INFANCY

THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES

Infancy, ***22***(4), 436–469, 2017

ISSN: 1525-0008 print / 1532-7078 online
DOI: 10.1111/infa.12186

WILEY Blackwell

Sample Size, Statistical Power, and False Conclusions in Infant Looking-Time Research

Lisa M. Oakes
UC Davis

Infant research is hard. It is difficult, expensive, and time-consuming to identify, recruit, and test infants. As a result, ours is a field of small sample sizes. Many studies using infant looking time as a measure have samples of 8–12 infants per cell, and studies with more than 24 infants per cell are uncommon. This paper examines the effect of such sample sizes on statistical power and the conclusions drawn from infant looking-time research. An examination of the state of the current literature suggests that most published looking-time studies have low power, which leads in the long run to an increase in both false positive and false negative results. Three data sets with relatively large samples (>30 infants) were used to simulate experiments with smaller sample sizes; 1,000 random subsamples of 8, 12, 16, 20, and 24 infants from the overall samples were selected, making it possible to examine the systematic effect of sample size on the results. This approach revealed that despite clear results with the original large samples, the results with smaller subsamples were highly variable, yielding both false positive and false negative outcomes. Finally, a number of emerging possible solutions are discussed.

There has been much discussion in the scientific literature broadly, and increasingly in the psychological literature in particular, about whether or not there exists a *replication crisis* in science (Crandall & Sherman, 2016; Open Science Collaboration, 2015; Pashler & Harris, 2012; Stroebe & Strack, 2014). Although it is debated whether the problem has arisen to the level of a crisis, one outcome of this discussion is an increased awareness of the effect of conventional scientific practices on the conclusions we draw from our studies. One area that has received attention is the effect of less-than-optimal sample sizes on false positive and false negative conclusions (Button et al., 2013; Fraley & Vazire, 2014; Schweizer & Furley, 2016; Vadillo, Konstantinidis, & Shanks, 2015). This is particularly an issue in fields or areas of inquiry in which subject populations are difficult to identify or recruit, or in which the testing of individual subjects is time-consuming and expensive. Research involving infants as participants fits these criteria, and thus, it is important to carefully consider how sample sizes are established and whether our current conventions should be adjusted.

Correspondence should be sent to Lisa M. Oakes, Center for Mind and Brain, UC Davis, Davis, CA 95618, E-mail: lmoakes@ucdavis.edu

There have been discussions and debate about the issue of sample sizes in science broadly, and how decisions about sample size contribute to statistical power (see, e.g., Desmond & Glover, 2002). Button et al. (2013), for example, examined the effect of low-powered studies on the field of neuroscience. They argue for changes in research practices to deal with low power, arguing that this is important for drawing strong conclusions from studies. Low power not only reduces the sensitivity to detect true differences, it also increases the likelihood of observing a false positive result (as a result of the bias to publish significant effects). Thus, studies with low power not only create the problem of having difficulty interpreting nonsignificant small effects, but also increase the proportion of studies in which a spurious effect is taken to reflect the truth.

That is, it is not often recognized that sample size—and statistical power—directly relates to the likelihood of making both type 1 and type 2 errors. By convention, we select our type 1 error rate (e.g., the likelihood of concluding falsely that a difference exists) as 5%—or $p = .05$. We select our type 2 error rate (e.g., the likelihood of failing to detect a true difference) as 20%—or *power* of .80. However, we rarely consider how factors, such as sample size, influence these rates. Moreover, although our estimates of power appear to be independent of our type 1 error rate, in reality *p*-values are much more variable (and less reliable) with low power (Halsey, Curran-Everett, Vowler, & Drummond, 2015), and *p*-values get smaller with increased sample size (Motulsky, 2015).

It would be unproductive to insist that all studies have very large sample sizes. Power and *p*-value depends both on the size of the effect and the sample size. If the true effect is large, a smaller sample would provide sufficient power to detect that effect. Moreover, power analyses are imperfect and requiring large samples may make it impossible for some research to be conducted at all (Bacchetti, 2013). More controversial is the possibility that extremely large samples may yield many statistically significant but very small effects that are not meaningful (Quinlan, 2013). Thus, it is important to not only consider the sample size, but also the effect size.

It is also to point out that studies with small sample sizes (and lower power) can be an important part of scientific discovery, and it is critical that we not abandon or reject all studies with low power. But, it is clear that researchers must carefully consider the implications of target sample sizes, both for the time and expense of conducting research and for the conclusions that can be drawn from the study once the target sample has been obtained. The bottom line message here is that it is important that a field not depend exclusively on studies with small samples and that research with small samples be considered in the context of a larger body of research.

The discussion of small samples and underpowered studies is particularly relevant to the study of infant development. Infant research is hard. Many variables can affect our measurement and our conclusions—we must identify and use reliable and valid measures, develop sensitive measurement procedures, train experimenters, and maintain well-trained experimenters. Problems in any of these will influence our measurement and may cause us to draw an incorrect conclusion. Another source of difficulty is the ability to recruit adequate samples of infant subjects. As in other areas of research with specialized populations or expensive and highly technical methods, it can be difficult, time-consuming, and expensive to identify, recruit, and test a large number of infant research participants. As a result, researchers examining infant development often opt for testing as few infants as possible per cell or condition. In this

demonstration, I focus on studies with infants using *looking time* as the dependent measure. I focus on these studies because they are widely used across many areas of infant development, they have been in use for decades (and thus standards and conventions are well established), and because of the relative ease of use, these methods are likely to continue to be a primary way of assessing infant development. To be clear, the specific conclusions drawn here about power levels and conventional sample size can only be directly applied to research using these methods; but the general conclusions about the relations between power, sample size, and conclusions can be applied to other methods.

Standards and conventions have evolved such that most published research using infant looking time is conducted with 8–24 infants per cell (see later section). However, it is not clear that these sample sizes were chosen on the basis of formal power analyses, nor it is clear that these sample sizes provide sufficient power to test the hypotheses under study. It might seem that only large and robust effects would be significant in studies with low power, and so we should have even more confidence in results from such studies (see Friston, 2012). However, given the bias to publish significant results and not nonsignificant results, low power actually increases the proportion of false positive results across the field (see later section).

The goal of this paper is to evaluate sample sizes in infant research by undertaking a careful consideration of looking-time research. In this context, I will delineate the effect of those sample sizes on effect sizes, statistical power, and the incidence of false positive and false negative conclusions in the literature. Several caveats must be made. First, the goal is not to advocate only for very large samples—this would eliminate much of the important work in infant development, and would mean that only some researchers could contribute to this field (and there is some concern about overpowered studies; see Friston, 2012; Quinlan, 2013). Rather, this paper is intended to inspire discussion within laboratories, among researchers, and across the field about the consequences of the decisions we make, how to best decide on what samples we should use, and about alternative approaches to data collection and analysis.

Second, although for simplicity I focus here on a relatively narrow set of studies from a methodological standpoint (i.e., only studies that measured infant looking time), the message here is not only about how to improve research using this method. Instead, the present discussion about these issues in a very well constrained problem provides a model for thinking about sample size, statistical power, and conclusions more broadly.

Finally, the goal is to provide a starting point for changing conventions, and for setting standards for reporting and interpreting results. The bottom line is that we should be considering not only *p*-values, but also effect sizes and power when drawing conclusions from our research (Fraley & Vazire, 2014).

The paper includes three sections. The first section is a discussion of the influence of sample size on statistical power in infant research. This is not a mathematical discussion of power, or a discussion of different ways to calculate power—there are other good sources for that information (Cohen, 1992; Halsey et al., 2015; Krzywinski & Altman, 2013). Rather, this section focuses on broader conceptual issues, discussing how low-powered studies might impact the field of infant development in nonobvious ways. The second section is a description of the state of the field with respect to power and samples size in infant looking-time studies. In this section, 70 papers are reviewed, examining the power and sample size for a single effect reported in each paper

(typically the first or main statistical test reported for the first experiment in the paper). The goal of this description is to demonstrate that sample sizes in this field are determined by convention rather than by formal estimates of power. Although the main conclusion from this section is that infant looking-time studies have low power, the goal is not to cast doubt on the conclusions from this body of work. Indeed, the assumption is that many reported studies provide an accurate understanding of infants' development. Rather, the goal is to describe common practices in the field as a starting point for a discussion of how the field might evolve in beneficial ways, and ways that would allow more confidence about the reproducibility of reported findings as well as in the conclusions we draw from our reported results.

In the final section, three example data sets are explored to establish the influence of sample size on the conclusions that are drawn from a given study. All three data sets were collected in my laboratory and are relatively large for this field (>30 infants per cell). The relatively large numbers of infants included in these data sets provide the opportunity to examine how the results would vary as a function of sample size. Specifically, by randomly drawing subsamples of different sizes from these larger samples, we simulate experiments with smaller samples, and can directly see what effect the sample size would have on the conclusions that could be drawn.

