Should I test more babies? Solutions for transparent data peeking

Esther Schott[1]

Mijke Rhemtulla[2]

Krista Byers-Heinlein[1]

[1]*Concordia University*

[2]*University of California, Davis*

Acknowledgements: This work was supported by graduate fellowships to ES from the Fonds de Recherche du Québec - Société et Culture and Concordia University, by a grant to KBH from the Natural Sciences and Engineering Research Council of Canada (402470-2011), as well as the Concordia University Research Chairs program.

Author Note: All materials required to reproduce the analyses reported in this article are available at https://osf.io/qjx7t/ . Supplementary materials are available at https://osf.io/qd25s/.

Correspondence can be addressed to: Esther Schott, esther.schott@mail.concordia.ca

Abstract

Research with infants is often slow and time-consuming, so infant researchers face great pressure to use the available participants in an efficient way. One strategy that researchers sometimes use to optimize efficiency is data peeking (or "optional stopping"), that is, doing a preliminary analysis (whether a formal significance test or informal eyeballing) of collected data. Data peeking helps researchers decide whether to abandon or tweak a study, decide that a sample is complete, or decide to continue adding data points. Unfortunately, data peeking can have negative consequences such as increased rates of false positives (wrongly concluding that an effect is present when it is not). We argue that, with simple corrections, the benefits of data peeking can be harnessed to use participants more efficiently. We review two corrections that can be transparently reported: one can be applied at the beginning of a study to lay out a plan for data peeking, and a second can be applied after data collection has already started. These corrections are easy to implement in the current framework of infancy research. The use of these corrections, together with transparent reporting, can increase the replicability of infant research.

Should I test more babies? Solutions for transparent data peeking

1. Introduction

Experimental research with infants is challenging, as testing infants is a slow and sometimes cumbersome process (Frank et al., 2017). Infants can be difficult to recruit, and bringing them into the lab is both time-consuming and expensive. Once in the lab, many infants do not contribute useable data, as they may be fussy or inattentive. Even when they do contribute useable data, the variability of infant behavior means that experimental measures have varying levels of reliability (Cristia, Seidl, Singh, & Houston, 2016). Finally, infant behavior can be highly sensitive to particular choices made by the experimenter, for example the specific stimuli used, the length and number of trials, and the particular age at which infants are tested. For these reasons, infant researchers take great steps to "make every baby count", and rightly so.

One approach that has long been in the informal toolkit of infant researchers is to perform analyses (whether a formal significance test or informal eyeballing) on preliminary data, either once or multiple times, and to adjust data collection as a function of the results. Throughout this paper, we will refer to all types of data-dependent sample size selection as "data peeking". The goal of data peeking is typically to decide whether a study appears promising or not, and to adjust the data collection plan accordingly. In some cases, preliminary results will be discouraging. If the first few participants do not show the expected effect, the paradigm might be tweaked or abandoned without too many "wasted babies". In a climate where publication of null results in mainstream journals is still exceedingly difficult, this strategy is thought to minimize the number of participants run in studies doomed for the file-drawer. In a recent survey of developmental researchers (Eason, Hamlin, & Sommerville, 2017), a third of respondents

reported stopping a study early if preliminary results did not show the expected effect. Ethnographic studies of infant laboratories have also reported such practices (Peterson, 2016).

In contrast, sometimes data peeking indicates promising results. In this case, researcher behavior will often depend on the number of participants tested, and the results obtained. If a researcher peeks after only a few participants, they might decide to continue data collection when most infants show the desired effect. If a researcher peeks after a small but minimally acceptable sample size, they might add a few more participants if statistical tests do not quite reach conventional levels of statistical significance. If tests are significant, they might decide to end data collection. In the above-mentioned survey of developmental researchers (Eason et al., 2017), 23% of researchers reported they would add participants once after finding a non-significant but promising *p*-value, and an additional 12% reported they would add participants multiple times.

Of course, the issue of data peeking is closely linked with the question of how researchers decide how many participants to test. The recommended way is to conduct a power analysis based on the expected effect size (see Lakens & Evers, 2014 for practical recommendations), and to collect the full planned sample size. In this calculation, meta-analyses like Metalab (Bergmann et al., 2018) are a useful resource for researchers when they determine their sample size through a power analysis, keeping in mind that meta-analyses will overestimate the true effect size due to publication bias. In practice, researchers are likely to also consider resources like time, funding, and competing demands when running multiple studies (Miller & Ulrich, 2016). Yet, most discussions of sample size selection do not consider data peeking, a common and potentially useful tool for infant researchers.

In this paper, we argue that while data peeking can be intuitively appealing, it can have unintended negative consequences unless thoughtfully and transparently implemented. First, we explain why typical, uncorrected approaches to data peeking can lead to the very situations researchers are trying to minimize: wrongly concluding that an effect is present when it is not (Type-I error or false positive), and ending a study early even though a true effect is present (Type-II error or false negative). We provide examples and simulations to illustrate these situations. Second, we review two solutions for transparent data peeking: a pre-registered data peeking plan (specified prior to the beginning of data collection), and a post-hoc data-peeking plan (applied to an ongoing study). Finally, we discuss how infancy research can move towards transparent reporting of sample size selection – including data peeking – within the context of current research practice. We believe that transparent data peeking should be part of the growing toolkit for improved research, which includes meta-analysis (Bergmann et al., 2018), large-scale multi-lab studies (ManyBabies Consortium, accepted pending data collection; Byers-Heinlein et al., accepted pending data collection), large-scale data collection online and in museums (e.g., Frank, Vul, & Saxe, 2012; Tran, Cabral, Patel, & Cusack, 2017), and pre-registration (Lindsay, Simons, & Lilienfeld, 2016). Transparent data peeking can help infancy researchers to marry their desire for efficient data collection with doing research that replicates across labs.

2. Status Quo: Sample Sizes in Infancy

As we reviewed at the beginning of this paper, recruiting and testing babies can be a slow process. Thus, one problem that infancy research is facing is that the sample sizes typically found in infancy research are often too small for the effect sizes typically found in infancy research (Bergmann et al., 2018, Oakes, 2017). A review of 12 meta-analyses of infant language acquisition showed a median sample size of 18 participants, with a median effect size of Cohen's

d = 0.45 (Bergmann et al., 2018). This estimate of the median effect size, like any effect size derived from a meta-analysis, is subject to publication bias, and thus is likely an overestimation of the true effect size. To detect an effect of that size with 80% power in a paired-samples *t*-test, a sample size of 40 participants is needed. For between-subjects designs, such as for comparisons between age groups, 78 infants are needed. This disconnect between observed sample sizes and effect sizes shows itself in the average observed power: only three of the 12 meta-analyses had an average observed power above 80%. The power for the remaining effects ranged from 8% to 61%. To improve replicability, infant researchers have been calling for larger sample sizes (e.g., Bergmann et al., 2018; Oakes, 2017). We think that the strategies for transparent data peeking reviewed in this paper can help researchers to aim for larger, more appropriate sample sizes while keeping some flexibility in their data collection.

