ideal way to supplement a verbal analysis plan. The rule of thumb should be that anyone would be able to replicate exactly the same analysis based on the information in this section.

2.3.1 | Data transformations

Data transformations are often only mentioned in passing, or not at all, in published papers. However, they add additional degrees of freedom. It is therefore important to think through all potential data transformations, and under which conditions to use them.

Dependent variables

The type of data transformation needed largely depends on the statistical models. For instance, parametric statistical models assume normally distributed residuals, and it might be recommended to transform the dependent variable logarithmically to fulfill this criterion. Since it may be difficult to estimate in advance whether transformations are needed, this is a good example of conditional preregistration: A researcher can preregister that after checking the assumptions of the statistical analyses, for instance, running a test for normality on looking-time data, they would transform them in case of non-normality (though see Williams & Albers, 2019, for arguments against such conditional decisions). Note that transformation is not the only possibility: A nonparametric statistical test might be more suitable if the data do not meet the criteria (for instance, a Mann–Whitney test instead of a *t* test).

Independent variables

A second type of data transformation concerns the transformation of independent variables in order to accommodate the statistical model. For instance, to interpret the intercept of a regression model, you might need to center your continuous independent variable. Similarly, for categorical independent variables, coding is key to the comparisons made in the statistical model (for details, see Daly, Dekker, & Hess, 2016).

2.3.2 | Statistical model

Deciding on a statistical model is one key piece of a preregistration. We illustrate how to preregister the statistical model of an infant study despite uncertainties.

Simple statistical models

For a within-subject comparison of two conditions, it may be sufficient to preregister a *t* test. However, even here there are some degrees of freedom: Is it a *t* test over mean or maximum looking times? Will the measure be based on all blocks and trials? Looking at the previous literature gives an indication where researchers' decisions differ. However, some of these decisions might not be explicitly mentioned. It is therefore always useful to create mock data and to run through an analysis. Here again, a conditional

preregistration might help. For instance, the preregistration could take into account that the final block potentially has noisy data, by declaring that the final block will be discarded if infants' attention as measured by their total looking time to the screen has dropped below half of their initial looking times.

More complex statistical models

The number of degrees of freedom increases tremendously in, say, a linear mixed-effects model. Preregistering "a linear mixed-effects model" is underspecified and leaves many analysis choices open after looking at the data. The most common method to choose a random effect structure is to use the "maximal effect structure *justified by the design* that allows the model to converge" (Barr, Levy, Scheepers, & Tily, 2013). However, this method leaves open many researcher degrees of freedom. If the model does not converge with the maximal effect structure, then how to choose which effects to remove? One option is to start with variables that are of little theoretical value (such as control factors)—but who will be the judge of whether the removed variables were really the least important? An alternative is to remove factors that explain the least variability. This is a more objective criterion, but this might concern factors that are of major interest. When preregistering the maximal effect structure to converge, descriptions of criteria that will be used to trim the model if it does not converge are key. A different option is to use a parsimonious mixed model (Bates, Kliegl, Vasishth, & Baayen, 2015), which in essence runs a principal component analysis on the random effects structure to determine the number of variance components and correlation parameters supported by the data (there is a function in R: rePCA in lme4, which was first introduced in RePsychLing, by Baayen, Bates, Kliegl, & Vasishth, 2015, that helps perform this calculation). The advantage of this procedure over the more prevalent use of the maximal effect structure is that it allows for a robust random effect structure without including unnecessary random effects and therefore losing power. Another advantage is that the procedure is automatic and leaves the researcher less room for interference based on their own biases. While further debate of the different methods to select your random effect structure is beyond the scope of this paper, it is important to know that different options exist and take the time to reflect which method best fits the research design.

Deciding on a statistical model under uncertainty

It can be difficult to decide a priori on crucial parameters of a statistical model. Infant time-series data, derived for instance from eye-tracking or event-related potentials (ERPs), are good examples. Consider the classical looking-while-listening paradigm (e.g., Fernald et al., 2008), where infants see two images side by side while one of them is named. A statistical test determines whether they looked to the named image significantly above chance as an indication of word recognition (see Figure 3). A key issue with this measure is the time window of the effect: The literature largely varies with respect to the time windows chosen for analysis (Von Holzen & Bergmann, 2019). Previous literature is often too scarce to reliably predict how the effect will unfold in a particular age group and language background. How, then, to preregister the analysis? We propose three possibilities. None of them is perfect, but each can lead to a solid preregistration if carefully put in place.