SAMPLE SIZE AND INFANT RESEARCH

Our understanding of infant development was dramatically altered by the development of looking-time measures by Robert Fantz in the 1950s and 1960s (Fantz, 1958, 1963, 1964). Adapting methods developed for use with chimpanzees, Fantz demonstrated that young infants' looking behavior is systematically related to sensory and cognitive factors. Fantz's first studies demonstrated that infants prefer to look at patterned stimuli than at unpatterned stimuli (Fantz, 1958, 1963) and that early preferences were for stimuli that resembled human faces (Fantz & Nevis, 1967). Although these revelations seem modest now, this early worked opened the door for the study of cognitive and perceptual abilities in infants, and led the way to the current state of the field in which we use looking time to draw conclusions about infants' perception of emotions (Peltola, Leppänen, Palokangas, & Hietanen, 2008; Young-Browne, Rosenfeld, & Horowitz, 1977), theory of mind (Onishi & Baillargeon, 2005), understanding of physical relations (Muentener & Carey, 2010; Spelke, Breinlinger, Macomber, & Jacobson, 1992), categorization (Oakes & Ribar, 2005), word learning (Graf Estes, Evans, Alibali, & Saffran, 2007), and much, much more.

Despite the advances that have allowed us to dig deeper into infants' developing abilities, there are many challenges to conducting infant work. In addition to the problem of working with uncooperative subjects who have few motor or voluntary abilities (e.g., we cannot ask them to fill out a questionnaire or complete a task on a computer), infant researchers must identify a pool of potential research participants, effectively recruit participants from that pool, and maintain a laboratory with trained personnel to conduct the studies. Each of these tasks is time-consuming, expensive, and difficult. Few infant researchers feel awash with data. Many researchers struggle to test enough infants to meet their goals at critical career points, such a completing a dissertation, conducting a body of work substantial enough to be awarded tenure, or making sufficient progress on a grant-funded project to be awarded a renewal.

An informal poll of the members of the International Congress for Infant Studies (via the listserv) in September 2016 revealed a significant amount of variability in the rate at which infant researchers can collect data. Some researchers indicated that it was impossible for them to recruit infant research participants and they had given up on testing infants altogether. Other laboratories reported being quite productive, testing 20–40 infants per week. However, to test large numbers of infants, a laboratory will simultaneously test infants of different ages and conduct several studies—often a laboratory is conducting 10 or more studies simultaneously. Indeed, although one laboratory indicated that a single experiment could be completed in a month, the most productive laboratories generally indicated that it takes at least 3 and often 6 months to complete a single experiment. Many researchers reported that they could test 10 infants (or fewer) in a given week (typically by recruiting multiple ages at once). These researchers feel lucky if they can accumulate data from 300 to 400 infants in a year, divided among many different studies. Even when the acquisition rate is low, laboratories are running several studies simultaneously and testing infants of several different ages. Given attrition, pilot testing, and other factors, this means that it may take many months to complete a single experiment even when sample sizes are low. Indeed, some researchers indicated that data collection for a single experiment can take a year or more. Because many papers include the data from multiple experiments, researchers must often collect data for over a year to complete the data collection for a single paper.

Perhaps in part as a result of these difficulties, we are a field that values effects observed with relatively small samples. In 2014, Wally Dixon conducted a survey of researchers in child development, asking them to endorse papers published since 1960 that were important, revolutionary, fascinating, or controversial. He published in SRCD *Developments* (the newsletter of the Society for Research in Child Development) a series of articles reporting these results in a series of the "Twenty most _____ studies in child psychology" lists—the 20 most important, revolutionary, fascinating, and controversial studies. Several infant looking-time studies appeared on these lists, in particular Baillargeon (1987), Baillargeon, Spelke, and Wasserman (1985), Hamlin, Wynn, and Bloom (2007), Onishi and Baillargeon (2005), Saffran, Aslin, and Newport (1996), and Wynn (1992). Clearly this is not an exhaustive list of all infant looking-time studies, and many other studies have had a significant impact on the field. However, given that these papers were identified as controversial, fascinating, important, and/or revolutionary, these studies have clearly had a significant impact. The sample sizes in these studies ranged from 12 per cell (Baillargeon, 1987; Hamlin et al., 2007) to 24 per cell (Saffran et al., 1996). These papers were published in top journals, have been widely cited, and have been the source of considerable discussion and debate in the field.

These studies illustrate the range of sample sizes of highly visible, influential studies of infant looking. If the field has relied on rules of thumb and convention to determine sample sizes, researchers will rely on studies like those in the previous paragraph to provide a standard for target sample sizes. Indeed, as described in the review of the recent literature described in the next section, these are the samples sizes that are most commonly used in infant looking studies.

However, relying on convention and rules of thumb—rather than formal power analyses—to determine sample sizes can result in studies with low power. It is well understood that this may be a problem for detecting a true effect—that is, lower power

by definition means lower likelihood that a real effect will be statistically significant. Counterintuitively, however, low power also decreases the likelihood that *significant* effects are true effects (Button et al., 2013). In other words, the proportion of published results with significant effects that reflect true effects will be lower if we decrease the probability that studies with true effects are significant. To make this more concrete imagine that 200 studies are conducted. Let's further assume that in 100 of these studies the null hypothesis is true and in the other 100 the null hypothesis is false. With the conventional alpha of .05, 5% of the 100 studies in which the null hypothesis is true will yield a significant effect (i.e., a false positive). When the null hypothesis is false, however, we expect a significant effect. But, if we have only 0.2 only 20 of the 100 studies (in which the null effect was false) will yield a significant result (i.e., a true positive). Further, given publication biases, we expect that only the studies with significant effects would be published, the five false positives and the 20 true positives. Thus, of the 25 published studies with significant effects, five (20%) would be false positives, which is far higher than the 5% rate one might expect with an alpha of .05. Moreover, because our hypothetical studies had low power, 80 studies were conducted that led to false negatives. Thus both the published literature and the "file drawer" results misrepresent the status of real and potentially important effects.

Of course this is an extreme example, and (as will be clear later) power is usually higher than 0.2. However, the conclusion remains the same: *p*-values are not independent of power (Halsey et al., 2015), and *p*-values should not be considered without also considering the sample size (Royall, 1986). Of course, power analyses are not without controversy (see McShane & Böckenholt, 2014; and Muthén & Muthén, 2009 for alternative ways to calculate and determine power and sample size), making it even more difficult to know how to evaluate a body of literature. However, it is important to characterize a research area using standard methods to examine the power and effect sizes in a collection of studies. This will allow us to better understand the scope of the problem. This was the goal of the next section.

POWER AND SAMPLE SIZE

Why do we care about sample sizes? Given that the highly influential studies described above yielded positive results with samples that ranged from 12 to 24 infants per cell, why is it a problem to use samples of this size? One might conclude that due to the nature of infant research (i.e., difficulty of measurement, reliability, and so on) only large effects can be reliably detected, and therefore, 12–24 infants per cell is a sufficient convention. Indeed, studies—such as the classic important studies referenced earlier—with such sample sizes have made significant and key contributions to the field. Moreover, error in measurement and experimental design also contribute to false conclusions in infant research. Investigators must carefully consider how the reliability and validity of their measurement, as well as other factors in their experimental design or procedure, make them more or less confident in their conclusions.

But sample size can have powerful effects on outcomes and is subject to conventions. As will become clear in the following paragraphs, conventional sample sizes have yielded many published infant looking-time studies with low power, potentially inflating the publication of false positive results. Indeed, it is possible that some controversies in the field—for example, about infants' developing numerical abilities

(Cohen & Marks, 2002; McCrink & Wynn, 2004; Simon, Hespos, & Rochat, 1995; Wynn, 1992)—reflect at least in part the use of small samples combined with a bias to publish positive findings. We may observe fewer conflicting results if power was routinely considered as a factor in evaluating work, especially when considering a body of research. To be clear, there may be cases when a study with a relatively small sample makes an important contribution to the literature. What I am advocating here is considering the consequences for the field when most studies have small samples, and adjusting our conventions accordingly.

The focus here is on infant looking-time studies, and as a result, the conclusions about specific samples sizes can only be applied to studies using those methods. However, the issues discussed here likely can be applied broadly to studies with infants using a variety of methods, and the examples and methods presented here may provide a model for evaluating other approaches. For the purposes of the present discussion, I focused on a constrained set of methods, measures, and procedures. Different methods, measures, and subject populations yield different levels of variability, there will be variation in effect sizes and sample size requirements as a function of what method is used or what measures are analyzed. For these reasons, the evaluation presented here focused narrowly on infant looking-time studies.