3. How Inferences Are Made

Most infant researchers analyze their data using inferential statistical tests such as *t*-tests, ANOVAs, and regressions. These tests typically rely on null hypothesis significance testing, where a *p*-value < .05 allows the researcher to reject the null hypothesis. In the following section, we review the concepts that underpin this framework, including *p*-values, sampling distributions, and Type-I errors. Readers comfortable with these concepts may skip directly to Section 3.2.

3.1. How *P*-values Work

Null hypothesis significance testing relies on the idea of a null hypothesis, which assumes that the true effect under investigation is zero. *P*-values are probabilities derived from the assumption that the null hypothesis is true. We will use a single-sample *t*-test as an example: Suppose that Dr. Stats performs an experiment with 24 infants, and wishes to determine if infants

show different-than-chance behavior (in this case, chance is 50%). To do so, Dr. Stats must first assume that the null hypothesis is true — that is, that the data were sampled from a distribution with a population mean of 50%. Then, she computes a *t*-statistic by scaling her observed difference by the standard error of the mean. This *t*-statistic will follow another distribution — the central *t* sampling distribution. The top left panel of Figure 1 shows the expected distribution of *t*-statistics in our hypothetical example when the null hypothesis is true. By comparing her observed *t*-statistic with this expected distribution, Dr. Stats finds a *p*-value. This *p*-value represents the long-run probability of the observed statistic (in this case, *t*) arising when the null hypothesis is true.

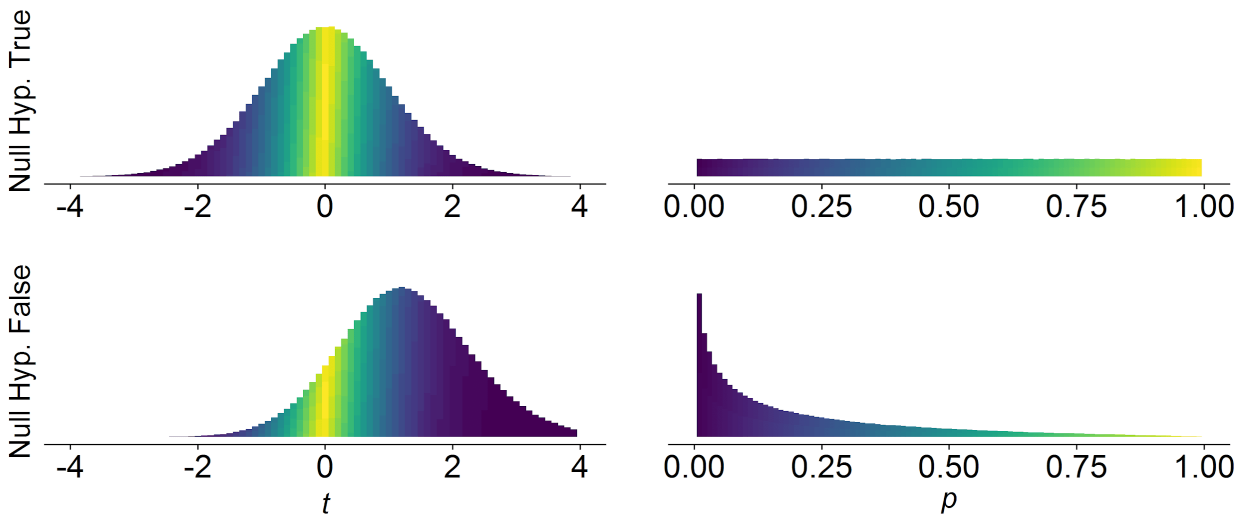

Figure 1. Distribution of *t*-statistics and *p*-values when the null hypothesis is true (top) and when there is a true effect size of .5 (bottom). Left: curves represent the likelihood of observing each value of *t*. Axes above these plots show *p*-values corresponding to values of *t*. Right: distributions represent the likelihood of observing each value of *p*. The color gradient

depicts the correspondence of particular *t*-values with *p*-values. The code used to generate these and all other figures are available at https://osf.io/qjx7t/.

Returning to Dr. Stats, she compares her obtained *p*-value to her cutoff alpha (traditionally .05, although see Benjamin et al., 2018, and Lakens et al., 2018, for a discussion on changing this standard). If Dr. Stats observes a *p*-value smaller than alpha, she rejects the null hypothesis, and concludes that her sample was drawn from a population with a mean other than 50%. The rationale is that if the null hypothesis was true, the obtained results would only arise 5% of the time. On the other hand, if Dr. Stats obtains a *p*-value larger than .05, she must conclude that her result is reasonably likely when the null hypothesis is true, so she retains the possibility that the population mean is equal to 50%.

Null hypothesis significance testing is a form of inferential statistics because the conclusion is simply an inference, and as such whatever conclusion Dr. Stats draws, it might be wrong. There are two types of error that Dr. Stats can make. A Type-I error (also known as false positive) occurs if she mistakenly rejects the null hypothesis, which is expected to happen *alpha* percent of the time when the null hypothesis is true. A Type-II error (false negative) occurs if she mistakenly retains the null hypothesis; the rate of Type-II errors depends on experimental power. Having reviewed the basics of inferential tests, we can now turn to how data peeking affects the conclusions we can draw from our data.

3.2 The Hidden Importance of Sampling

A key component of inferential tests like the *t*-test and *F*-test are the *t*- and *F*-distributions. These are called *sampling distributions* because they describe the shape of the distribution of test statistics over many repetitions of the same experimental procedure. The typical *t*- and *F*-distributions that we refer to for inference describe the sampling distribution

when data are sampled in a particular way, namely, when observations constitute a simple random sample of predetermined sample size *N* from a population with certain assumptions met (e.g., homogeneous variances). It is well-known that violations of key assumptions of these statistical tests (e.g., homogeneity of variance), can result in invalid inferences. It may be less well understood that using a non-random sampling procedure – such as data peeking – can also result in invalid inference.

4. Problems Resulting from Uncorrected Data Peeking

Having reviewed the basic concepts of null hypothesis significance testing in Section 3, in Section 4 we show how data-peeking violates a key assumption of null hypothesis significance testing. Violating assumptions can affect *p*-values in unanticipated ways, and increase Type-I error. We also discuss cases where data peeking can result in a Type-II error if a researcher abandons a study too early.

4.1 Why Data Peeking Can Lead to More False Positives: Type-I Error Inflation

To understand the effects of data peeking on the sampling distribution, we will explore the example of Dr. Marginal. Like Dr. Stats, her key question is whether participants in her study perform above chance (50%). Since we are interested in the effect of data peeking on Type-I error (which can only occur if the null hypothesis is true), we will assume that the true performance in her study is at chance, thus the standardized effect size in the population is $\delta = 0$. Dr. Marginal starts by testing 16 infants, which is what she assumes is a minimally acceptable sample size (although many would disagree with her assumption; see Bergmann et al., 2018; Oakes, 2017). She would, however, be open to collecting more data if necessary. If the result is not statistically significant at 16 infants, Dr. Marginal would add data in groups of 4 participants

until she either reaches a statistically significant result or until she has tested 32 infants. Each time Dr. Marginal peeks at her data, she evaluates her obtained *t*-statistic against the usual *t*-distribution (i.e., the *t*-distribution assuming random sampling), gets the associated *p*-value, and decides whether to continue data collection.

