INFANCY

THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES

Infancy, ***22***(4), 470–491, 2017

ISSN: 1525-0008 print / 1532-7078 online
DOI: 10.1111/infa.12183

WILEY Blackwell

A Survey of Common Practices in Infancy Research: Description of Policies, Consistency Across and Within Labs, and Suggestions for Improvements

Arianne E. Eason
Department of Psychology
University of Washington

J. Kiley Hamlin
Department of Psychology
The University of British Columbia

Jessica A. Sommerville
Department of Psychology
University of Washington

Psychology is currently in a period of unprecedented self-reflection, with particular attention being paid to research practices and policies. Although infancy researchers have a history of attending to research practice in the form of papers outlining how to best implement various methods and paradigms, less is known about the particular practices currently being used by researchers in the field, making it difficult to identify areas for improvement. To address this issue, we developed an online survey for measuring methodological practices in infancy laboratories around the world. Our results suggest that infancy researchers adopt a range of a priori research practices. Individuals earlier in their training (particularly within the first year) were more likely to report not knowing their lab's policies, especially regarding issues that arise late in the research process. Members of the same lab agreed on lab policies at above-chance levels; however, improvements could be made. The use of clearly problematic research practices was relatively rare; however, several "risk-permeable" practices were identified that may, under select circumstances, threaten data integrity. We discuss how our results could be used to improve practice in infant research.

Correspondence should be sent to Jessica A. Sommerville, Department of Psychology, University of Washington, Campus Box 351525, Seattle, 98195, WA. E-mail: sommej@uw.edu
Note that all authors contributed equally to the publication; names are listed in alphabetical order.

Psychology, along with many other scientific disciplines, is currently undergoing a period of unprecedented self-reflection (see special issues and sections in *Perspectives on Psychological Science*: Pashler & Wagenmakers, 2012; Ledgerwood, 2014a,b, 2016). A critical part of this self-reflection is an examination of research practices, with the aim of optimizing practice to most effectively pursue scientific truth. The field of infancy has a long history of examining research practices within various methodological paradigms, as evidenced by the large number of "tools of the trade" articles in the literature. These articles outline particular paradigms, explain their unique challenges, and propose effective solutions (e.g., looking time: Aslin, 2007; Aslin & Fiser, 2005; Horowitz, 1974; Oakes, 2010; Spelke, 1985; eye-tracking and head-mounted cameras: Aslin, 2012; Oakes, 2012; Smith, Yu, Yoshida, & Fausey, 2015; conditioned responses: Werker, Polka, & Pegg, 1997; neurological measurement: Csibra, Kushnerenko, & Grossmann, 2008; DeBoer, Scott, & Nelson, 2007; Gervain et al., 2011; just to name a few). Tools of the trade articles are invaluable for helping infancy researchers make research decisions as they pertain to specific paradigms and allow researchers to effectively implement various research methodologies.

In addition to developing an understanding of which practices infancy researchers *should* use, optimizing practice in infancy research presumably also requires some knowledge of the practices infancy researchers actually *do* use. One reason for this is straightforward: Pursuing best practices requires knowing both which areas are in need of improvement and which are already functioning well. Indeed, as any developmental researcher will tell you, an understanding of how best to reach some ideal end-state requires an accurate picture of the nature of the initial state. Furthermore, there may be legitimate disagreements among expert infancy researchers as to what practices are indeed "best" for a given methodology, and so cataloging variation among researchers might help to elucidate to what extent this is the case. Finally, although many infancy methods have a long and well-established history, other methods are newer and may still be evolving. In these cases, it is particularly important to understand the practices infancy researchers adopt and the ways in which they approach methodological decision-making.

Unfortunately, to our knowledge no comprehensive account of practices in infancy research currently exists. This study was designed to address this gap in the literature, with the goal of helping infancy researchers understand the current state of our field so that we might effectively pursue best practices. We reasoned that the most straightforward way to determine current practice in infancy would be to simply ask infancy researchers to report what they do in their own laboratories. To this end, we designed and conducted an online survey aimed at measuring methodological policies and practices in infancy laboratories around the world.

As our overarching goal was to identify research practices and policies common to a fairly wide range of infant-specific paradigms, we identified six aspects that we saw as central to the infancy research process. First, infancy research typically involves a large degree of pilot work to determine how to effectively pursue a research question with infant populations. Therefore, we asked questions central to the piloting process such as the degree to which researchers pilot and why they pilot. Second, because infancy research can be expensive and resource intensive, it is common for considerable attention to be paid to sample size; accordingly, we queried respondents about their practices for determining sample size, such as whether

sample sizes are predefined and the basis on which sample sizes decisions are made. Third, because infancy research often involves running a range of experimental and control conditions to accurately interpret data, we queried researchers regarding how they assign participants to conditions and regarding whether and how they engage in blinding practices. Fourth, given that attrition in infancy research can be higher than in other fields due to fussiness and noncompliance, and because infant-specific paradigms can be challenging to enact, we asked respondents questions about inclusion and exclusion criteria, such as the circumstances under which a participants' data would be excluded from the final data set. Fifth, because statistical analyses are germane to all aspects of psychological research, we queried respondents about their statistical approaches, such as the extent to which decisions about statistical approaches are made a priori and whether all dependent measures are reported. Sixth and finally, because infancy labs typically rely on a large range of students, staff, and volunteers, training procedures are often critical to success; thus, we asked respondents about their training practices and how lab practices are disseminated. Given that our goal was to assess both basic methodological variation (i.e., different methodological choices researchers might adopt, of which none are clearly problematic) as well as the presence of more problematic research practices, we generated a range of possible responses for each question, some of which we considered more troublesome than others.