The general point is that statistical power is critically important for interpreting the results of empirical studies. Higher power is not only important for sufficient sensitivity to detect true effects, higher power also is associated with more accurate estimates of effect sizes and lower probability of false positive results (Fraley & Vazire, 2014). However, behavioral scientists often lack a clear understanding about the importance of statistical power for interpreting their findings (Vankov, Bowers, & Munafò, 2014), or even what sample sizes are required to obtain sufficient power (Bakker, Hartgerink, Wicherts, & van der Maas, 2016). Moreover, it is tempting to conclude that when a field is restricted to small sample sizes, our science is more likely to report only large effects. However, the bias to publish only significant results means that the published literature likely overestimates effect sizes, and the reported effects in published papers may be twice the true effect sizes (Brand, Bradley, Best, & Stoica, 2008; Lane & Dunlap, 1978; Open Science Collaboration, 2015).

Despite the possibility that publication practices and biases may create inaccuracies in our understanding of phenomena, the problem of low power may be pervasive in science. Recent reviews suggest that many published studies have low power in neuroscience (Button et al., 2013) and psychology (Fraley & Vazire, 2014). Given the difficulty of recruiting infant subjects, it would not be surprising if the conventional sample sizes in infant looking-time studies often yield relatively low power. However, this topic has not been discussed much in the context of infant looking-time studies.

Table 1 lists all 70 articles using looking-time studies published in the years 2013 to 2015 in a collection of psychology journals that publish large numbers of articles focusing on infants. The articles were published in *Child Development* ($N = 11$), *Cognition* ($N = 3$), *Cognitive Development* ($N = 3$), *Developmental Psychology* (10), *Developmental Science* (4), *Frontiers in Psychology* ($N = 3$), *Infancy* ($N = 7$), *Infant Behavior and Development* ($N = 13$), *Journal of Cognition and Development* ($N = 3$), and *Journal of Experimental Child Psychology* ($N = 13$). Although there are other journals in which the kind of data evaluated here might be reported, I focused on journals that commonly publish this type of work. These papers were scrutinized by experts in infant

TABLE 1
All 70 Articles Using Looking-Time Studies Published from 2013 to 2015 in Top Developmental Journals (see text for details), Including Information About Sample Size, The Selected Statistical Test, Calculated Effect Size, Observed Power, and Sample Sizes Required for the Effect Size. An Asterisk by the Citation Indicates That the Calculated Effect Size Differed from That Reported in the Paper.

Citation	*N per cell*	*Statistical test*	*Calculated effect size*	*Observed Power*	*50% effect size*	*Sample size needed given effect size*	*Sample needed for 50% effect*
Althaus and Plunkett (2015)	29	$t(28) = 4.037$, $p < 0.001$	0.76	0.98	0.38	16	57
Bahrick, Lickliter, and Castellanos (2013)	16	$t(15) = 5.26$, $p < .0001$	1.32	0.99	0.66	7	21
Bahrick, Lickliter, Castellanos, and Todd (2015)	16	$t(15) = 3.27$, $p = .005$	0.82	0.87	0.41	14	49
Baker, Mahamane, and Jordan (2014)*	28	$t(27) = 2.474$, $p = .02$	0.49	0.70	0.24	35	139
Baker, Pettigrew, and Poulin-Dubois (2014)	24	$F(1, 88) = 6.56$, $p = .012$	0.07	0.43	0.03	108	120
Bardi, Regolin, and Simion (2014)	12	$t(11) = 2.768$, $p < .05$	0.79	0.70	0.40	15	52
Benavides-Varela and Mehler (2015)	22	$t(21) = 2.123$, $p < .05$	0.45	0.52	0.23	41	151
Bidet-Ildei, Kitromilides, Orliaguet, Pavlova, and Gentaz (2014)	12	$Z = 1.96$; $p < .05$	1.37	0.99	0.69		
Biro, Verschoor, Coalter, and Leslie (2014)	12	$t(22) = 2.54$, $p = .019$	1.08	0.72	0.54	15	55
Bremner, Slater, Mason, Spring, and Johnson (2013)	12	$F(1, 10) = 10.95$, $p < .008$	0.52	0.91	0.26	11	26
Brower and Wilcox (2013)	10	$t(18) = -0.58$, $p = .58$	0.26	0.09	0.13	234	930
Cantrell, Boyer, Cordes, and Smith (2015)	20	$t(19) = 3.05$, $p = .007$	0.68	0.82	0.34	19	70
Casasola and Park (2013)	16	$F(1, 15) = 5.38$, $p < .05$	0.26	0.49	0.13	26	56
Cashon, Ha, Allen, and Barna (2013)	23	$t(22) = 3.44$, p	0.72	0.91	0.36	18	63
Coubart, Izard, Spelke, Marie, and Streri (2014)	16	$F(1, 12) = 12.8$, $p = .004$	0.52	0.97	0.26	11	26
Esteve-Gibert, Prieto, and Pons (2015)	24	$F(1, 20) = 7.262$, $p = .014$	0.27	0.31	0.13	25	56
Ferry, Hespos, and Gentner (2015)	11	$t(10) = 3.577$, $p = .005$	1.07	0.89	0.54	9	24
Flom, Janis, Garcia, and Kirwan (2014)*	20	$t(19) = 2.5$, $p = .02$	0.56	0.66	0.28	28	103
Frick and Möhring (2013)	20	$F(1, 36) = 6.94$, $p < .05$	0.16	0.27	0.08	14	26
Frick and Wang (2014))	7	$t(13) = 2.82$, $p = .01$	1.45	0.70	0.73	9	31

Table 1 *(Continued)*

Citation	*N per cell*	*Statistical test*	*Calculated effect size*	*Observed Power*	*50% effect size*	*Sample size needed given effect size*	*Sample needed for 50% effect*
Gazes, Hampton, and Lourenco (2015)	32	$F(1, 28) = 10.21$, $p = .003$	0.27	0.84	0.13	25	56
Graf Estes and Hay (2015)*	16	$t(15) = 2.28$, $p = .038$	0.57	0.57	0.29	27	96
Gustafsson et al. (2015)	30	$F(1, 29) = 4.588$, $p = 0.041$	0.14	0.32	0.07	52	108
Henderson and Scott (2015)*	16	$t(15) = 2.20$, $p < 0.05$	0.55	0.54	0.28	28	103
Hernik and Csibra (2015)	16	$t(15) = 2.65$, $p = .018$	0.66	0.69	0.33	21	75
Heron-Delaney, Quinn, Lee, Slater, and Pascalis (2013)	18	$t(17) = 2.44$, $p = .025$	0.58	0.64	0.29	26	96
Hillairet de Boisferon, Uttley, Quinn, Lee, and Pascalis (2014)	23	$t(22) = 2.89$, $p < .01$	0.60	0.78	0.30	24	90
Hillairet de Boisferon et al. (2015)	18	$t(17) = 2.29$, $p < .05$	0.54	0.58	0.27	29	110
Hock, Kangas, Zieber, and Bhatt (2015)	15	$t(14) = 2.84$, $p < .02$	0.73	0.75	0.37	17	60
Imura, Masuda, Shirai, and Wada (2015)	19	$t(18) = 4.39$, $p = 0.0004$	1.01	0.98	0.51	10	33
Kampis, Somogyi, Itakura, and Király (2013)	22	$F(1, 21) = 7.03$, $p = 0.015$	0.25	0.61	0.13	27	56
Kavsek and Marks (2015)	16	$F(1, 15) = 55.13$, $p \leq .001$	0.79	1.00	0.39	5	16
Kwon et al. (2014)	36	$t(35) = 4.32$, $p < .001$	0.72	0.99	0.36	18	63
Lee, Cheal, and Rutherford (2015)	23	$t(21) = 2.60$, $p = 0.02$	0.55	0.71	0.28	28	103
Lewkowicz (2013)	35	$F(1, 67) = 4.30$, $p < .05$	0.06	0.09	0.03	126	257
Lewkowicz, Minar, Tift, and Brandon (2015)*	24	$t(23) = 0.52$, $p = .30$	0.11	0.08	0.06	651	2183
Libertus and Needham (2014)	16	$t(15) = 2.7$, $p < .02$	0.67	0.71	0.34	20	70
Liu et al. (2015)*	11	$t(10) = 3.16$, $p = .01$	0.95	0.81	0.48	11	37
Longhi et al. (2015)	15	$F(1, 14) = 6.015$, $p < .028$	0.30	0.58	0.15	22	48
Loucks and Sommerville (2013)	13	$t(12) = 2.38$, $p = .035$	0.66	0.59	0.33	21	75
Mackenzie, Graham, Curtin, and Archer (2014)	16	$t(15) = 3.56$, $p = .003$	0.89	0.91	0.45	12	41
Matatyaho-Bullaro, Gogate, Mason, Cadavid, and Abdel-Mottaleb (2014)	12	$t(11) = 6.44$, $p < .001$	1.94	1.00	0.97	5	11