First, and probably most frequent, researchers could choose to base their analysis on the previous literature, either basing their time window on a study that comes close in design to theirs, or interpolating based on the time windows from multiple previous studies. An example can be seen in Figure 3 (green line in panel a; panel b): a preselected time window of 2,000 ms starting 400 ms after target word onset. The advantage of this method is that provided previous studies are a valid benchmark for future ones, chances are high that the time window preregistered actually corresponds to the relevant time window. However, what if the time window does not correspond to the results, as is the case for the real data in Figure 3? In that case, what is a reasonable way to adjust the analysis? As mentioned above,

a preregistration should never preclude later justified changes. Thus, if inspection of the data suggests that a different time window would be more suitable, it can be reasonable to add additional analyses. Determining a time window arbitrarily based on visual inspection is not recommended, but could be done in exploratory analyses as the starting point of a replication study using this time window.

The most objective ways to determine a new time window post hoc would be the next two solutions. First, researchers might choose to perform the analysis over the duration of the whole trial (orange line in Figure 3, panel a; panel c). Alternatively, they might choose an analysis such as the nonparametric permutation test (Maris & Oostenveldt, 2007; see Havron, de Carvalho, Fiévet, & Christophe, 2019, for an implementation in a visual world task with young children), which will determine the time windows of statistically significant differences in a bottom-up fashion without inflating false-positive rates (the time window in which the permutation analysis shows a significant effect is illustrated by the red shaded area in Figure 3, panel a). These choices have the advantage of not being tied to any specific predetermined time window. However, it comes with a caveat: Both methods might need a large effect in order to detect statistically significant differences, which is not necessarily the case in infant research. Therefore, these methods risk not detecting a true effect. In the present example in Figure 3, the whole-window analysis indeed does not reach significance, while the permutation test results in detection of a significant cluster—but the reverse can also happen. The second solution is to preregister a way to determine the time window conditionally on the data. For instance, researchers could preregister the time window as one second around the peak of the proportion of target looks (blue line in Figure 3, panel a; panel d). In our example, this approach leads to the selection of a window close to the significant cluster detected with the permutation approach, and finds a significant effect. This method needs to be used carefully: When comparing two conditions, choosing the time window around the peak difference between the conditions would inflate results. It is less biased to choose the peak mean looking time across conditions, or the peak for the baseline condition. This method has the advantage of accounting for both the necessary precision of the measure and unforeseeable divergences. However, it might not be implementable if the choice cannot be made based on a criterion independent of the effect of interest.

2.4 | Updating a preregistration

While researchers strive to make all possible decisions about the data, design, or analysis in advance, there are many unknowns when conducting an experiment. In any such case, it is always possible to upload an updated or amended version of a preregistration. One good example is exclusion criteria. There could be good reasons to exclude an infant for reasons that a researcher did not foresee. For example, a study might use online questionnaires about infants and preregister a set of exclusion criteria. When starting to process the data, the researcher might find that some of the questionnaires were not filled in by the parents of the infant, but by another relative who is not one of the primary caregivers (e.g., a grandmother). Excluding such questionnaires might make sense in some cases. When the researcher submits an amendment to a preregistration during or after data collection but before knowing the results, this is by no means a questionable research practice.

3 | CONCLUSION

Preregistration is a promising way to reduce the probability of a researcher—voluntarily or involuntarily—engaging in QRPs. However, a preregistration is only useful if done well. We elaborated on

the necessary degree of precision a preregistration should include, while providing suggestions specifically relevant for infant research.

Although a preregistration considering all the complexities and what-ifs of infancy research might seem daunting, it is an incredibly fruitful way to exhaustively and honestly examine a given research project before any resources are invested into data collection. We recommend that a researcher who cannot preregister all abovementioned decisions will still preregister whatever decisions they do find possible to make in advance. Moving forward, public preregistrations will illustrate all the decisions and uncertainties most infant research projects involve, and that stand behind the often streamlined story of a published article. We hope that increasing the number of preregistrations in infant research will not only directly work against questionable research practices, but also lead the field toward more transparency and openness.