In addition to illustrating the range of practices currently being used by infancy researchers, the analyses reported below explore several questions. First, best practices minimally require that each individual/lab adopt policies that, whenever possible, are consistently applied across studies and situations. Indeed, the dangers of post hoc decision-making in science have recently received increased attention (Simmons, Nelson, & Simonsohn, 2011), and some have even suggested that post hoc decision-making is particularly likely to occur in infancy research (e.g., Peterson, 2016). Thus, here we explore whether infancy researchers report having policies or not, as well as whether this differs depending on which aspect of the research process is being asked about. Relatedly, given that consistently implementing policies requires knowing what they are, we explore to what extent researchers report knowing vs. not knowing their lab's policies and to what extent reported policy knowledge varies based on experience (e.g., years in a lab). Second, we explored whether researchers within the same lab report adopting the *same* policies and practices. Given the broad ages and theoretical questions that general infant methodologies lend themselves to, we expected some policy variation would be reported, even among individuals within a lab who use the same general method. Despite this, we reasoned that measuring within-lab, within-method variability would help to estimate both the existence of policies within a lab and how effectively those policies are being disseminated and followed. Third, and finally, we provide details on the rate that researchers report engaging in various specific practices. Rather than simply listing the rate at which various practices are followed (which we provide in supplementary tables), we chose to focus our discussion specifically on whether and to what extent participants report adopting practices that may pose risk to the integrity of published research. We divide these risky practices into two types: those that (in our opinion) are "clearly problematic" and those that are merely "risk-permeable." Clearly problematic practices are practices that we view as posing inevitable threat to the integrity of the published research and should therefore always be avoided.

Risk-permeable practices, in contrast, are practices for which we can identify both situations wherein the practice would be problematic and situations wherein the practice would be fine. We hope that our analysis will help to illustrate that whether or not a particular practice is clearly problematic, risk-permeable, or risk-free is often a matter of both opinion and context, and encourage researchers to carefully consider their own research context in deciding whether or not a particular practice is or is not problematic.

METHODS

Survey participants

Infancy researchers were recruited to participate in the survey via several developmental psychology email listservs. The recruitment email specified that we hoped that as many individual members of a lab as possible would participate, including lab managers, graduate students, post-docs, and faculty members. We chose not to include undergraduate research assistants, as we reasoned that variability in role, time in lab, and immersion level would mean that at least some undergraduates would not be sufficiently aware of lab policies to accurately report on them. We also specified that all survey responses would remain anonymous and requested that individuals within each lab not discuss the survey until all interested lab members had submitted their responses. So that responses from members of the same lab could subsequently be identified, the first member of each lab to begin the survey chose a unique alphanumeric code that he/she subsequently provided to other lab members; we specified that original codes would be de-identified to maintain anonymity. Although we did not incentivize participants to tell the truth, previous research suggests that psychologists are willing to report engaging in questionable research practices even without such incentives (although incentives may increase reporting rates to some extent; e.g., John, Loewenstein, & Prelec, 2012).

Our final sample consisted of 151 fully completed surveys, coming from at least 72 distinct laboratories (five respondents did not provide a lab code, making it impossible to know whether they were from a previously represented lab or not). There were slightly more nonfaculty respondents (i.e., postdoctoral researchers, graduate students, research coordinators; 54%) than faculty respondents (44%). In particular, 7% of respondents were lab managers/research coordinators, 31% were graduate students, 13% were postdoctoral researchers, 14% were assistant professors, 13% were associate professors, and 18% were full professors. The remaining 5% indicated a different academic rank (e.g., "research scientist") or did not answer.

Survey design and response options

The full survey is available at https://nyu.databrary.org/volume/239/slot/11764/-/asset/47180; an overview of major survey categories with a summary of the question contents appears in Table 1. The survey was drafted by the authors and subsequently revised based on invaluable feedback from several others in the field (J. Colombo, L. Oakes, M. Soderstrom). This process was completed in a relatively short period, from late February through late March 2016, so that it might be administered and results analyzed prior to the "Building Best Practices in Infant Cognition Research"

TABLE 1
Overview of Major Survey Categories

Survey category	*Description*	*Number of questions*
Piloting policies	Questions about: whether, why, and how researchers conduct pilot studies?	4 (2 MC; 1 CA; 1 OE)
Sample size	Questions about: How final sample size is determined? what happens when extra subjects tested? what happens when $p > .05$ but $<.10$?	5 (all MC)
Condition and blinding	Questions about: How are infants assigned to conditions? What are experimenters and parents blind to?	10 (7 MC; 3 CA)
Inclusion/Exclusion	Questions about: What constitutes a procedural error and what are the consequences? who decides on inclusion/exclusion of participants? are procedures verified?	17 (14 MC; 2 CA; 1 OE)
Statistical analyses	Questions about data analyses prior to first submission, including data integrity and reporting and choice of statistical tests	8 (all MC)
Training procedures	Questions about documenting, updating, and disseminating lab policies about lab policies	9 (4 MC; 3 CA; 2 OE)

Key: CA, check all that apply; MC, multiple choice; OE, open ended.

preconference at the May 2016 meeting of the International Congress for Infant Studies. In total, the survey included 66 multiple-choice (some of which, $n = 9$, allowed participants to chose all options that applied; see Table 1) and four open-ended questions, in which participants first answered various demographic questions about themselves, their lab, and the methods they use. Given that many labs/researchers use multiple methods for which they might have different policies, after completing demographic questions participants were asked to indicate which method they would base their subsequent answers on. Participants were invited to fill out the survey multiple times for different methods if they wished (19 did so). Results reported below include only participants' primary methodologies (all data are provided in the data files; see below for how to access all data).

In addition to question-specific response options, the majority of questions included options to answer either "I don't know my lab's policy," "my lab does not have a policy," or "N/A," so that participants could indicate that the question did not apply to the particular method they was answering about. Finally, to ensure that responses were not unduly influenced by the particular response options we generated, the majority of questions included an "Other – please describe" option, where participants could write in their own policy and/or provide a comment.

When examining written responses to the "Other" category, we noted that participants chose this option for several different reasons. In some instances, the participant had written a description that clearly fell into a response option that was already represented (but perhaps they also included a justification for the response). In these cases, we recoded the response to the appropriate option. In other instances, multiple people provided the same policy that appeared perfectly plausible, but was not one we had initially generated. In these cases, we created a new category to represent the option (most commonly, this happened because we had initially failed to provide an N/A option). Responses that stayed in the "Other" category included (1) cases in

which only 1–2 people provided a policy we had not initially generated, (2) cases in which participants stated they did not understand the question, and (3) cases in which we felt the response did not provide evidence of a policy. The data file with the recoded values can be found at: https://osf.io/tdhmf/?view_only=ccf997c0b35d45aaab8bc68f3c852695; and the (pre-recoded) raw data file can be found at: https://nyu.databrary.org/volume/239/slot/11764/-/asset/47181.

RESULTS

Laboratory descriptives and methodologies

Below we describe key laboratory characteristics and methodologies. Other characteristics of individual respondents, and labs they belong to, are described in the Supplement (Table S1).