Table 1 *(Continued)*

Citation	*N per cell*	*Statistical test*	*Calculated effect size*	*Observed Power*	*50% effect size*	*Sample size needed given effect size*	*Sample needed for 50% effect*
May and Werker (2014)*	24	$F(1, 22) = 10.67$, $p < .01$	0.33	0.87	0.16	19	45
Moher and Feigenson (2013)	18	$F(1, 16) = 21.996$, $p < .001$	0.58	1.00	0.29	9	23
Novack, Henderson, and Woodward (2013)*	16	$t(15) = 3.50$, $p < .003$	0.88	0.91	0.44	13	43
Oakes and Kovack-Lesh (2013)	12	$t(11) = 2.06$, $p = .06$	0.59	0.46	0.30	25	90
Otsuka et al. (2013)	12	$t(11) = 3.04$, $p < .01$	0.88	0.79	0.44	13	43
Ozturk, Krehm, and Vouloumanos (2013)	12	$t(11) = 2.77$, $p < .02$	0.80	0.71	0.40	15	52
Park and Casasola (2015)	15	$F(1, 32) = 20.86$, $p < .001$	0.39	0.80	0.20	16	35
Perone and Spencer (2014)*	39	$t(38) = 3.50$, $p < .001$	0.56	0.93	0.28	28	103
Pruden, Roseberry, Göksun, Hirsh-Pasek, and Golinkoff (2013)*	23	$t(22) = 3.12$, $p < .05$	0.65	0.84	0.33	21	75
Quinn and Liben (2014)	12	$t(11) = 5.14$, $p =$	1.48	0.99	0.74	6	17
Rigney and Wang (2013)	18	$F(1, 17)=10.83$, $p < .004$	0.39	0.87	0.19	16	37
Robson, Lee, Kuhlmeier, and Rutherford (2014)*	24	$t(23) = -2.305$, $p = .031$	0.47	0.60	0.24	38	139
Sanefuji, Wada, Yamamoto, Mohri, and Taniike (2014)	12	$t(10) = 2.79$, $p = .019$	0.84	0.75	0.42	14	47
Sato et al. (2013)	14	$t(13) = 3.04$, $p < 0.01$	0.81	0.80	0.41	15	49
Skerry and Spelke (2014)	32	$F(1, 31) = 8.524$, $p = 0.006$	0.27	0.84	0.13	25	56
Slone and Johnson (2015)	20	$t(19) = 2.76$, $p < .05$	0.62	0.75	0.31	23	84
Soley and Sebastián-Gallés (2015)	32	$t(31) = 2.42$, $p = .02$	0.42	0.63	0.21	47	180
Starr, Libertus, and Brannon (2013)	20	$t(19) = 3.50$, $p < .005$	0.78	0.91	0.39	15	54
Takashima, Kanazawa, Yamaguchi, and Shiina (2014)	13	$t(12) = 2.46$, $p < .05$	0.68	0.62	0.34	19	70
Tham, Bremner, and Hay (2015)	12	$t(10) = 2.342$, $p = .041$	0.71	0.61	0.35	18	67
Träuble and Bätz (2014)*	15	$t(14) = 4.09$, $p < .001$	1.06	0.97	0.53	10	30
Tsuruhara, Corrow, Kanazawa, Yamaguchi, and Yonas (2014)	12	$t(11) = 3.94$, $p < .01$	1.14	0.95	0.57	9	27

Table 1 *(Continued)*

Citation	*N per cell*	*Statistical test*	*Calculated effect size*	*Observed Power*	*50% effect size*	*Sample size needed given effect size*	*Sample needed for 50% effect*
Turati, Gava, Valenza, and Ghirardi (2013)	17	$F(1, 14) = 5.489$; $p = 0.034$	0.28	0.58	0.14	24	52
Vukatana, Graham, Curtin, and Zepeda (2015)	35	$F(2, 68) = 3.27$, $p = .044$	0.09	0.19	0.04	83	192
Woods and Wilcox (2013)	15	$F(1, 14) = 7.57$, $p = .02$	0.35	0.71	0.18	18	39
Yamashita, Kanazawa, and Yamaguchi (2014)*	12	$t(11) = 4.55$, $p = .0008$	1.31	0.98	0.66	7	21
Zieber, Kangas, Hock, and Bhatt (2014)*	16	$t(15) = 2.54$, $p = .02$	0.64	0.67	0.32	22	79
Zieber, Kangas, Hock, and Bhatt (2015)	11	$t(10) = 2.71$, $p < .03$	0.81	0.63	0.41	15	49

research in the peer review process, and we can have confidence that the methods reported have "passed muster" by a broad set of experts from our peer group. Thus, although this is not an exhaustive list of all looking-time studies published during these 3 years, I selected all the studies published in the journals listed above that met the criteria listed in the following paragraphs (I sincerely apologize if I inadvertently omitted from this list any papers published in the listed journals that did meet those criteria). Thus, this sample will provide a good indication of standard, accepted practices in the field.

The goal was to narrow the range of variability to allow us to evaluate the effect of sample size on a well-defined, constrained set of studies. Thus, the articles in this list were selected using the following inclusion criteria. First, because it is plausible that sample size requirements and effect sizes change with the age of the subjects, I included papers only if the infants tested were younger than 18 months. Second, the method involved must have involved observer-recorded looking time. Although I included studies that used a wide range of methods used to record looking time—for example, online recording by one or more observer, offline coding in real time, frame-by-frame coding from recordings of the session—in all of the papers included here the dependent measure was related to looking time (e.g., the duration of looking, the proportion of looking, or difference in looking). Other measures—such as reaction time and number of looks—may have different levels of variability, and therefore, the sample size requirements may be different. In addition, I excluded studies in which an eye-tracker was used. In these studies, eye gaze is recorded quite differently than when coded by a human observer, and the scale of measurement is often quite different (e.g., millisecond level recording with eye-trackers as compared to tenths of second recordings by human observers). Moreover, eye-tracking procedures often involve more trials, finer spatial resolution, and other factors that change both the nature of the measure and the variability observed. Factors such as the validity and reliability of looking-time measures may also differ when looking time is coded by human observers versus automatically

by an eye-tracker. Future work may examine the effects of power and sample size in eye-tracking studies. In all the studies evaluated here, infants' a priori preference, changes in preferences (after familiarization), or response to novelty was assessed. These procedures have been widely used in the field.

There were several other inclusion criteria. Only work that examined development in typically developing, healthy, full-term infants was included. In addition, papers were excluded if their main focus was individual differences or if they tested infants longitudinally. In these instances, the hypotheses were quite different from the studies listed in Table 1, making it difficult to know how power and effect size would compare. It would be extremely useful for a future investigation to examine such studies. Finally, only a single statistical test was evaluated in each study. For many studies, multiple experiments were reported. In this case, only information about the first experiment reported was included. Often this experiment reported the main finding of the study. In three papers, the first experiment reported was a control condition or included adults as participants. In these cases, I included the information about the second reported experiment.

For the present purposes, I selected a single statistical test that was key for the conclusions of the paper. Typically, this was the first reported test that evaluated infants' responding on the test trials. However, if there were multiple analyses, the statistical test associated with the largest effect size was selected. This decision was made to bias the sample toward larger effects, which will favor smaller samples. In two cases a nonsignificant effect is included because that effect was critical for the conclusions of the paper. The specific test used is listed in Table 1. The distribution of group or cell

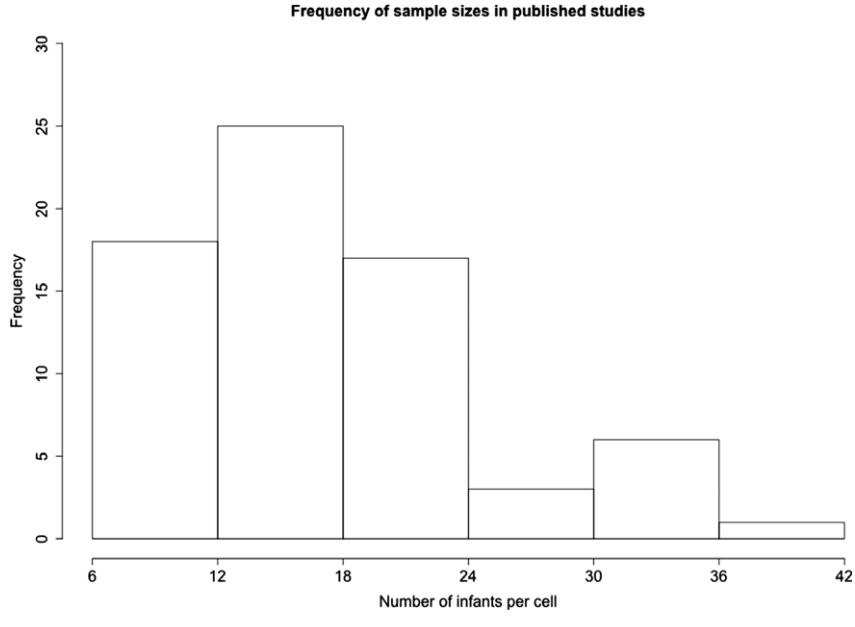

Figure 1 Histogram of the frequency of sample sizes (number of infants per cell) for the studies listed in Table 1. The number indicates the lower value of the adjacent range (e.g., the bins correspond to samples from 6 to 11, 12 to 17, 18 to 23 infants, and so on).

sample sizes is presented in Figure 1. This histogram indicates the frequency of each sample size across the 70 studies.