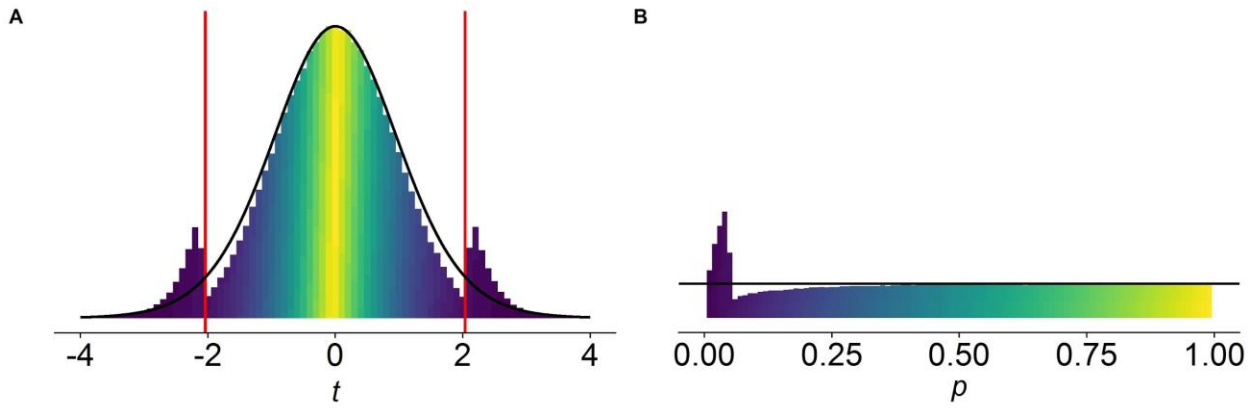

Figure 2. Distribution of *t*-statistics and *p*-values when the following data peeking rule is used: Begin with $N = 16$, then if $p > .05$, add $N = 4$ participants and recompute *p*, and so on until $N = 32$ or $p < .05$. Left: Distribution of *t*-statistics under the data peeking procedure. The black curve shows the regular *t*-distribution with 31 *df*; vertical lines represent critical values for a 2-tailed *t*-test based on the regular *t*-distribution. The shading of the distributions represent the correspondence between *t*- and *p*-values. Right: Distribution of *p*-values under the data peeking procedure. The black line shows the expected (uniform) distribution of *p*-values under a fixed-*N* procedure when the null hypothesis is true.

Dr. Marginal's data peeking procedure results in a sampling distribution that is different than the canonical *t*-distribution (see Figure 2). The *p*-values generated from the canonical *t*-distribution are no longer uniformly distributed when Dr. Marginal's sampling plan is used; she will observe more *p*-values below .05 than expected if the null hypothesis is true. Despite using an alpha of .05 for each *t*-test she evaluates, Dr. Marginal's actual Type-I error rate in our

simulation was 11.5% because her sampling procedure violates the assumptions of the *t*-distribution. For Dr. Marginal, this means that 11.5% of her studies will show statistically significant findings despite testing ideas with a true effect size of 0. The degree to which the Type-I error increases depends on the details of the data peeking procedure, as each procedure will lead to its own sampling distribution. Importantly, the sampling procedure itself is not inherently invalid. What is invalid is deriving *p*-values from the wrong sampling distribution.

As an even more extreme example, imagine Dr. Infinity, who is willing to continue adding participants indefinitely until the desired *p*-value is obtained. Recall that under the null hypothesis, each *p*-value is equally likely. As the sample size increases, the *p*-value will wander around randomly between 0 and 1. Dr. Infinity is guaranteed to eventually observe a *p*-value that dips below the alpha level. This means that with repeated data peeking and infinite potential participants, Dr. Infinity's Type-I error rate is 100%! In Figure 3, we show several simulations where the null is known to be true, and yet *p* crosses below the .05 threshold several times. Of course, hardly any researcher would collect data from thousands of participants waiting for a significant *p*-value, but less extreme decisions also matter. Thus, a researcher's data-dependent decisions about the final sample size affect the Type-I error rate (see also Sagarin, Ambler, & Lee, 2014, for further examples of how data peeking affects Type-I error rate).