Lab locations

Participants were predominantly from North America (64% located within the United States and 9% located within Canada). The remaining laboratory locations included the European Union (21%), Australia and the Pacific (4%), Asia (1%), and the Middle East (1%).

Lab institutions

Participants indicated that they worked at public institutions that were focused on both research and teaching (39%), public institutions focused on primarily research (24%), private institutions focused on both research and teaching (18%), and private institutions focused on primarily research (15%). A small minority worked at public or private institutions focused on teaching (4%).

Lab size

More than half of the labs reported that their lab consisted of more than four graduate students and postdocs (58%). Many (44%) had between 6 and 10 undergraduates working in the lab at one time, with the next most common response (29%) being 0–5 undergraduates.

Publication rate

Perhaps indicative of the relatively slow speed at which infant research proceeds, the majority of participants reported that their lab publishes 0–1 (22%), 2–3 (27%), or 4–5 papers per year (27%).

Age groups tested

Participants were asked to indicate all of the age groups that their laboratory regularly studies. Give that our main focus was on infant research practices, it is unsurprising that 92% of participants report running studies with children between 0 and 1 year

old and that 78% report running studies with children between 1 and 2 years old. In addition, 60% of labs also run studies with 2–3-year-olds, 54% with 3–5-year-olds, and 9% with children older than 5 years of age.

Methodologies

Participants generally reported that their lab uses multiple methods (see Table S2 for a breakdown of the rate of choosing particular methods). The most common methodologies were preferential responding paradigms (e.g., preferential looking or preferential reaching; 82% of respondents indicated using these), computerized habituation and familiarization paradigms (74%), and eye tracking (64%). These same three methods were most often identified as individuals' primary methodology (29% computerized habituation/familiarization, 23% eye tracking, 21% preferential responding).

Do infancy labs have policies?

The broadest question our survey allowed us to ask was to what extent participants reported having policies about various aspects of the research process vs. not having policies (e.g., making research decisions on an ad hoc, case-by-case basis). Indeed, best practices minimally require that each individual/lab adopt policies that, whenever possible, are consistently applied across studies and situations. Recently, however, it has been suggested that given the unique challenges of testing infants, infancy research may be particularly susceptible to post hoc decision-making (Peterson, 2016). Our survey allowed us to address this question by examining the reports of infancy researchers themselves.

We first identified which survey questions asked about policies and excluded those questions that did not (e.g., questions about lab demographics and primary methodologies). Subsequently, for each relevant question we excluded responses indicating that the participant did not believe the question to be applicable to his or her chosen method (for instance, eye tracking does not require reliability coding; not all researchers use "trials" in their methods). All other responses to each question were categorized as either reflecting a policy (e.g., the participant identified one of our policy options or chose "other" and wrote in their own specific policy) or not reflecting a policy. Responses that we considered not reflecting a policy included "my lab does not have a specific policy on this issue," "I do not know my lab's policy," "prefer not to answer," and "other," in which the participant provided a response that we felt did not reflect a policy, for the reasons outlined above. Note that for the purposes of this analysis, we made no distinctions based on the *quality* of the policies participants identified: We were solely interested in whether researchers identified a policy or not. We will return to the question of policy quality in a later section.

Results from these analyses are outlined in Table 2. They demonstrate that infancy researchers report having policies the vast majority of the time; specifically, participants identified policies that they/their lab follow over 80% of the time in each of the five question categories (average = 82%). Moreover, rates of having policies were quite consistent across different aspects of the research process (range = 81–85%). When no policy was identified, participants reported that their lab has no policy an average of

TABLE 2
Average Percentage of Responses Chosen for Each Section, Broken Down by Whether or Not a Policy was Identified and (if not) for What Reason. Columns do not Consistently add to 100% due to Rounding Error

	Identified a policy we offered, or chose "other" and provided alternative	*Did not identify a policy*			
		Chose "my lab has no policy"	*Chose "I do not know my lab's policy"*	*Chose "other" and response was not interpretable*	*Chose "prefer not to answer"*
Piloting policies	85%	10%	4%	0%	2%
Sample size	82%	7%	8%	0%	2%
Condition assignment & blinding	81%	9%	7%	0%	2%
Inclusion & exclusion	83%	7%	8%	1%	2%
Statistics	81%	5%	10%	1%	3%
Overall	82%	8%	7%	0%	2%

8% of the time and that they did not know their lab's policy an average of 7% of the time. "Other" and "prefer not to answer" responses were very rare. Thus, these findings suggest that researchers are primarily engaging in a priori, rather than post hoc, decision-making about the research process.

Do infancy researchers know what their lab policies are?

In order for infancy researchers to effectively implement policies, they must first and foremost know what their lab policies are. As highlighted above, the rate at which participants indicated that they did not know their lab's policy was low: Across all questions, the average IDK score was 3.03 (SE = .45) of 44 questions, or 7%. This figure is encouraging, as it suggests that the vast majority of participants operate according to known lab policies.

Although the rate of IDK responses was low overall, it is possible that different lab members differ in their knowledge of lab policies. For instance, given that faculty are primarily responsible for setting lab policies, it would be surprising if faculty regularly reported not knowing their own lab's policies. Thus, we explored whether faculty and nonfaculty would report different rates of IDK responses, anticipating that nonfaculty would report higher rates of IDK responses than would faculty. Confirming our hypothesis, comparison of faculty (n = 67) to nonfaculty (n = 82) respondents revealed that nonfaculty had a significantly higher rate of IDK responses (M = 5.27 of 44; SE = .70), than did faculty (M = .06 of 44, SE = .03), $t(81.28) = 7.43$, $p < .0001$. On average, nonfaculty chose the IDK response 12% of the time, whereas faculty did so only .001% of the time.

Nonfaculty IDK rates may have been higher than faculty IDK responses for several reasons. One possibility is that faculty are generally ineffective at communicating policies to their lab members. On the other hand, perhaps the increased rate of IDK response by nonfaculty was most strongly driven by students, staff, and postdocs that

are relatively new to their labs. Indeed, policies are often learned and reinforced not merely via communication about particular policies, but through actually putting those policies into practice. From this perspective, lab members with less time in a lab may be less knowledgeable about particular lab policies because they have not yet had a chance to encounter certain aspects of the research process.