Several things are immediately apparent. First, the data in Table 1 show that only four of the experiments had 10 or fewer subjects per cell; in three of these experiments, conclusions were drawn by comparing two groups of infants (14–20 infants in total). Ironically, such between-subjects comparisons typically require larger sample sizes to achieve the same power as within-subjects comparisons (Bramwell, Bittnerjr, & Morrissey, 1992; Charness, Gneezy, & Kuhn, 2012), and thus, these studies probably should have had larger sample sizes. Just 11 (15%) of the experiments had 25 or more infants per cell. The vast majority of the experiments had between 11 and 24 infants per cell.

Why have we adopted the convention of testing 11–24 infants per cell in this type of study? One possibility is that the controversial, fascinating, important, and revolutionary studies described earlier used similar sample sizes, and thus, a convention was established because of these influential studies. It is also possible that these sample sizes reflect sufficient power to detect the kinds of effects typically observed in infant research—so the convention is not based on an arbitrary decision, but actually reflects the kind of power needed to detect the true effects that exist. Indeed, in the original studies and most of the studies listed in Table 1, significant effects were observed with these sample sizes. But, conclusions about sample size and power must be drawn carefully when relying solely on published findings. The widespread bias to publish only significant effects makes it much more difficult to determine what the *true* effects is, and therefore whether the power in these studies was sufficient. It is commonly understood that the bias to publish significant effects creates a file drawer problem, in which nonsignificant findings are not reported. When studies have low power, it is likely that there are more such file drawer studies, making it even less likely that the significant finding reflects a true effect. To be clear, assuming the absence of p-hacking (or engaging in practices that inflate the *p*-value, see Head, Holman, Lanfear, Kahn, & Jennions, 2015; Lakens, 2015; Ulrich & Miller, 2015), many published findings must reflect true effects. The problem is that the level of power influences our confidence about whether a particular finding reflects a true effect. Thus, if the convention in a field is to conduct low-powered studies, it becomes less clear what proportion of published findings reflect true effects.

This discussion raises an interesting paradox regarding the pressures of difficult-to-obtain subject populations on conducting studies with low power. If in fact running studies with small sample sizes is more likely to yield nonpublishable, file drawer results, researchers may get more bang for their buck if they ran fewer studies with larger sample sizes. To be clear, I am not advocating for a complete rejection of studies with small samples sizes—there may be some cases and some study populations where small samples sizes are the only option. However, it may be that our reliance on small sample sizes *in general* has actually created a more difficult situation for researchers who have limited access to infants. That is, by creating a culture in which small sample sizes are widely accepted, researchers who have difficulty recruiting infants may fall into the trap of testing many underpowered studies that become "file drawer" studies; these researchers may have more success in general if our conventional sample size yielded studies with higher power. Changing our convention to expect larger samples sizes in general would obviously mean that it take longer to collect the data for a single study. However, these better powered studies may be more likely to yield interpretable (and publishable) findings, reducing the number of file drawer studies.

To address the question of whether our conventional sample sizes provide sufficient power, I calculated the observed effect size in each study using the approach described by Lakens (2013). Although effect size was reported in many studies, Lakens's method was used to calculate the effect size for all the statistics listed here to ensure that effect size was calculated in the same way for all experiments. A handful of calculated effect sizes differed from those reported in the published papers (indicated by an * in the table), perhaps reflecting a different method for calculating effect sizes, an error in calculation, or a typo. The incidence of these inconsistencies suggests that editors and reviewers often are satisfied with the presence of effect sizes, and do not double-check the effect sizes (note that there has recently been a discussion of the prevalence of errors in statistical reporting more broadly, Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015). The development of a tool like statcheck (Epskamp & Nuijten, 2016, a package for R that operates like spellcheck, except for statistical reporting, but does not evaluate effect size) that could detect such errors would be a significant benefit to the field.

Figure 2 provides a histogram of the frequency of the calculated Cohen's *d* scores for the 49 *t*-tests reported in Table 1. We chose to plot only Cohen's *d* because they were more numerous than the η_p^2 reported for *F*-tests. To make sure that the data included in the following figures and discussion were comparable, my evaluation focused only on *t*-tests, as they were the most frequent statistic sampled from the studies (many studies reported both ANOVAs and *t*-tests, but main conclusions were drawn from *t*-tests comparing infants' looking at two tests, or comparing infants' preference to chance). Recall that, by convention, Cohen's *d* of .2 is a small effect, .5 is a medium effect, and .8 is a large effect (Cohen, 1992). Sawilowsky (2009) further

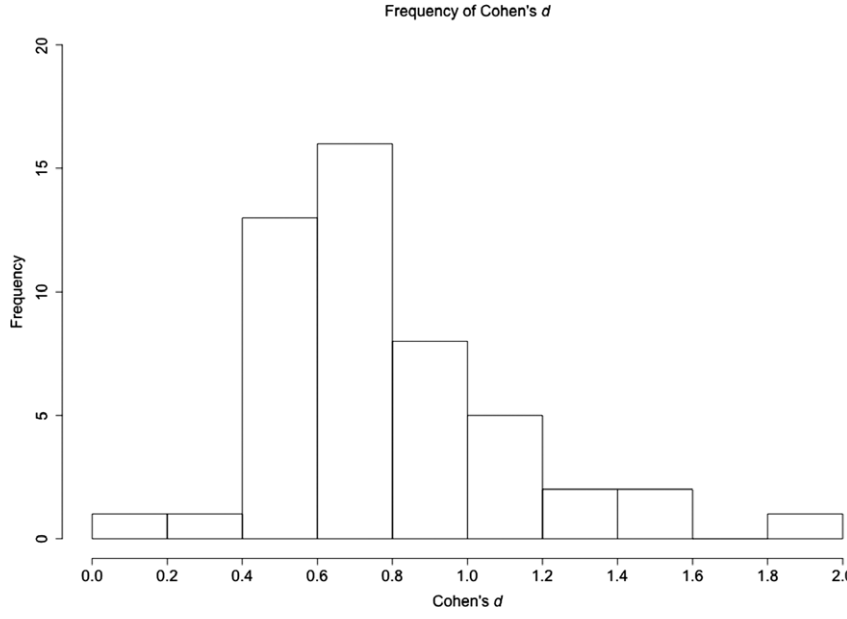

Figure 2 Histogram of the frequency of Cohen's d for the 49 *t*-tests listed in Table 1. The number indicates the lower value of the adjacent range (e.g., the bins correspond to Cohen's *d* from 0 to .19, from .20 to .39, and so on).

suggested that d scores of 1.2 and 2 be considered very large and huge effect sizes respectively. These effect sizes indicate that the means differ by at least 1.2 standard deviations, and indeed are quite large.

Figure 2 shows that 28 of the 49 effect sizes (57%) fall between .4 and .8, and thus fall the medium category Fourteen effect sizes (29%) were between .8 and 1.2 (large), and only five effect sizes (10%) were very large or huge. The two small effect sizes were the two cases in which the *t*-test did not reveal a significant difference, and conclusions were drawn based on a null finding. Thus, these data suggest that the impression that ours is a field of large effect sizes is incorrect. Moreover, given that effects sizes are typically overestimated in studies with small sample sizes (Hedges & Vevea, 1996; Lane & Dunlap, 1978), the data presented here suggest that research evaluating infants looking times is (at best) mainly a field of medium estimated effect sizes and is likely a field in which actual effect sizes are often small.