PREREGISTRATION IN PRACTICE

In order to facilitate putting the above steps into practice, we suggest consulting our openly available preregistration checklist: https://osf.io/ekp4x/

We also created a Frequently Asked Questions document that provides more in-depth discussion of some issues mentioned here: https://osf.io/hf7zy/

ACKNOWLEDGMENTS

This work was partially funded by ANR contracts ANR-17-EURE-0017, ANR-11-0001-02 PSL, and ANR-12-DSSA-0005-01.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest with regard to the funding source for this study.

ORCID

Naomi Havron https://orcid.org/0000-0001-6429-1546
Christina Bergmann https://orcid.org/0000-0003-2656-9070
Sho Tsuji https://orcid.org/0000-0001-9580-4500

REFERENCES

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*(11), 1547–1562. https://doi.org/10.1177/0956797617723724

Baayen, H., Bates, D., Kliegl, R., & Vasishth, S. (2015). *RePsychLing: Data sets from psychology and linguistics experiments*. R package version 0.0.4.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. arXiv preprint arXiv:1506.04967.

Bergmann, C., Rabagliati, H., & Tsuji, S. (2019). *What's in a looking time preference?* https://doi.org/10.31234/osf.io/6u453

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. https://doi.org/10.1111/cdev.13079

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333–342. https://doi.org/10.1177/1745691614529796

Byers–Heinlein, K. (2015). Methods for studying infant bilingualism. In J. W. Schwieter (Ed.), *The Cambridge handbook of bilingual processing* (pp. 133–154). Cambridge, UK: Cambridge University Press.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610.

Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F., & Vanpaemel, W. (2019). *Preregistration: Comparing dream to reality*. https://doi.org/10.31234/osf.io/d8wex

Crimon, C. (2019). *Social cues and word learning in 12-month-old infants* (Unpublished master's thesis). Laboratoire de Sciences Cognitives et Psycholinguistique, DEC, École Normale Supérieure, PSL, Paris, France.

Daly, A., Dekker, T., & Hess, S. (2016). Dummy coding vs effects coding for categorical variables: Clarifications and extensions. *Journal of Choice Modelling*, *21*(September), 36–41. https://doi.org/10.1016/j.jocm.2016.09.005

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. Retrieved from Center for Early Literacy Learning website: http://www.earlyliteracylearning.org/cellreviews/cellreviews_v5_n1.pdf

Eason, A. E., Hamlin, J. K., & Sommerville, J. A. (2017). A survey of common practices in infancy research: Description of policies, consistency across and within labs, and suggestions for improvements. *Infancy*, *22*(4), 470–491. https://doi.org/10.1111/infa.12183

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. *Developmental Psycholinguistics: On-line Methods in Children's Language Processing*, *44*, 97.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Technical Report Department of Statistics, Columbia University.

Goldfeld, K. (2019). *simstudy: Simulation of Study Data*. R package version 0.1.15. Retrieved from https://CRAN.R-project.org/package=simstudy

Goulet, M. A., & Cousineau, D. (2019). The power of replicated measures to increase statistical power. *Advances in Methods and Practices in Psychological Science*, *2*(3), 199–213. https://doi.org/10.1177/2515245919849434

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Havron, N. (2019). *BayesianSimulations*. GitHub repository. Retrieved from https://github.com/NaomiHavron/BayesianSimulations

Havron, N., Babineau, M., & Christophe, A. (2020). *18-month-olds fail to use recent experience to infer the syntactic category of novel words*. Retrieved from psyarxiv.com/gak96

Havron, N., de Carvalho, A., Fiévet, A., & Christophe, A. (2019). Three- to four-year-old children rapidly adapt their predictions and use them to learn novel word meanings. *Child Development*, *90*(1), 82–90. https://doi.org/10.1111/cdev.13113

Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Rasmussen, J. V., Hilden, J., … Brorson, S. (2014). Observer bias in randomized clinical trials with time–to–event outcomes: Systematic review of trials with both blinded and non–blinded outcome assessors. *International Journal of Epidemiology*, *43*, 937–948. https://doi.org/10.1093/ije/dyt270

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of null effects of large NHLBI clinical trials has increased over time. *PLoS One*, *10*(8), e0132382. https://doi.org/10.1371/journal.pone.0132382

Lakens, D., & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*(3), 278–292. https://doi.org/10.1177/1745691614528520

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19. https://doi.org/10.1037/1082-989X.7.1.19

Maillot, S., Havron, N., Dautriche, I., Spelke, E., Ashur, P., & Christophe, A. (2019). *Fourteen month-old infants' understanding of sentences*. The 44th Annual Boston University Conference on Language Development, Boston, USA.

ManyBabies Consortium (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

McClelland, G. H. (2000). Increasing statistical power without increasing sample size. *American Psychologist*, *55*(8), 963–964. https://doi.org/10.1037/0003-066X.55.8.963

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., & Lakens, D. (2014). Registered reports. A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141. https://doi.org/10.1027/1864-9335/a000192

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, *31*(3). Retrieved from: https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science/comment-page-1

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Paul, M., Govaart, G. H., Craddock, M., Schreiner, M. S., & Schettino, A. (2020) *Preregistration of EEG projects – Discussion and guidelines*. [Manuscript in preparation].

Reinagel, P. (2019). *Is N-Hacking Ever OK? A simulation-based study*. BioRxiv, 2019.12.12.868489. https://doi.org/10.1101/2019.12.12.868489

Roettger, T. B. (2019). *Preregistration in linguistic research: Applications, challenges, and limitations*. https://doi.org/10.31234/osf.io/vc9hu

Rouder, J. N., Haaf, J. M., & Snyder, H. K. (2019). Minimizing mistakes in psychological science. *Advances in Methods and Practices in Psychological Science*, *2*(1), 3–11. https://doi.org/10.1177/2515245918801915

Schönbrodt, F. D., & Wagenmakers, E. (2016). *Bayes factor design analysis: Planning for compelling evidence*. https://doi.org/10.3758/s13423-017-1230-y

Schott, E., Rhemtulla, M., & Byers-Heinlein, K. (2019). Should I test more babies? Solutions for transparent data peeking. *Infant Behavior and Development*, *54*, 166–176. https://doi.org/10.1016/j.infbeh.2018.09.010

Scott, K., & Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14. https://doi.org/10.1162/OPMI_a_00002

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Slaughter, V., & Suddendorf, T. (2007). Participant loss due to "fussiness" in infant visual paradigms: A review of the last 20 years. *Infant Behavior and Development*, *30*(3), 505–514. https://doi.org/10.1016/j.infbeh.2006.12.006

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, *21*(10), 736–748. https://doi.org/10.1016/j.tics.2017.06.007

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2019). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *2*, 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science*, *9*(6), 661–665. https://doi.org/10.1177/1745691614552498

Tsuji, S., Cristia, A., Frank, M., & Bergmann, C. (2019). *Addressing publication bias in meta-analysis: Empirical findings from community-augmented meta-analyses of infant language development*. Accepted for publication in Zeitschrift für Psychologie.

Uri, J. F. (2017). *How to properly preregister a study*. Data Colada. Retrieved from http://datacolada.org/64

Von Holzen, K., & Bergmann, C. (2019). *The development of infants' responses to mispronunciations: A meta-analysis*. Preprint. Retrieved from https://psyarxiv.com/dze29

Wagenmakers, E. J., & Dutilh, G. (2016). Seven selfish reasons for preregistration. *APS Observer*, *29*(9). Retrieved from https://www.psychologicalscience.org/observer/seven-selfish-reasons-for-preregistration

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Williams, M. N., & Albers, C. (2019). Dealing with distributional assumptions in preregistered research. *Meta-Psychology*, *3*, 1–15. https://doi.org/10.15626/mp.2018.1592

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Havron N, Bergmann C, Tsuji S. Preregistration in infant research—A primer. *Infancy*. 2020;00:1–21. https://doi.org/10.1111/infa.12353