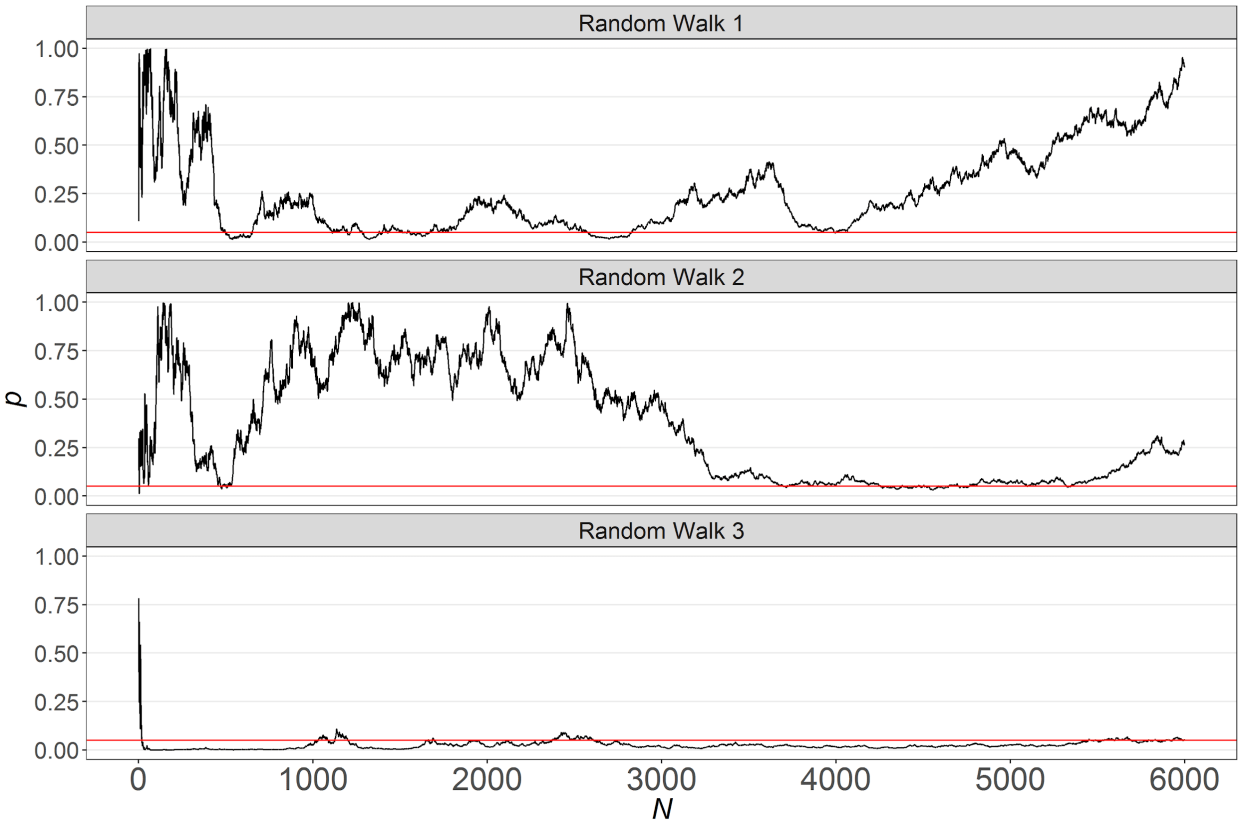

Figure 3. Three random walks of a *p*-value as each data point is added to the data set from $N = 1$ to $N = 6000$, where the null hypothesis is true. The red horizontal line indicates $p = .05$. Note that in each case the *p*-value dips below .05 numerous times, and in some cases even stays below .05 for a long stretch, despite the null hypothesis being true. Each different simulated example has a different random trajectory.

4.2 Data Peeking Without Statistical Tests: No *P*-values but Same Problem

In all examples above, the researcher performed a statistical test with preliminary data before deciding whether to end data collection or to increase sample size. What about when a researcher peeks at the data without performing a statistical test? As an example, imagine Dr. Headcount decides to pilot a few participants to see if a study is "working". She pilot tests six infants, and decides that if four or more of the infants show the expected direction of results

(here, for example, score above chance), she will test 10 more infants. If fewer than four infants show results in the expected direction, she will tweak the study (e.g., find more salient objects, add practice trials, etc.) and start data collection again (see e.g., Peterson, 2016). This procedure is also called "pilot dropping" (Simmons, Nelson, & Simonsohn, 2018).

Unfortunately, even though she does not perform statistical tests, Dr. Headcount nonetheless changes her data collection plans based on observed data, and will have similar problems as if she had conducted a formal statistical test. Figure 4 shows the distribution of test statistics under the null hypothesis when Dr. Headcount's procedure is used. Comparing Dr. Headcount's procedure (in red) to the *t*-distribution using a random sampling procedure (in black) shows that Dr. Headcount is more likely to commit a Type-I error; specifically, in 11.4% of simulations Dr. Headcount ended up with a false positive result. In cases where there is an explicit pilot dropping procedure rule, the sampling distribution under the null hypothesis can be simulated (as we did in Fig. 4), and a legitimate p-value can be obtained by comparing the observed test statistic to the simulated distribution. However, because informal pilot dropping procedures are rarely as explicit as the 4-out-of-6 rule described above, it is difficult to estimate their effect precisely, and thus there is no straightforward statistical correction for inflated Type-I error rates. A further consideration is the number of participants that end up being tested using a pilot dropping procedure (see Fig. 4B). Under Dr. Headcount's pilot dropping procedure, she would end up testing a median of 22 of babies (although since this is the median, half the time she ends up testing even more). Yet, because a number of these babies were tested in failed pilots, her final sample is always $N = 16$. Thus, when using pilot dropping, there is a risk of running through a lot of participants without increasing the power of the final study. In many cases, those "wasted" infants would be better invested in a single, larger-sample study.

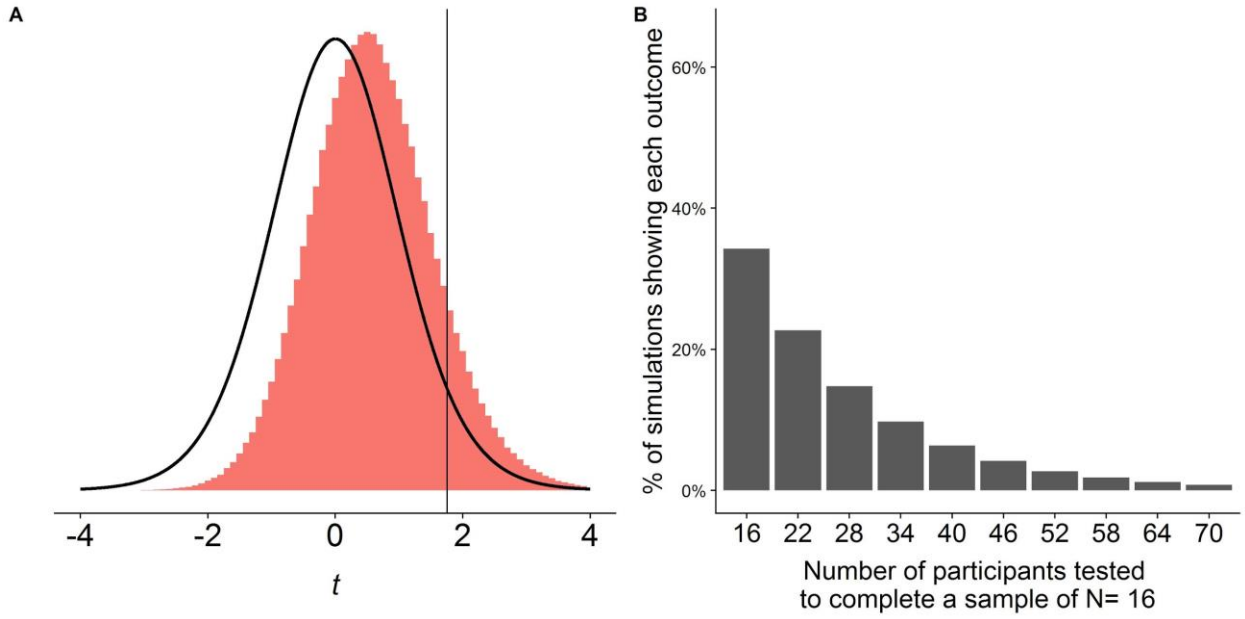

Figure 4. Pilot dropping when the null hypothesis is true. A. The black curve shows the distribution of test statistics when N = 16 infants are randomly sampled without inspecting the results of the first six participants; in red, the distribution of test statistics when the first six participants are kept only if at least four of them have a result in the anticipated direction. B. Distribution of the number of participants collected to achieve a single complete dataset of N = 16when the pilot dropping procedure is followed. The x-axis was truncated at N = 70.