To investigate this possibility, we examined IDK responses for nonfaculty as a function of how long they had been in the lab, focusing on nonfaculty who had been in their position for <1 year up to 5 years. Inspection of the distribution of IDK responses revealed that the distribution significantly differed from normal, Shapiro–Wilk normality test, $W(82) = .78$, $p < .0001$ and was positively skewed (skewness = 1.53). Thus, we conducted an ANOVA on average log-transformed IDK responses with years in lab (1 year or less, 2–3 years, 4–5 years) as the between-subject variable. This analysis revealed a significant effect of years in lab, $F(2, 58) = 5.85$, $p = .005$. Whereas the average IDK rate was 18% for those in the lab for 1 year or less, it was 10% for those in the lab for 2–3 years and 6% for those in years 4–5. Planned comparisons (with Bonferroni correction) revealed a significant difference in IDK response between nonfaculty of 1 year or less and those in years 4–5 ($p = .004$), but no significant differences between all other comparisons ($p > .19$).

While the log transformation improved the normality of the distribution, even with this transformation the distribution still significantly differed from normal (Shapiro–Wilk normality test, $W(65) = .93$, $p = .002$). Thus, we conducted a Kruskal–Wallis test (which does not assume a normal distribution) on the raw data. This test also revealed an effect of years in lab, X^2 ($df = 2$) = 6.45, $p = .040$. Table 3 provides a breakdown of the average percentage of IDK responses as a function years in lab for each major survey category. At a descriptive level, it can be noted that most question categories follow the same general trend, whereby more years in a lab are associated with fewer IDK responses.

Thus, these findings suggest that IDK responses are not evenly distributed across all nonfaculty, but rather become less prominent the longer nonfaculty have been in a given lab. Our results suggest that students, staff, and postdocs learn lab policies as they need to apply them to various different steps of the research process. In further support of this claim, inspection of the descriptive stats in Table 3 reveals that IDK rates for nonfaculty are lowest for policies that need to be implemented earlier in the research process (e.g., piloting, sample sizes) and higher for those policies that may not need to be implemented until later in the research process (e.g., statistical analyses).

TABLE 3
Percentage "I Don't Know" Responses as a Function of Nonfaculty Time in Lab

Survey section	*1 year or less*	*2–3 years*	*4–5 years*
Piloting policies	11%	1%	1%
Sample size	16%	10%	4%
Condition assignment and blinding	12%	10%	5%
Inclusion/Exclusion criteria	18%	11%	6%
Statistical analyses	32%	16.0%	9%
Training procedures	6%	4%	7%
Overall	18%	10%	6%

Taken together, these findings suggest that while there is some room for improvement in dissemination of lab policies (ideally, all new members of a lab would encounter and learn lab policies immediately), some of the gaps in knowledge revealed in our survey can be accounted for by the fact that nonfaculty respondents are in the very process of learning and applying lab policies.

Do members of the same lab report using the same policies?

The previous sections demonstrated that infancy researchers report having policies and that knowledge of policies increases with more time spent in the lab. But do individual members of the same infant lab agree on which policies their lab uses? Of course, it is not only desirable for lab policies to be known, but also for lab members to have a shared understanding of what the particular lab policies are; if so, individual lab members should tend to choose the same response option for any given question. Presumably, high levels of agreement among people within a lab signal that the policies in place are being translated into practices, whereas low agreement signals that practices may differ.

To assess within-lab reliability, we calculated Krippendorff's alphas (α), a highly robust reliability measure that satisfies the unique constraints of our survey. In particular, unlike a percent agreement measure or Cohen's κ, Krippendorff's α can be used to assess reliability between more than two observers (required in this case because several labs had more than two lab members to fill out the survey). Furthermore, Krippendorff's α can be used on nominal data in which questions have different numbers of response options, as many of our questions did. Finally, Krippendorff's α is unbiased by the number of respondents and can handle missing data (Hayes & Krippendorff, 2007). Krippendorff's α ranges between -1 (systematic disagreement) and 1 (perfect agreement), with a value of 0 reflecting chance-level concordance among individuals.

Analyses focused on the 21 (of 72) labs that had two or more participants who reported using the same primary methodology (as one would expect users of different methodologies to have different policies). On average, there were 2.86 respondents per method per lab (SD = 1.15, Range = 2–6 respondents). When calculating the reliabilities among these individuals, we removed responses of "I don't know my labs' policy" and "Prefer not to respond" as we reasoned that these responses would negatively bias the estimates. (Note that concordance statistics do not change when "I don't know" responses are included).

Across the entire survey, the average reliability (M_α = .512, SD = .07, Range = .420 to .706, 95% CI [.48, .54]) was significantly greater than chance, $t(20) = 32.78$, $p < .001$, $d = 7.15$. This suggests that individuals in lab were systematic in their responses, choosing the same answers in response to the same policy questions. Despite above-chance level consistency, there was still room for improvement, as the average reliability also significantly differed from perfect agreement, $t(20) = 31.20$, $p < .001$, $d = 6.81$.

To investigate whether particular parts of the research process led to more (or less) systematic responding, we also investigated the reliabilities for each section of the survey (see Table 4 for means). We conducted a repeated-measures ANOVA on reliability within each section of the survey (piloting, sample size, condition assignment and blinding, inclusion and exclusion criteria, statistical analyses, and undergraduate

TABLE 4
Krippendorff's Alpha Values for the Respondents Within the Same Lab Who Reported the Same Primary Methodology

Survey section	*Average α (SD)*	*Minimum α*	*Maximum α*
Piloting policies	.459 (.22)	.096	1.00
Sample size	.310 (.24)	−.154	.811
Condition assignment and blinding	.454 (.17)	.167	.775
Inclusion/Exclusion criteria	.504 (.10)	.253	.711
Statistical analyses	.355 (.21)	−.083	.793
Training procedures	.499 (.28)	−.136	1.00
Overall	.512 (.07)	.420	.706

training/dissemination). One lab had missing data in one section of the survey; therefore, they were not included in the following analyses. Overall, there was a significant effect of survey section, $F(5, 95) = 2.695$, $p = .025$, $\eta_p^2 = .124$, suggesting variability in agreement by major survey category. However, follow-up tests (Bonferroni corrected) revealed that only the category of highest agreement (inclusion/exclusion) and lowest agreement (sample size) was significantly different from one another ($p = .015$). See Table 4 for the average alpha for each section of the survey.

Overall, these data suggest that even though individuals are reporting that their labs *have* policies on various issues, agreement regarding the particular policies within a lab shows above-chance systematicity, but also room for improvement.