What does this mean about the conclusions that we can draw from the reported results? After all, these *t*-tests were significant. Perhaps this means that the sample sizes used provided sufficient power to detect those effects, assuming the effect sizes were an accurate estimate of the true effect size (which is likely a generous assumption). To test this possibility, I used G*Power to determine the sample size need to achieve the conventional power level of .80 (assuming *alpha* = .05) using the estimated effect sizes listed in Table 1. In addition, because observed effect sizes are often twice as large as true effect sizes when small sample sizes are used (Lane & Dunlap, 1978; Open Science Collaboration, 2015), Table 1 also includes effect sizes that are 50% of the observed effect size, as well as the sample sizes need to achieve .80 power for these reduced effect sizes.

Figure 3 presents the distribution of sample sizes in the original studies (a), the sample sizes required to achieve .80 power to detect the calculated effect sizes (b), and the

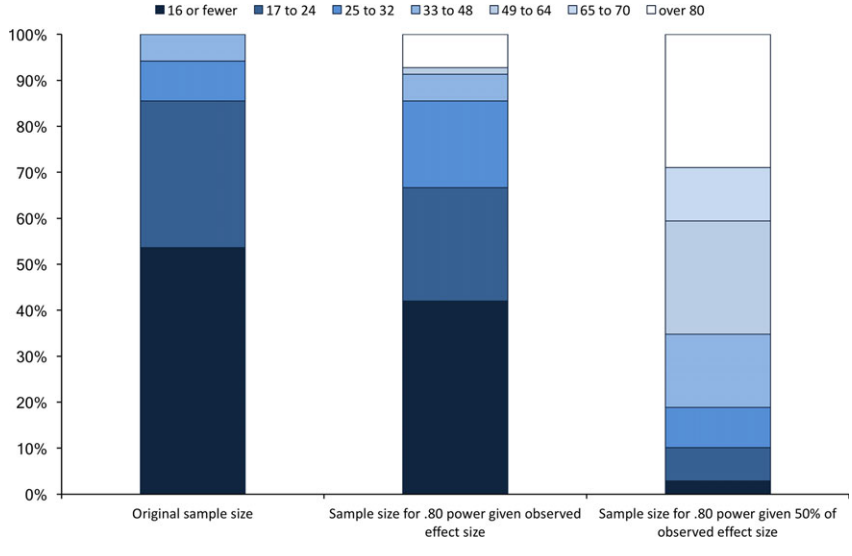

Figure 3 The proportions of sample sizes in the studies reported in Table 1 (a), required to achieve .80 power given the effects calculated from the statistic for each study reported in Table 1 (b), required to achieve .80 power given 50% of the effect size calculated from the reported statistic in Table 1 (c).

sample sizes required to achieve .80 power to detect 50% of those calculated effect sizes (c). It is immediately clear that the three distributions are quite different. Most published studies included 24 or fewer infants per cell (over 80%). However, the histogram in Figure 3 shows that these sample sizes would have provided sufficient power to detect 67% of the observed effect sizes. The samples used in the published studies were rarely large enough to achieve .80 power given the reduced effect sizes (e.g., if the true effect sizes were 50% of those observed in the published studies). Moreover, although few of the original studies had sample sizes of over 32, these larger samples were required to achieve .80 power for detecting a significant effect in the vast majority (81%) of cases if we assume the reduced effect sizes. Thus, assuming that the reported effect sizes are approximately twice the size of the true effect (Lane & Dunlap, 1978; Open Science Collaboration, 2015), it is clear that studies using infant looking time generally have low power.

What is the takeaway message from this analysis? Given that the reported studies tend to have low power, it seems that in this area of research, like many areas of psychological research, has relied on convention and rules of thumb to determine sample sizes. It is tempting to argue that this is not a problem because we are a field of large effects—that is, our sample sizes give us the sensitivity to detect relatively large effects. However, inspection of Table 1 shows that even when experiments yield relatively large effects (>.60), the studies often still have low power. Clearly the likelihood that a true *positive* effect is observed decreases with decreased power (see Krzywinski & Altman, 2013 for a nice description). However, because *p*-values vary more in low-powered studies (Halsey et al., 2015), decreased power may be problematic for conclusions from both positive and null findings. The particular effect of sample size on effect size and power is further explored in the final section through the examination of three different data sets collected in my laboratory.

AN EXPLORATION OF THREE DATA SETS

One challenge with analyses based on published studies is that they necessarily reflect the biases of the publication process (e.g., the strong tendency to focus on significant effects). Moreover, one cannot easily explore the effects of sample sizes in published studies because many other factors may covary with sample size. This section therefore examines the relationship between power and sample size in three relatively large data sets, using a Monte Carlo approach in which experiments with smaller sample sizes were simulated by selecting random subsets of the subjects from these actual experiments.

I selected three studies with relatively large samples sizes; 33 in the first set (published in Kwon, Luck, & Oakes, 2014, Experiment 2), and 32 in the other two sets (Experiment Action and Experiment Sound, both unpublished). The relevant details of each study will be described below. Importantly, because these are real data sets, the true effect size is unknown. But, because the sample sizes are relatively large (in the context of most infant research with looking time as the dependent measure), it is possible to simulate the effects with a variety of sample sizes smaller than the original sample.

The first experiment was a paired preference study with 6-month-old infants; the data represent their preferences on two types of trials. Experiment Action and

Experiment Sound were habituation experiments with 10-month-old infants; the data represent their looking time to familiar and novel test items following a habituation sequence. All data reported were from infants who met the relevant inclusion criteria (e.g., completed all six trials in the Kwon et al. experiment; met the habituation criterion in Experiment Action and Experiment Sound). All infants were healthy, full-term, and had no history of vision problems. No infants were statistical outliers (e.g., all responding was within 3 *SD* of the mean).

To demonstrate the effect of sample size in these experiments, 1,000 subsamples of 8, 12, 16, 20, and 24 infants were drawn without replacement from the full data set using the *samp* function in R (yielding 5,000 subsamples in total for each experiment). These sample sizes were selected to be representative of the sample sizes used in the published literature. As shown in Figure 1, most studies have sample sizes between 6 and 24 infants per cell. The goal here was to examine the mean response and *t*-values across the subsamples of a given size, making it possible to determine how variations in sample size could influence the outcome of the experiment. The simulated sample sizes were based on the typical range of values from the meta-analysis described earlier, and the sample sizes selected for these simulations approximate those that are typically used in the recent literature. By varying the sample size systematically within this range (8–24 infants per cell) it is possible to see how power and *p*-values change over this range (e.g., is the change linear).

Example 1

Kwon et al. (2014) reported the results of a *simultaneous streams* change detection task with 33 six-month-old infants (Experiment 2). Infants were presented with six trials, in which a changing stream was paired with a nonchanging stream (see Figure 4). On each trial, infants' preference for the changing stream was assessed by measuring how long they looked to each stream. A change preference score was calculated by dividing the amount of time infants looked at the changing stream by their total looking (e.g.,

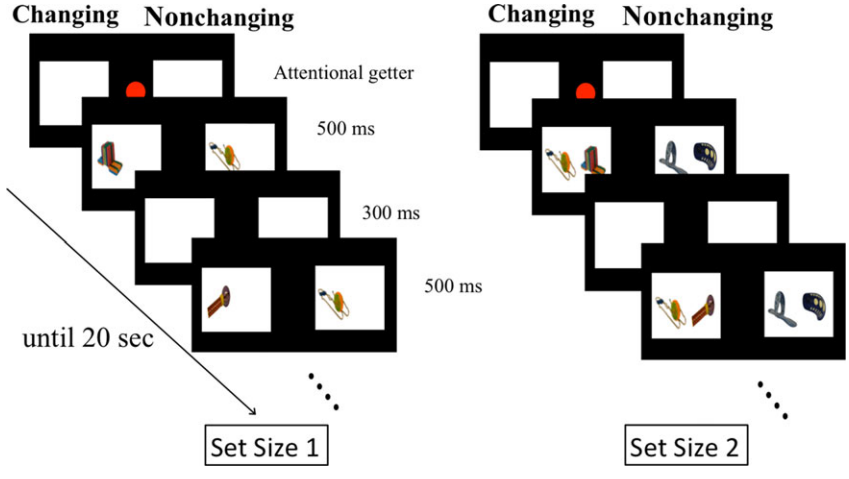

Figure 4 A schematic depiction of the procedure used by Kwon et al. (2014). Reprinted with permission from Wiley.

the changing and nonchanging streams combined). If infants significantly preferred the changing stream, their preference score would be significantly greater than chance, or .50 (i.e., equal looking to the two streams). We tested infants' preference for the changing stream when each stream contained only a single item (set size 1; left side of Figure 4) and when each stream contained two items (set size 2; right side of Figure 4).