This pilot dropping approach can also result in a Type-II error, as it is possible to observe unpromising early headcount results even if the study tests a true effect. Imagine Dr. Relentless, who dedicatedly pursues a great research idea. She uses the same 4-out-of-6 piloting procedure as Dr. Headcount. Every time she has a "failed" pilot, Dr. Relentless tweaks her study design before testing the same effect again with another six pilot infants. We will assume that the effect she tests is true (δ = .5 for this example), and her methodological adjustments do not affect the

true effect size. Our simulations show that Dr. Relentless ends up running an average of 19 participants to get to a final sample of 16 participants. Pilot dropping does increase her power to detect an effect for the times that she ends up augmenting the promising pilot data to a full sample of 16 (compared to running a single study with 19 participants without peeking). Yet, this increase in power does come at the expense of running participants in failed pilots. In Dr. Relentless' situation, even though she is testing a true effect, she will only observe a "successful" pilot around 73% of the time, thus about every fourth pilot will be ended as "failed". This example assumes that Dr. Relentless never gives up on the idea she is testing, and keeps piloting new tweaked versions even after another "failed" pilot. In reality, a researcher cannot know if the effect they are testing is true or not, and may end up abandoning the idea after a "failed" pilot, thereby committing a Type-II error.

The take-home message is that it can be counter-productive to peek at data too early: researchers risk increasing their Type I error, but may also commit a Type II error if they abandon their research idea after a failed pilot (recall that power is always lower with a smaller sample size). If researchers want to pilot a new experimental paradigm, the appropriate way to proceed is to exclude pilot data from the final sample, and begin data collection afresh with the experimental method locked down.

5. Data Peeking: The Right Way

Despite the negative consequences of uncontrolled data peeking reviewed in Section 3, the motivation behind data peeking is usually well-intentioned: researchers want to use their limited resources efficiently. Researchers are balancing competing goals, which include maximizing both the number of studies conducted and their experimental power, while minimizing both Type-I and Type-II errors (Miller & Ulrich, 2016). Pressure is particularly acute

in a climate where a lot of value is placed on a researcher's number of publications and on publishing statistically significant findings. Thankfully, the climate is changing, and initiatives such as registered reports make publishing null results from well-designed studies more commonplace (e.g., Simons, Holcombe, & Spellman, 2014). In the meantime, since interpreting and publishing significant results is still the most common strategy, infancy researchers will benefit from carefully planned data peeking which enables flexibility in sample size. Planned approaches to data peeking have long been advocated in other fields, for example in clinical trials (Pocock, 1977), but have only more recently gained attention as a solution for psychology research (Lakens, 2014; Oakes, 2017).

We introduce the term "transparent data peeking" to refer to statistically corrected and transparently reported data-dependent sample size selection. Statistical corrections are necessary because, at their core, problems with data peeking stem from using test statistics that assume random sampling to analyze data derived from non-random sampling. In the next section, we review two solutions for corrected data peeking: one that can be applied in the planning stages of a study (the pre-registered data peeking plan), and one that can be administered after data collection has begun (the post-hoc data peeking plan).

5.1. Pre-registered Data Peeking Plan

Under the pre-registered data peeking plan, the researcher creates a roadmap for data collection at the beginning of a study, and then determines the necessary corrections to their alpha level. The decision steps can be seen on the left side of Fig. 5. The first step is to decide on the maximum sample size that a researcher is willing to test for a study. Then, the researcher decides when during data collection they want to run statistical tests. As reviewed in Section 4.2, the first peek should occur only after a reasonably large sample size is achieved, to avoid "wasting" alpha at a point where stopping would not result in a publishable study. As an example of a data peeking plan, a researcher might plan for two peeks and a final analysis at the maximum sample. Then, the alpha level for each statistical test is adjusted such that the overall alpha level stays below 5% even with the three planned statistical tests. In this case, the researcher would end collection prior to reaching the maximum sample if the *p* value was below the adjusted alpha at the first or second peek (assuming this peek was at a reasonable minimum sample size). The *p*-value calculated at each peek is only used to decide whether data collection can be terminated or not. Once data collection is ended, the researcher computes a final adjusted *p*-value that carries the meaning of a traditional *p*-value, that is, the probability of an effect as or more extreme under the null hypothesis (see Appendix). This adjusted *p*-value takes into account the overall data peeking plan, and is what should be reported with reference to the pre-registered plan (see Lakens, 2014 for more detail).

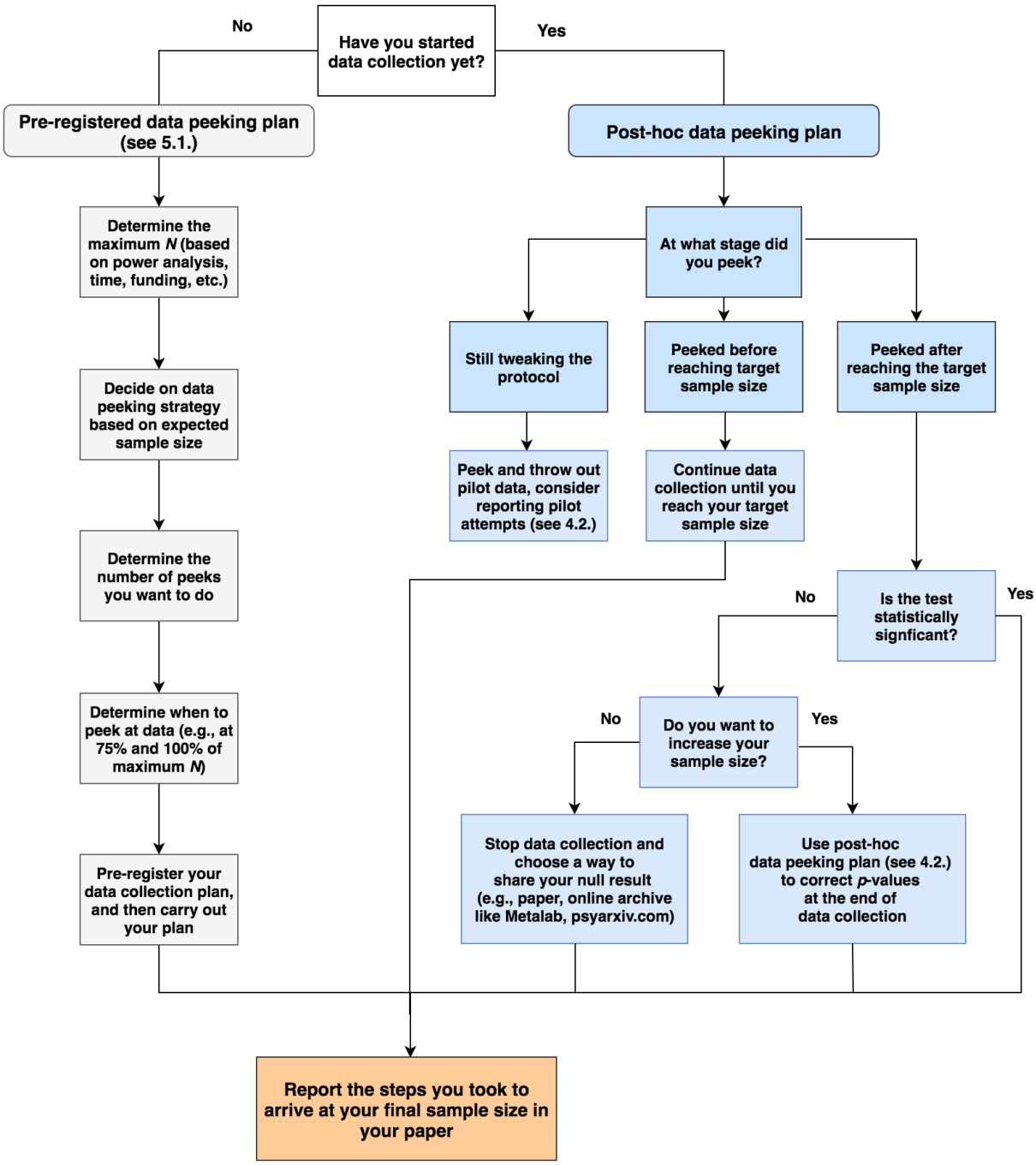

Figure 5. Step-by-step guide for more transparent data peeking. The pre-registered data peeking plan (in grey) is shown on the on the left and the post-hoc data peeking plan (in blue) is shown on the right. Both plans rely on transparent reporting in the final publication.