Description and evaluation of lab policies

In the section that follows, we describe the rate at which participants chose each response option for each question in the survey, broken down by survey section exploring different parts of the research process. The majority of these data are provided in supplementary tables (one table per survey section). As is obvious by the respondent breakdown for each question, particular policies implemented varied across labs. We anticipated this variability, given that there are often multiple ways to solve particular research problems, and because what is considered the best policy or practice for a particular research problem is often contested within any science.[1] Thus, one of our goals here is simply to describe the variability currently present in infancy researchers' policies and practices.

In addition to describing variability in infancy research practices, because the ultimate goal of exploring practice was to maximize the integrity of the published research, in this section we mainly focus on to what extent researchers reported engaging in policies that may threaten data integrity. As described in the introduction, for each survey section we divide these risky practices into clearly problematic and risk-permeable, wherein risk-permeable practices are those for which the extent that they threaten research integrity varies depending on the context under which they are implemented. Notably, given that our analysis of what is clearly problematic vs. risk-permeable vs. risk-free was based on (in addition to various methodological papers cited below) little more than our own opinions, it is certainly not our intention that what

[1]Furthermore, some variability may have resulted from the idiosyncratic word/item choices we made in designing the survey.

follows be taken as any sort of authority on which practices are ok vs. not ok for infancy researchers to implement. Rather, we hope that our examples and discussion will serve to illustrate various examples of how certain practices may be problematic in some circumstances and not others, thereby encouraging readers to actively consider whether any of their own practices are risk-permeable. Notably, the use of clearly problematic practices was rarely reported, with an average response rate of 4%. This evidence suggests that the prevalence of clearly problematic practice is low in infancy research.

Piloting policies

See Table S3 for the questions participants were asked about piloting practices, and the percentage of participants that selected each response option.

Cleary problematic practices

We identified the response option of "piloting in order to determine whether the data conform to your hypothesis" as a clearly problematic practice. A central consideration of our identification of this practice as clearly problematic was the fact that participants had the opportunity to select a different response option, in which piloting is used to determine whether preliminary data are "interpretable," irrespective of whether or not the data conform to the participants' hypothesis. Given that these participants did not choose this option, we reasoned that participants who selected the "conform to your hypothesis" option might selectively fail to pursue preliminary evidence that is interpretable, but runs counter to their hypothesis. This practice clearly introduces bias to the published literature.

Risk-permeable practices

Fifty-six percent of respondents to our survey reported that they distinguish between pilot and test data after they have stopped making procedural changes and count all participants from the last procedural change as test data. Although this policy/practice is not problematic if the decision to transition to test participants is prospective (as in: "We are confident that we have now made the last change to the procedure we will make. Starting tomorrow, all participants will count as test data."), it introduces bias when the decision about which participants count as test data are retrospective (as in: "The procedure has been running smoothly since we made the last procedural change. Thus, we will count all participants since that change as test data."). The second, retrospective decision-making option means that the first few subjects included in final data sets will almost certainly have gone "well" (given that these subjects formed the basis of the decision to stop changing the procedure), introducing positive bias into the results. Given the context-based nature of whether or not this practice is problematic, we identified it as risk-permeable.

Sample size

See Table S4 for the questions participants were asked about sample size, and the percentage of participants that selected each response option.

Cleary problematic practices

Three percent of participants reported not setting their sample size in advance of starting data collection, but instead running "enough subjects to get a good idea of what the effect looks like" and then stopping. Relatedly, 2% of participants report that when results look promising, but p is not less than .05, they add subjects until p is <.05. Others report that when results are promising but p is not less than .05, they may add additional groups of infants multiple times (1%). Each of these responses is clearly problematic in that they all reflect the use of "optional stopping": Running fewer or more infants than was initially planned based on how the data looks. Optional stopping has been shown to create spurious results, given the likelihood that at some point during data collection, p will dip below .05 simply by chance (Simmons et al., 2011). Furthermore, whereas it has been shown that adding additional participants to a sample just once does relatively little to change one's Type 1 error rate (assuming the initially observed p-value is in fact promising), doing so multiple times can greatly inflate rates of Type 1 error (Sagarin, Ambler & Lee, 2014).

Despite the fact that the large majority of participants reported setting sample sizes ahead of time (although with some problematic adding of subjects when initial results are promising), 1% of participants who preset sample size in advance run the minimum number of infants possible that, in their experience, yields a significant effect. We identified this practice as clearly problematic because not only do small sample sizes often fail to detect true effects, they also overestimate effect sizes when significant effects are detected (e.g., Berger & Selke, 1987; Button et al., 2013; Cohen, 1992; Ioannidis, 2005).

Risk-permeable practices

Eleven percent of participants report that in cases where p-values are promising but not less than .05, they add infants until they are "confident that that there is or is not an effect." We included this practice as risk-permeable, given that it is permissible in certain cases; for instance, if researchers are using Bayesian statistical methods (see Csibra, Hernik, Mascaro, Tatone, & Mengyel, 2016; for discussion of Bayesian analyses of infancy data) or sequential analyses (e.g., Lakens, 2014; see also Sagarin et al., 2014). However, to the extent that participants add infants in this manner while using traditional forms of null hypothesis statistical testing, this practice would be clearly problematic.

Condition assignment/blinding

See Table S5 for the questions participants were asked about condition assignment and blinding, and the percentage of participants that selected each response option.

Cleary problematic practices

Participants reported that presenters (7%), online coders (3%), and offline coders (3%) are not blind to any relevant study factors and that this could influence results. Although there are certainly situations in which individual experimenters simply cannot be blinded (e.g., it can be difficult for a live presenter to be blind to their own actions; we assume this is why more participants report using unblind presenters vs.

unblind coders), experimenter blinding is critical to producing unbiased results. Thus, in situations in which experimenter blinding is not possible, further checks should be introduced to ensure that the study results were not inadvertently influenced by experimenter bias. Without such checks, these practices are clearly problematic.

When asked what is one's policy when an experimenter inadvertently becomes unblinded, 7% of participants reported that violations to experimenter blinding are not identified in published manuscripts. Similarly, 11% of participants reported that violations to online coder blinding are not identified. These answers suggest that readers are sometimes not informed of situations in which participants were run in potentially biased circumstances; we view this as clearly problematic.