Statistical analyses of the entire group of infants showed that infants significantly preferred the changing stream at set size 1, $t(32) = 3.07$, $p = .004$, $d = .54$ (see Figure 5), but they failed to prefer the changing stream at set size 2, $t(32) = -.34$, $p = .84$, $d = .06$. We are assuming that the preference is a true effect at set size 1 and a null (or negligible) effect at set size 2, and these assumptions are based on two sources of evidence. First, several previous studies have found significant effects at set size 1 but not at set size 2 in 6-month-old infants using variants of this procedure (Oakes, Baumgartner, Barrett, Messenger, & Luck, 2013; Ross-Sheehy, Oakes, & Luck, 2003). Second, we conducted a Bayes factor analysis (Rouder, Speckman, Sun, Morey, & Iverson, 2009) which indicated that the data from set size 1 were 8.9 times more likely to arise from a true effect than to arise from a null effect, and the data from set size 2 were 5.1 times more likely to arise from a null effect than from a true effect. Thus, we are justified in assuming that the data from this experiment reflect a true effect at set size 1 and a null or negligible effect at set size 2.

The estimates of the sample size needed to have 80% power given these effect sizes (assuming *alpha* = .05) is 18 for set size 1. Although infants as a group preferred the changing side at set size 1 (the blue circles) but not at set size 2 (the red circles), it is also clear that there was significant variability in infants' responding (the individual

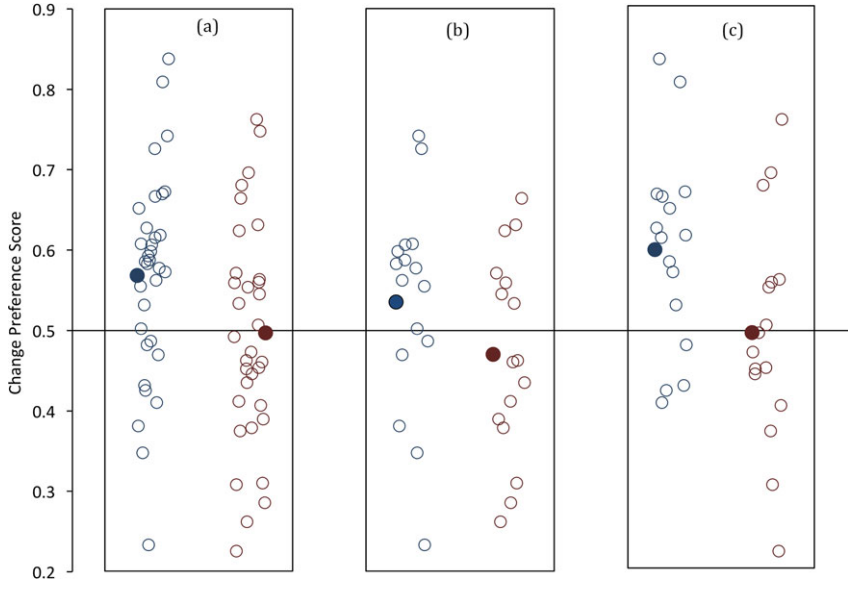

Figure 5 The results of Experiment 2 in Kwon et al. (2014). Means for each sample are given in solid circles and the data from individual subjects are presented in open circles. The proportion expected by chance is .50 (equal looking at the changing and non-changing stream). Panel (a) represents all 33 infants reported in Kwon et al., and panels (b) and (c) represent infants in two nonoverlapping subsamples of 16 infants from the whole sample.

circles in the graph). Moreover, one infant appears to be an outlier at set size 1; his or her mean responding was .233, which is 2.62 *SD*s below the mean. However, our standard exclusion criterion is for values that are 3 *SD* from the mean, and deciding whether and how to exclude outliers after having looked at the results can significantly affect the probability of type 1 error (Bakker & Wicherts, 2014a,b). Therefore, we did not exclude this infant from our analyses.

As a first step, we extracted from this sample of 33 infants two nonoverlapping subsamples of 16 infants to simulate what might happen in experiments with a sample size of only 16 infants. The results from these two subsamples are presented in Figure 5b and 5c. Note that we would have drawn different conclusions if we had sampled only one of these two sets of infants. Specifically, for the first sample of 16 infants, the *t*-test comparing mean preference score at set size 1 was not significant, $t(15) = 1.07$, $p = .30$, $d = .268$ (excluding the potential outlier infant changed the *t*-test to $t(14) = 1.99$, $p = .07$, $d = .52$) nor was the *t*-test comparing the infants' mean change preference score for set size 2 trials to chance, $t(15) = -.07$, $p = .94$, $d = .02$. Thus, from this sample, the results are ambiguous at best, and do not provide clear evidence that infants prefer the changing stream. For the second sample, the *t*-test comparing infants' set size 1 preference scores to chance was significant, $t(15) = 3.22$, $p = .006$, $d = .81$, indicating that they did prefer the changing stream. Their preference for the changing stream at set size 2 was not significant, $t(15) = -.07$, $p = .94$, $d = .02$. If we had tested only the first sample of infants, we would have concluded that we had no evidence that infants significantly preferred a change at set size 1—despite the fact that most of the infants in that sample had change preference scores above .50.

Of course, it is possible that these two subsamples are not representative of what would typically occur with a sample size of 16 in this experiment. The first subsample described may have been particularly nonrepresentative, and the results may have been nonrepresentative of subsamples of 16 infants in general. To test this possibility, we examined the effect of sample size more systematically using the Monte Carlo approach described earlier. For each of five different sample sizes ($N = 8$, $N = 12$, $N = 16$, $N = 20$, and $N = 24$), 1,000 experiments were simulated by randomly sampling (without replacement) from the larger sample of 33 infants. The result was 5,000 subsamples of infants, and each subsample contained the data from 8, 12, 16, 20, or 24 infants. For each subsample, the mean change preference scores for both set sizes, as well as the *t*- and *p*-values when comparing each mean score to chance, were calculated. The distributions of mean change preference scores for set size 1 and set size 2 are presented in Figure 6. Each simulated experiment is represented in the figure as an individual diamond (one representing the mean change preference score for set size 1 and another representing the mean change preference score for set size 2). Importantly, when selecting infants to include in a subsample, their change preference scores for both types of trials were selected. Thus, the mean change preference scores for set size 1 and set size 2 presented in Figure 6 represent the change preferences from the same subsamples; any differences between the scores do not reflect differences in the samples selected, but rather reflect differences in how those same subsets of infants responded on the two types of trials.

This visualization illustrates that the size of the sample has a significant impact on the distributions of the mean change preference for the samples. The distribution of the means for the simulated experiments with eight infants has a larger spread than the distribution of means for the simulated experiments with larger numbers of infants

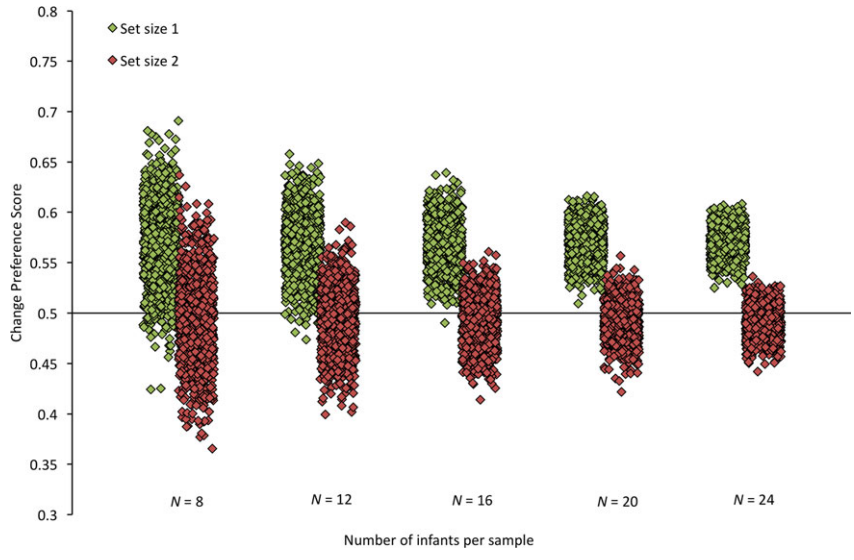

Figure 6 Mean change preference calculated from 1,000 simulated experiences with samples (selected without replacement) of 8, 12, 16, 20, or 25 infants for set size 1 (green) and set size 2 (red) trials. Each individual diamond represents the mean of one subsample. The *x*-axis crosses the *y*-axis at .50, which is chance responding. Any diamond above .50 indicates a preference for (i.e., longer looking at) the changing stream, and any diamond below .50 indicates a preference for (i.e., longer looking at) the nonchanging stream.