The approach of the pre-registered data peeking plan is to allocate alpha across the planned points of significance testing, and there are different strategies for allocating alpha when peeking (Lakens, 2014). We will highlight two here. These strategies differ in whether they "spend" more of total alpha during earlier peeks (Strategy 1) or later peeks (Strategy 2). Note that with either strategy, the more peeks that are planned, the lower the allocated alpha will be at each peek, making it harder to find a statistically significant result (i.e., reducing power).

Strategy 1 is best when a large effect size is expected, as it yields more power to detect an effect early in data collection, with the tradeoff of less power later in data collection. Under Strategy 1, alpha is distributed proportionally to the number of participants added between statistical tests (linear spending function; DeMets & Lan, 1994). For example, a researcher who decides on a maximum sample of 32 with peeks at 50% and 75% would need to observe a *p* lower than alpha = .025 at the first peek (16 infants), alpha = .022 at the second peek (24 infants), and alpha = .026 at the maximum sample (32 infants), respectively, for the result to be statistically significant. If the significant result occurs at the first peek (or the second, third, …), they could end data collection then, else they continue to the maximum sample size. Even if the researcher continues until the maximum sample size and evaluates statistical tests at the two peeks and final analysis, the adjustment ensures that the overall alpha level does not exceed .05. Other sampling plans can be seen in Fig. 6 (left panel).

Strategy 2 is best when a small effect size is expected, because in this case early estimates of the effect will be highly variable. Under Strategy 2, a greater proportion of alpha is allocated to later peeks than to earlier peeks (O'Brien-Fleming spending function; DeMets & Lan, 1994). This means that Strategy 2 has less power to detect an effect early in data collection and greater power later in data collection, compared to Strategy 1. Although a researcher is less

likely to end data collection early using Strategy 2 compared to Strategy 1, it is nevertheless possible to do so if the effect studied turns out to be larger than expected.

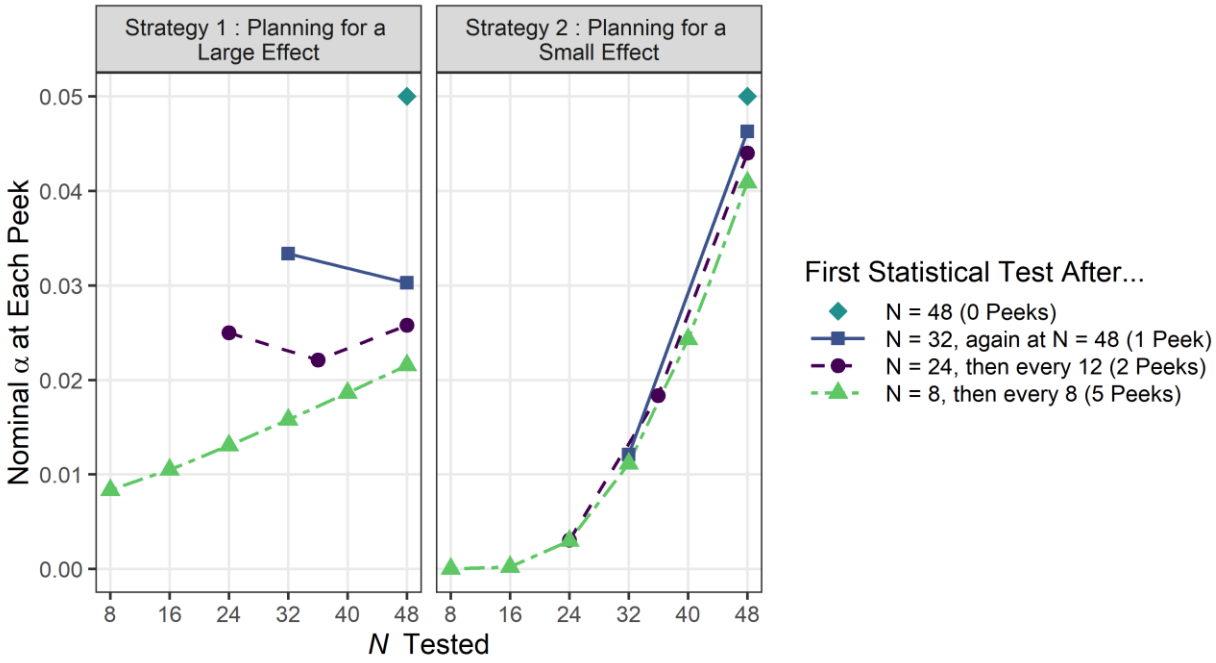

Figure 6. Illustrations of different approaches to correcting the significance threshold under the pre-registered data peeking plan. Using Strategy 1 (left panel), significance thresholds are distributed proportionally to the number of participants tested since the last peek. Using Strategy 2 (right panel), adjusted alpha levels are set very low at early peeks, and increase when the researcher gets closer to the maximum sample size. In both strategies, increasing the number of peeks lowers the alpha levels at each peek as well as at the final analysis, and decreases power.

As an example of Strategy 2, consider a researcher who decides on a maximum sample of 32 with two peeks, at 50% and 75% of the maximum sample. They would need to observe a *p* lower than alpha = .003 at the first peek (16 infants), alpha = .018 at the second peek (24

infants), and alpha = .044 at the maximum sample (32 infants), respectively, for the result to be statistically significant. An example of how this could be written up in a publication is shown in Box 1. Other examples of this strategy can be seen in the right panel of Fig. 6. To customize this approach, the GroupSeq package in R (Pahl, 2017) can be used, which has a graphical user interface (see Supporting Materials https://osf.io/qd25s/ for instructions).

> We pre-registered a sequential analysis (https://osf.io/example-link/) with a maximum sample of 32 infants, and two points of planned sequential analyses after 16 and 24 infants. Significance thresholds for the analysis at each sample size were calculated using the O'Brien-Fleming spending function. After testing 16 infants, no significant difference was found, t (15) = 1.85, p = .084, falling above our adjusted nominal alpha of .003. After collecting data from 8 additional infants and calculating a t-test on the sample of 24 infants, a significant difference was found, t (23) = 2.56, p = .0175. This p-value falls below our pre-registered adjusted nominal alpha of .0183 at the second analysis point, hence data collection was ended. The overall p-value adjusted for the sequential analysis procedure we used is p = 0.019. This value represents the probability of observing an effect that is at least as extreme as the one observed at the second analysis point, given that there was no significant effect at the first analysis point (for more information, see Lakens, 2014).