When asked about parental interference policies, 8% of participants reported that parents would need to interfere with their infant significantly and for much of the study in order for the infant's data to be excluded from final samples. Note that an alternative option allowed participants to indicate that they include infants in final samples if parents interact with infants during the course of the study in ways that *do not* appear to interfere with infants' performance. That is, choosing this option indicates that the participant includes infants in final samples who almost certainly did not respond independently of their parents, which we view as clearly problematic.

Risk-permeable practices

Nineteen percent of participants reported that infants sit on parents' laps during testing, but that parents are not blinded in any way. This raises the possibility that parental perception and interpretation of experimental events may influence infants' responding. Of those participants who report that it is their policy to blind parents to the experimental situation, 21% report that if parents fail to follow blinding instructions, they would exclude the infant only if the parent continued to show continued noncompliance to blinding instructions. Note that as above, by choosing this option participants did not choose an option in which infants are retained in final samples if noncompliance to blinding instructions is minor (e.g., parents are almost certainly still blind). This once again indicates that are infants being retained in samples in cases in which parents had the opportunity to influence results.

Here, we note that these practices could be viewed as clearly problematic rather than risk-permeable: Parents have the potential to influence results, which clearly introduces risk for bias. That said, presumably the extent to which parents can influence study results likely varies quite a lot depending on study details; in particular, how obvious the "correct" answer for a given study is. Therefore, we identified the practices listed here as risk-permeable.

Inclusion/Exclusion

See Table S6 for the questions participants were asked about inclusion/exclusion, and the percentage of participants that selected each response option.

Cleary problematic practices

One percent of participants reported that they include data in a final sample even if there is a major deviation during the course of the procedure. Notably, these participants did not choose an alternative option, whereby researchers retain infants in the final sample so long as any procedural deviation is viewed as unlikely to influence infants' interpretation of the stimuli. Thus, participants choosing this option include infants in final samples even when procedural deviations likely led to changes in infants' interpretation or performance.

When identifying who determines whether an infant's data are included in or excluded from the final sample due to a procedure error, 20% of participants indicated that decision makers are often aware of how the infant responded during the experiment when making their decision. Note that these participants did not choose an alternative option, wherein individuals making decisions about the inclusion of data are blind whenever it is practical, suggesting that labs do not always ensure that (when possible) blind individuals make inclusion/exclusion decisions. Obviously, this introduces the possibility for selecting data on the basis of whether or not infants' data conform to the hypothesis, which would lead to biased results.

Two percent of participants reported including babies who, for whatever reason, received fewer habituation trials than required. Given that habituation paradigms specify that infants must either meet a preset habituation criteria or view a minimum number of trials, this response indicates that data are sometimes included in final samples that effectively represent a procedural error.

In cases of equipment failure leading to a lack of a video record for infants' performance, 1% of participants indicated that, so long as online coding of infants' performance was available, infants' data are retained irrespective of how the procedure went. Note that an alternative option was retaining the infants' data so long as everyone involved agrees that everything went smoothly; this suggests that in fact data are retained even when one or more individuals believe there were procedural problems. This represents a problematic practice because there is reason to suspect there were procedural issues but no way to independently verify this.

Risk-permeable practices

We asked several questions about who can make the decision to exclude an infant's data based on various issues, such as "fussiness" or "inattentiveness." In response, 13% of participants reported that any lab member can exclude an infants' data due to fussiness, and 67% reported that only the primary experimenter can exclude an infants' data due to fussiness. Similarly, 10% of participants reported any lab member can exclude an infants' data due to inattentiveness, and 51% reported that only the primary experimenter can exclude an infants' data due to inattentiveness. Note that in choosing each of these options, participants did not choose the options indicating that decision makers must be blind, leading us to believe that in fact these decision makers are often not blind. We identified this as risk-permeable rather than clearly problematic given that many labs reported having a priori explicit criteria for what constitutes fussiness or inattentiveness, which presumably makes blindness less of an issue. Nevertheless, this practice may introduce more risk than necessary, given that unblind experimenters may have difficulty making objective decisions regarding the a priori criteria.

When asked about how procedures are verified, 33% of participants reported that procedural deviations are only identified online or perhaps during reliability coding. Given that some procedural errors will not be caught in the moment, and given that not all labs reliability code 100% of their data, a better choice might be to have independent procedural checks performed by a blind coder.

For those infants who were retained in samples despite procedural deviations, 21% of participants reported not noting procedural deviations in the methods sections of their manuscripts. Although many of these deviations may be quite minor, presumably reporting these would provide readers with a more accurate sense of any variability in procedures, as well as how particular researchers tend to make inclusion and exclusion decisions.

Statistical analyses

See Table S7 for the questions participants were asked about statistical analyses, and the percentage of participants that selected each response option.

Cleary problematic practices

When asked about whether or not researchers include all dependent measures in their published papers, 5% of participants reported that they sometimes or often exclude dependent measures that yielded nonsignificant results. Relatedly (and more concerning), 1% of participants reported excluding results from dependent variables if they were inconsistent with participants' initial hypotheses. These practices clearly bias the published literature to only featuring significant results and/or results that are consistent with particular hypotheses.

When asked about reliability coding policies, 1% of participants reported only reliability coding data when asked to do so by reviewers. Given that establishing whether findings are robust requires that coding methods are reliable, this practice is clearly questionable.

When asked about data transformation policies, 1% of participants reported that they explore various different transformations on their data and use the one that makes their results look best. The dangers of this and other forms of *p*-hacking, and the impact on Type 1 error rate, have been clearly established (Simmons et al., 2011).

When asked about outliers, 5% of participants reported that they have an outlier policy, but that they do not consistently follow it (because they sometimes forget to check for outliers). Given that we might be biased to check for outliers solely in situations in which the data look somewhat less than ideal, inconsistent implementation of outlier policies provides a problem both for data interpretation and for comparing data across studies.

Risk-permeable practices

Five percent of participants reported that they only consider or plan statistical analyses once their data are in hand. This practice allows for the possibility that analyses are chosen based on how they make the data look. This clearly introduces bias into study results, given that different tests may yield different statistical results for the

same data set, and given that excessive unplanned data analysis allows one to report unpredicted results as if they had been predicted all along.

Training and dissemination

See Table S8 for the questions participants were asked about training and dissemination, and the percentage of participants that selected each response option.

Cleary problematic practices

We did not identify any clearly problematic practices for this section.

Risk-permeable practices

When asked about how policies are recorded, 16% of participants reported not having a lab manual. Presumably, this means that lab policies are passed down through tacit or informal means, which may hinder dissemination and implementation and/or create miscommunications. Furthermore, 17% of individuals who reported having a lab manual also reported that their manual needs updating, suggesting that policies that are documented may be incomplete or out of date. Relatedly, 18% reported that they/their lab has no official system for documenting study progress. Given the long timeline that is typical of infant studies, and the fact that human memory is fallible, a lack of a system for study documentation may mean that important parts of study progress might be lost, forgotten, or misreported.

DISCUSSION

In this paper, we set out to answer several specific questions regarding the state of lab policies in infant research. First, we asked about the extent to which researchers reported actually using set policies (vs. post hoc decision-making) to guide their research. Despite the fact that infancy research involves working with a challenging population, and despite suggestions that infancy researchers often make decisions based on individual discretion on a case-by-case basis (e.g., Peterson, 2016), our findings suggest that by and large infancy researchers reported being guided by a priori policies. Indeed, the vast majority of individuals, for the vast majority of items queried, identified particular policies that guide their research.

Second, our findings suggest that there is some within-lab inconsistency in terms of policies being applied to the same methods. Although presumably some of this inconsistency is due to the use of different ages and procedures within larger methodological groupings, it is likely that it also reflects gaps in within-lab dissemination of policies. These dissemination gaps may be partially accounted for by the fact that students, staff and postdocs learn lab policies as they need to implement these policies; indeed, newer members of labs reported not knowing policies at a much higher rate than did older lab members. Thus, our findings suggest that there is room for improvement in terms of when and how policies are disseminated within labs; we will return to this issue below.

Third, our review of the use of particular policies reveals that different labs regularly implement different policies. This is unsurprising for several reasons: We not only surveyed researchers using a wide array of different methods, but there are also presumably myriad ways to make objective scientific decisions within a particular method. Indeed, a hallmark of science is that what constitutes the exact best policy or practice for a given area of research is often passionately debated; therefore, we view variability in reported practice to be neither surprising nor concerning. Keeping this in mind, we also evaluated particular practices with an eye toward identifying those that are clearly problematic or risk-permeable (practices that may be problematic under select circumstances), to examine whether and to what extent infancy researchers engage in questionable research practices. Our review suggests that there is a fairly low incidence of practices that are obviously harmful to scientific integrity: The overall average of clearly problematic practices was 4%, with the modal response being even lower (1%). Although these results suggest that clearly problematic practices are rare in infancy research, we did identify a number of practices as risk-permeable and sought to identify the particular situations in which they would be problematic.

Here, we note that in raising the issue of risk permeability, it is not our intention to suggest that risk-permeable practices should never be used or used only in specific contexts. Indeed, given the inherent challenges of studying infant populations (infants are challenging to recruit and expensive to test; there is a much more restricted range of methods and paradigms that can be used with infants vs. older children and adults), we contend that there is particular need for methodological flexibility in infancy research and that it would be harmful to our field (and to others') to restrict researchers' ability to adapt their practices to the unique needs of their population and circumstance. Our challenge as infancy researchers, then, is to maintain an appropriate balance between methodological flexibility and rigor; essentially, we must ensure that our flexibility is principled. We hope that our discussion will encourage researchers to evaluate their specific research practices within the contexts in which they are implemented, and will further the conversation regarding how best to navigate the balance between flexibility and rigor in infancy research.

Limitations

There are several limitations to our survey. First, to capitalize on an opportunity to present findings from our survey at an upcoming conference (International Congress for Infant Studies, 2016), we created the survey relatively quickly, and external feedback could only be sought from a handful of outside readers. Consequently, the survey may not be broadly representative of the field, both in terms of the questions asked and in terms of the response options provided. Moreover, because we sought to provide a broad overview of policies within the field, our survey was not methodology-specific and was likely influenced by the particular methods most frequently used in our own work. Finally, due to the quick turn around to complete the survey, the sample size, while not small, did not capture data from all infancy researchers in all infancy research labs. Future work should focus on large-scale, in-depth, and methodology-specific surveys, to accurately capture lab policies as they may differ across infancy subfields.

In addition, although these survey responses provide a first step toward understanding the policies in place across a wide range of infant laboratories, self-report is inherently imperfect: Asking participants about their policies does not guarantee that

participants actually have these policies, nor does it elucidate the extent to which policies are reliably translated into practice. While we focused on gauging researcher policies in this initial survey, more work needs to be done to understand the rates at which policies are manifested in practice.

Suggestions

Based on what we have learned from our survey, we have several suggestions for improving lab policies. We are certainly not the first to make these suggestions, but raise them here again as relatively simple solutions to some of the issues highlighted by our survey. First and foremost, we suggest that labs have a formal means of recording lab policies, either in the context of a written lab manual or as its own separate document. Having lab policies explicitly articulated in a manner that all members of the lab have access to will likely produce greater dissemination and accuracy than more tacit means of communicating lab policies (e.g., direct communication between advisor and student). Fortunately, the vast majority of survey respondents (81%) stated having a lab manual, although the majority of lab manuals are geared toward undergraduate training only. Thus, we suggest that researchers add their lab policies to the lab manual and establish policies for ensuring manuals are kept up to date. Various updatable online formats may be especially suited to this suggestion (e.g., wikis).

Second, we encourage public posting of lab manuals/policies on lab and other Web sites (e.g., Open Science Framework; osf.oi; see osf.io/5cu9q for an example). Greater sharing of lab policies across labs provides the opportunity for researchers to compare and contrast policies and also provides for greater accountability for adherence to lab policies. In fact, to the extent that information about lab policies is publically available, researchers will be poised to empirically evaluate the extent to which researchers follow their lab policies in published work.

Third, we suggest that labs not only carefully record and distribute their lab policies, but also have a means to regularly discuss them. These discussions should help to ensure that dissemination occurs and will allow for policy revision, particularly as informative lessons from meta-science become available (e.g., many of us have recently learned more about the dangers of small sample sizes and changed our minimum sample size policies as a result). Relatedly, labs can consider exercises that encourage students to learn lab policies prior to the need to actively apply them, for example, discussing vignettes or scenarios in lab meetings and considering how lab policies would apply to these scenarios. Engaging in such exercises within a lab may also help to clarify instances in which lab members disagree on lab policies and thus help move labs toward more within-lab consistency.