Box 1. Example write-up of a study using a pre-registered data peeking plan

Pre-registration of the chosen data peeking plan using a platform like the Open Science Framework (https://osf.io/) is an important component of this approach, because lowering alpha affects the study's power to detect a true effect. Imagine Dr. Marginal decides to use a pre-registered data peeking plan where she collects a maximum of 48 participants with peeks after 24

and 36 participants and uses Strategy 1 to adjust her alpha. In a one-sample two-tailed *t*-test, Dr. Marginal has a power of 0.75 (based on simulations, see Figure 7) to observe an effect of a size that would have yielded a power of 0.8 without data peeking. If Dr. Marginal observes $p = .03$ at the maximum sample size $N = 48$, she must retain the null hypothesis because her p is larger than 0.026 (the nominal threshold for her peeking plan). Yet, had she initially decided to run 48 participants without peeking, $p = .03$ would have meant a statistically significant result. This may be frustrating to Dr. Marginal, however it is important to be transparent about her decisions during the data collection process. As we outlined above, the data peeking procedure Dr. Marginal actually used means she cannot use the usual *t*-distribution (which assumes no peeking) to determine significance. To avoid any ambiguity, we recommend pre-registration (see Lindsay et al., 2016, for options), which allows documentation of *a priori* versus post-hoc data collection and analysis decisions (Simmons, Nelson, & Simonsohn, 2011).

5.2. Post-hoc Data Peeking Plan

What can researchers do if they did not pre-register a data peeking plan, but once the target sample size is collected observe an "almost significant" result? One possibility is to run an exact replication with a larger sample size. However, in many cases this approach will be unattractive for infant researchers, given the high costs in money and time for participant testing.

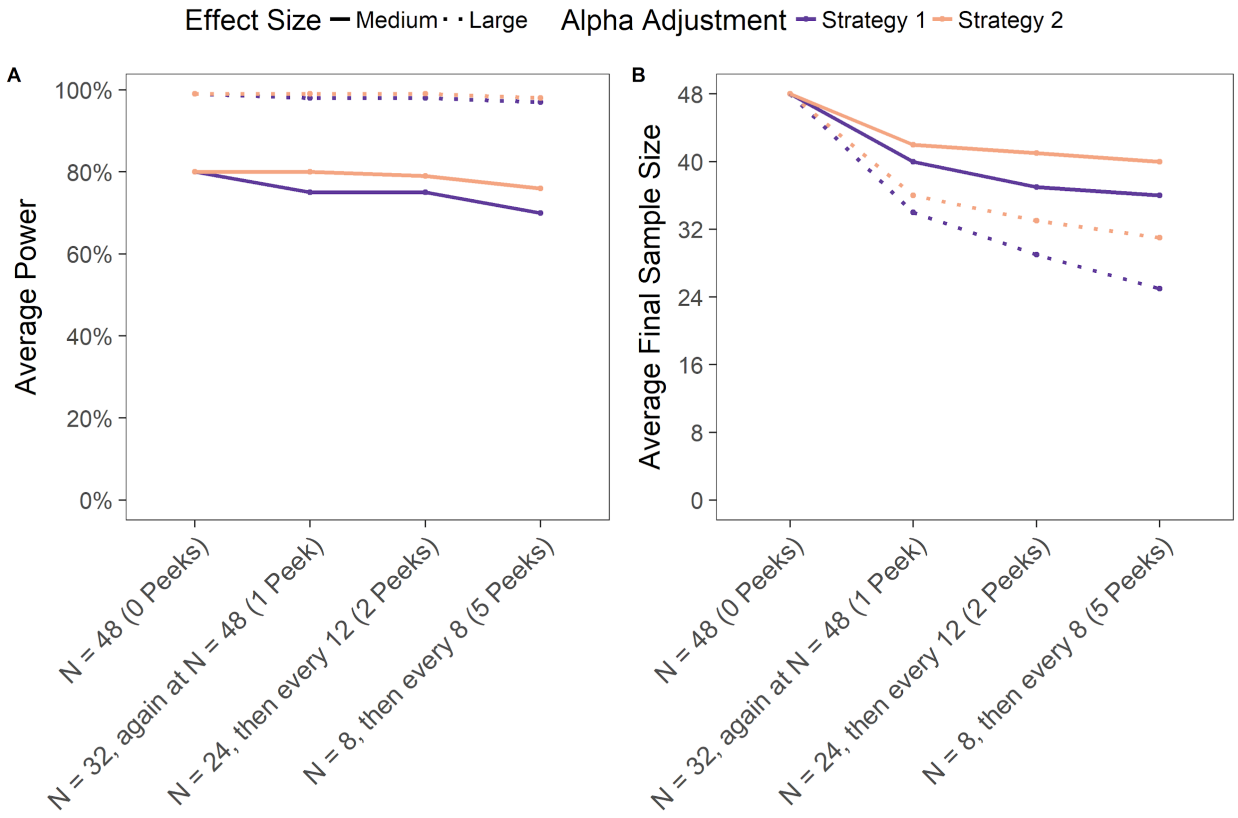

Figure 7. Comparison of the average power and sample size at completion for different sampling plans. A medium effect size indicates the effect size at which a one-sample two-tailed *t*-test with 48 participants has 80% power ($\delta = 0.41$), and a large effect size is a 50% larger effect size ($\delta = 0.62$). When the effect size is large, power for both Strategy 1 and 2 is close to 100%, but Strategy 1 is slightly more advantageous because it results in a smaller average sample size at completion. With a medium effect size, power decreases more in Strategy 1 than 2, and thus Strategy 2 is more advantageous.

Instead, it is possible to collect additional participants and calculate a *p*-value that estimates the impact of the data peeking procedure used (Sagarin, Ambler, & Lee, 2014). The calculation takes into account how many participants were tested before the first statistical test,

how often and by how much the sample was increased, and the observed *p*-values. The formulas to make the necessary calculations using R or Excel are available online (www.paugmented.com).

To illustrate the logic of the post-hoc data peeking plan, imagine Dr. Hindsight conducts a study with 24 infants and observes a final *p*-value of .20. To see if the effect holds up with a larger sample size, she decides to run 8 more infants, yielding an uncorrected *p*-value of .04. To estimate how peeking at the data and deciding to increase sample size affected the *p*-value, the calculation requires the maximal *p*-value, p_{max}, at which Dr. Hindsight would have decided to add additional participants to the dataset. As Dr. Hindsight did not set out with the plan to collect additional data, it is impossible to know at what *p*-value she would have abandoned her study instead of collecting additional participants. To address this issue, Sagarin and colleagues (2014) use a range for p_{max} from the highest observed *p*-value (for Dr. Hindsight $p_{max} = .20$) to the highest possible value for the *p*-statistic ($p_{max} = 1$). Consequentially, the corrected *p* statistic also represents a range, from the corrected *p* calculated for the lower limit of p_{max} to the corrected *p* value calculated for the upper limit of p_{max}. In this example, the corrected *p*-statistic ranges from $p = [0.061 – 0.068]$, and Dr. Hindsight would report this range in addition to the original *p*-value (see Box 2 for an example). The value for *p* corrected for data peeking will always bigger than the nominal alpha level, because once the first statistical test has been evaluated at the nominal alpha level, any additional tests bring *p* beyond that level. Nevertheless, this provides a method to collect additional data points after the start of data collection and report the process transparently.