Fourth, because our survey revealed that trainee knowledge about statistical practices was relatively weak, students, staff, and postdocs can be encouraged to write analytic plans that specify analyses prior to data collection. These plans can either be filed "in house" or preregistered via online Web sites/tools (Open Science Framework: osf.io; AsPredicted: aspredicted.org). Doing so will help trainees and faculty alike to "think ahead" to consider, and make, a priori decisions regarding data analyses. Such practices often have additional positive consequences for study design (e.g., the realization that a particular study design or dependent measure does not lend itself particularly well to a given analysis).

Finally, while ideally lab policies would be as consistent as possible across studies (at least within a given method), in the course of research new methodological paradigms and techniques arise, and thus, policies and practices can (and should!) change over time. In situations in which there is good reason to veer from lab policies (e.g., in the course of the development of a new paradigm), general lab policies can be supplemented with preregistrations that detail any variations from the (already publically posted) usual lab policy. Indeed, preregistration effectively achieves many of the same goals as having and using particular lab policies, providing a means for decision-making that is not contingent upon data. Although it can sometimes be difficult to preregister all aspects of an infant study, particularly when using a new paradigm, a starting approach to preregistration can involve preregistering whatever aspects of a procedure or study one can, even if these details are fairly restricted or minimal (see Poldrack, 2016).

CONCLUSIONS

A critical first step toward improving practices in any scientific domain is a thorough understanding and knowledge of practices that are currently in place. Our survey of infancy researchers and research labs provides us with a starting place for articulating what those practices are and how they can be improved. Survey results revealed considerable strengths within the field of infancy: Research is largely guided by scientifically grounded policies, and the reported use of clearly problematic research practices is low. Results also revealed some areas for improvement: Our field can work toward both greater dissemination of policies within labs and increasing within-lab adherence to lab policies.

The field of infancy research has a long history of deep and critical consideration of methodological practices aimed at providing accurate and rigorous information regarding the nature and state of infants' perception, cognition, and behavior (Aslin, 2007, 2012; Aslin & Fiser, 2005; Csibra et al., 2008, 2016; DeBoer et al., 2007; Gervain et al., 2011; Horowitz, 1974; Oakes, 2010; Smith et al., 2015; Spelke, 1985; Werker et al., 1997). It is our hope that the findings from our survey, as well as the accompanying suggestions, will provide a next step for continuing this tradition of scientific vigilance and self-scrutiny, toward the ultimate end of maximizing the evidentiary value of our science.

ACKNOWLEDGMENTS

This research was supported by a grant to the National Science Foundation, Award # BCS 1639747 to the third author. We would like to thank John Columbo, Lisa Oakes & Melanie Soderstrom for their comments on the original survey, members of the Social Cognitive Development and Early Childhood Cognition Labs at the University of Washington for comments on an earlier draft of the manuscript, as well as Susanne Kirchner-Adelhart for her help with data and manuscript preparation.

REFERENCES

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

Aslin, R. N. (2012). Infant eyes: A window on cognitive development. *Infancy*, *17*, 126–140.

Aslin, R. N., & Fiser, J. (2005). Methodological challenges for understanding cognitive development in infants. *Trends in Cognitive Sciences*, *9*, 92–98.

Berger, J. O., & Selke, T. (1987). Testing a point null hypothesis: The irreconcilability of *P* values and evidence. *Journal of the American Statistical Association*, *82*, 112–122.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Mengyel, M. (2016). Statistical treatment of looking time data. *Developmental Psychology*, *52*, 521–536.

Csibra, G., Kushnerenko, E., & Grossmann, T. (2008). Electrophysical methods in studying infant cognitive development. In C. A. Nelson, & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (2nd edn, pp. 247–262). Cambridge, MA: MIT Press.

DeBoer, T., Scott, L. S., & Nelson, C. A. (2007). Methods for acquiring and analyzing infant event-related potentials. In M. de Haan (Ed.), *Infant EEG and event-related potentials* (pp. 5–37). New York, NY: Psychology Press.

Gervain, J., Mehler, J., Werker, J. F., Nelson, C. A., Csibra, G., et al. (2011). Near-infrared spectroscopy: A report from the McDonnell infant methodology consortium. *Developmental Cognitive Neuroscience*, *1*(1), 22–46.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89.

Horowitz, F. D. (1974). Visual attention, auditory stimulation, and language discrimination in young infants. *Monographs of the Society for Research in Child Development*, *39*, 5–6.

International Congress for Infant Studies. (2016). *Best Practices in Infant Research Pre-conference*. New Orleans: LA, May 2016.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi: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*, 524–532.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*, 701–710.

Ledgerwood, A. (2014a). Introduction to the special section on advancing our methods and practices. *Perspectives on Psychological Science*, *9*, 275–277.

Ledgerwood, A. (2014b). Introduction to the special section on moving toward a cumulative science: Maximizing what our research can tell us. *Perspectives on Psychological Science*, *9*, 610–611.

Ledgerwood, A. (2016). Introduction to the special section on improving research practices: Thinking deeply across the research cycle. *Perspectives on Psychological Science*, *11*, 661–663.

Oakes, L. M. (2010). Using habituation of looking time to assess mental process in infancy. *Journal of Cognition and Development*, *11*, 255–268.

Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy*, *17*(1), 1–8.

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

Peterson, D. (2016). The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius*, *2*, 2378023115625071. doi:10.1177/2378023115625071.

Poldrack, R. (2016). Why pre-registration no longer makes me nervous [Blog post]. Retrieved from http://www.russpoldrack.org/2016/09/why-preregistration-no-longer-makes-me.html

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.

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*, 1359–1366.

Smith, L., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*, 407–419.

Spelke, E. S. (1985). Preferential-looking methods as tools for the study of cognition in infancy. In G. Gottlieb, & N. A. Krasnegor (Eds.), *Measurement of audition and vision in the first year of postnatal life: A methodological overview* (pp. 323–363). Westport, CT, USA: Ablex Publishing.

Werker, J. F., Polka, L., & Pegg, J. E. (1997). The conditioned head turn procedure as a method for assessing infant speech perception. *Early Development and Parenting*, *6*, 171–178.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article:

Table S1. Laboratory characteristics and participants demographics.
Table S2. Laboratory methodologies.
Table S3. Piloting practices.
Table S4. Sample size.
Table S5. Condition assignment and blinding.
Table S6. Inclusion and exclusion criteria.
Table S7. Statistical analyses.
Table S8. Training and dissemination.