> In our initial sample of 24 infants, no significant difference was found, t (23) = 1.44, p = 0.164. After testing 8 more infants, the comparison for the whole sample of 32 infants was significant, t (31) = 2.11, p = 0.044. The range for the p-value adjusted for this post-hoc sample size increase is 0.064 – 0.069 (see Sagarin et al., 2014 for more detail).

Box 2. Example write-up of a study using a post-hoc data peeking plan

6. Transparent Data Peeking: Going Too Far or Not Far Enough?

As we have shown, uncorrected data peeking violates the assumptions underlying null hypothesis significance testing and affects the conclusions drawn from the observed data. The two solutions reviewed here allow researchers to get the benefits of more flexible data collection while correcting for some of the negative consequences of data peeking.

The solutions proposed here could be critiqued from two sides: for not going far enough or for going too far. Proponents of a not-far-enough critique may bring up Bayesian statistics as a solution. Bayesian statistics such as Bayes Factors are an alternative to null hypothesis significance testing (Wagenmakers, 2007). Many researchers have argued that Bayesian statistics are less affected by data peeking (Dienes, 2016; Rouder, 2014; but see also Sanborn & Hills, 2014; Yu, Sprenger, Thomas, & Dougherty, 2014). However, switching to a Bayesian framework requires a high upfront investment for learning a new method, and thus is unlikely to be adopted widely in the short- to medium-term. In contrast, the simple-to-use approaches for transparent data peeking outlined here can be applied immediately.

Proponents of the going-too-far critique of data peeking may be concerned that sample sizes are already small in infancy studies (Oakes, 2017), and that encouraging researchers to implement data peeking may lead to even smaller sample sizes. Given the evidence that at least

some infancy researchers have or are currently using data peeking (Eason et al., 2017), we argue that using transparent data peeking is a useful strategy. The pre-registered data peeking plan might also incentivize researchers to plan for a larger maximum sample size, because they have the option to stop early if the effect is large.

7. No More Ambiguity: Embracing Transparency in Data Peeking

Any infancy researcher would acknowledge the complexities of our research topic – infants are hard to recruit, have short attention spans, and it is often difficult to pinpoint the best experimental design for a research question. These complexities are rarely reflected in our manuscripts, where we typically attempt to craft a clean narrative about what infants can and cannot do. It is usually only in informal settings that we discuss the "messy" details about sample size decisions, the months or years it took to collect a sample, or our failed attempts to observe an effect. Not discussing these messy details in a publicly-accessible manner makes it harder for other researchers to reproduce our findings or to avoid our mistakes. Registered reports are one way to be transparent about the research process, but even for studies that were not pre-registered, we argue that there is value in conveying the complexities and challenges of infant research. Increased transparency will benefit the field, but currently it might come at a cost for the individual researcher. Researchers have valid concerns that being honest about the sometimes messy process of science will reflect negatively on the quality of their science, particularly in high-stakes contexts such as peer review. As researchers write their manuscripts in a more transparent way, the responsibility of reviewers and editors is to encourage and embrace this transparency, even if sometimes this openness can lead to a more nuanced – and, frankly, more complex – scientific story. We hope that the solutions proposed here will enable infant

researchers to use data peeking in an effective and valid manner, and to feel comfortable transparently reporting their data peeking practices.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6. doi: 10.1038/s41562-017-0189-z

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development.* Advance Online Publication. doi: 10.1111/cdev.13079

Byers-Heinlein, K., Bergmann, C., Black, A., Carbajal, M. J., Fennell, C. T., Frank, M. C., Gervain, J., Gonzalez-Gomez, N., Hamlin, J. K., Kline, M., Kovács, A. M., Lew-Williams, C., Liu, L., Polka, L., Singh, L., Soderstrom, M., Tsui, A. S. M. (accepted pending data collection). A multi-lab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science.*

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test-retest reliability in infant speech perception tasks. *Infancy*, *21*(5), 648–667. doi: 10.1111/infa.12127

DeMets, D. L., & Lan, K. K. G. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine*, *13*, 1341–1352. doi: 10.1002/sim.4780131308

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. doi: 10.1016/j.jmp.2015.10.003

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. doi: 10.1111/infa.12183

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. doi: 10.1111/infa.12182

Frank, M. C., Vul, E., & Saxe, R. (2012). Measuring the development of social attention using free-viewing. *Infancy*, *17*(4), 355–375. doi: 10.1111/j.1532-7078.2011.00086.x

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. doi: 10.1038/s41562-018-0311-x

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*(7), 701–710. doi: 10.1002/ejsp.2023

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.

Lindsay, D. S., Simons, D. J., & Lilienfeld, S. O. (2016). Research preregistration 101. *APS Observer*, *29*(10). Available at: https://www.psychologicalscience.org/observer/research-preregistration-101

Manybabies Consortium. (accepted pending data collection). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science.*

Miller, J., & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science, 11*(5), 664–691. doi: 10.1177/1745691616649170

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4). doi: 10.1111/infa.12186

Pahl, R. (2017). GroupSeq: A GUI-based program to compute probabilities regarding group sequential designs. Retrieved from https://CRAN.R-project.org/package=GroupSeq

Peterson, D. (2016). The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius: Sociological Research for a Dynamic World*, *2*,1–10. doi: 10.1177/2378023115625071

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191–199. doi: 10.1093/biomet/64.2.191

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. doi: 10.3758/s13423-014-0595-4

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*(3), 293–304. doi: 10.1177/1745691614528214

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*(2), 283–300. doi 10.3758/s13423-013-0518-9

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at perspectives on psychological science. *Perspectives on Psychological Science*, *9*(5), 552–555. doi: 10.1177/1745691614543974

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. doi: 10.1177/0956797611417632

Simmons, J., Nelson, L., & Simonsohn, U. (2018, January 25). Pilot-dropping backfires (so Daryl Bem probably did not do it). [Blog post]. Retrieved from http://datacolada.org/68

Tran, M., Cabral, L., Patel, R., & Cusack, R. (2017). Online recruitment and testing of infants with Mechanical Turk. *Journal of Experimental Child Psychology*, *156*, 168–178. doi: 10.1016/j.jecp.2016.12.003

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi: 10.3758/BF03194105

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review*, *21*(2), 268–282. doi: 10.3758/s13423-013-0495-z