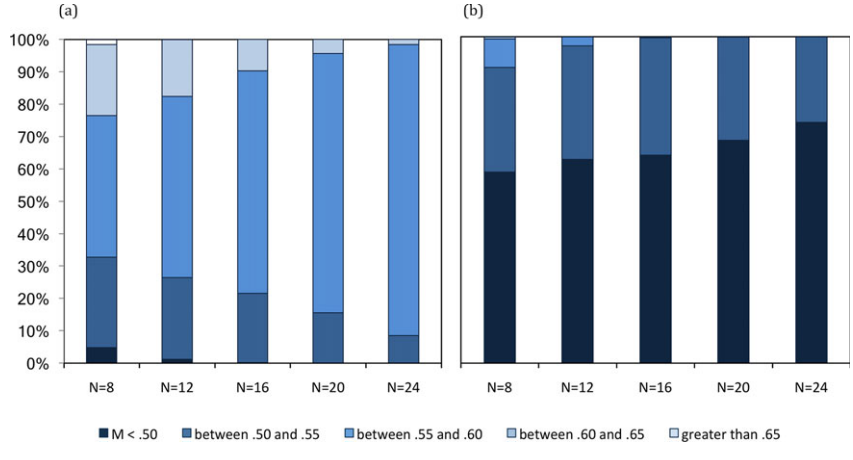

Figure 7 The distribution of change preference scores for set size 1 (a) and set size 2 (b) trials for the 1,000 simulated experiments with subsamples of 8, 12, 16, 20, and 24 infants from the sample reported in Kwon et al. (2014). The height of each region in the columns represents the proportion of samples with mean change preference scores within a given range (e.g., <.50, between .50 and .55, and so on).

—for both the positive outcome (set size 1) *and* the negative outcome (set size 2). This is further illustrated in Figure 7, which presents the proportion of subsamples with means in particular ranges. Note that many subsamples of 16 or fewer infants yielded

mean change preferences scores that are below .55 at set size 1 (33% of the subsamples of eight infants, 26% of the subsamples of 12, and 22% of the subsamples of 16 infants) and above .55 at set size 2 (9% of the subsamples of 8 infants and 3% of the subsamples of 12 infants). If one of these subsamples had been the sample reported in the paper, we may have concluded that infants failed to detect a change at set size 1 and/or did detect a change at set size 2. Of course, this may be a legitimate conclusion for individual infants, but what is clear from the subsamples that included at least 20 infants, the group of infants as a whole did not show this pattern of responding.

The issue is further illustrated by the distribution of *p*-values when comparing the mean change preference scores for each subsample to chance (.50). This distribution is presented in Figures 8 and 9. In Figure 8, it is clear that the distribution of the *p*-values changes as the sample size increases. In Figure 9, it is clear that this is particularly true for set size 1. Interestingly, as the sample size decreases, the proportion of *p*-values that are <.05 *decreases* for set size 1 and *increases* for set size 2, confirming that low-powered studies both decrease true positives and increase false positives.

The take-home message is clear. If our study had included one of these randomly selected sets of 8, 12, or 16 infants in the actual experiment (rather than the entire group of infants), there is a nontrivial chance we would have concluded that we have no evidence of change detection at set size 1 or that we have evidence of change detection at set size 2. In other words, by having a sample size that is too small, we increase the

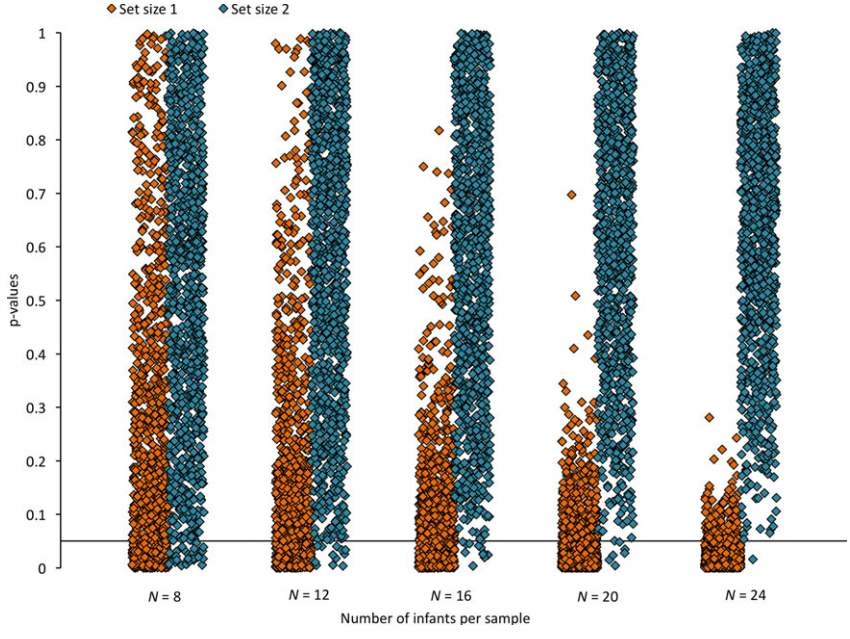

Figure 8 The *p*-values associated with one-sample comparisons of each change preference score to chance (.50) for each of the 1,000 random samples (without replacement) of 8, 12, 16, 20, or 25 infants. The *p*-values for set size 1 trials are given in orange, and the *p*-values for set size 2 trials are given in teal. Each individual diamond represents the *p*-value for the *t*-test comparison for a single sample. The *x*-axis crosses the *y*-axis at .05; diamonds between zero and .05 represent $p < .05$ and would be considered significant, and diamonds above .05 represent $p > .05$ and would be considered nonsignificant.

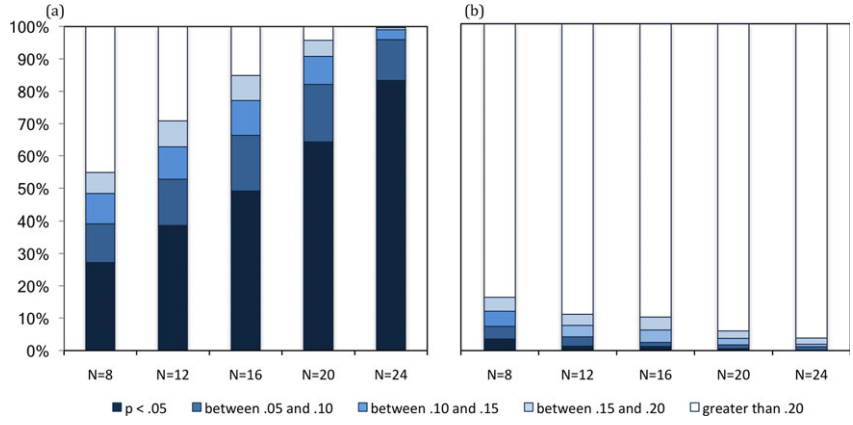

Figure 9 The distribution of *p*-values when comparing to chance (.50) the change preference scores on set size 1 (a) and set size 2 (b) trials for the 1,000 random samples selected of 8, 12, 16, 20, and 24 infants from the sample reported in Kwon et al. (2014). The height of each region in the columns represents the proportion of *t*-tests with *p*-values within a given range (e.g., <.05, between .05 and .10, and so on).

likelihood of a study with ambiguous or false results. Considering again the problem of researchers whose abilities to recruit infants is limited. The data presented here suggest that collecting the data from a single study with 24 infants would be more likely to yield interpretable results than collecting the data from two studies with 12 infants each.

Examples 2 and 3

To confirm that these observations were not specific to a single study, sample, procedure, and/or dependent measure, we took this same approach in evaluating two unpublished data sets from my laboratory. These second and third samples are two experiments that were conducted using the same audiovisual stimuli in habituation procedures using the *switch* design. In each experiment, 32 ten-month-old infants were habituated to two multimodal dynamic events; infants saw these events until they reached a habituation criterion of a 50% decrement of their initial looking to the events. Following habituation, infants were tested with one of the two familiar events and a switched event, in which the features of the two events were combined in a new way. The two experiments only differed in the ways the features were combined; in the first Experiment infants' attention to the *Action* was observed and in the second infants' attention to the *Sound* was observed. Therefore, in the following discussion I will refer to them as Experiment Action and Experiment Sound.

Mean looking times for test events in Experiment Action are presented in Figure 10; the results from the whole sample are presented in Figure 10a, and the results of two nonoverlapping subsamples of 16 infants are presented in Figures 10b and 10c. Statistical analyses of the whole sample revealed that this groups of infants significantly increased their looking to the novel stimulus compared to the familiar, $t(31) = 3.36$, $p = .002$, $d = .59$. The Bayes factor indicated that the data were 17.1 times more likely to come from a true difference than to come from a null effect, so we can be quite confident that these data reflect a real increase in looking. The first of the two subsamples

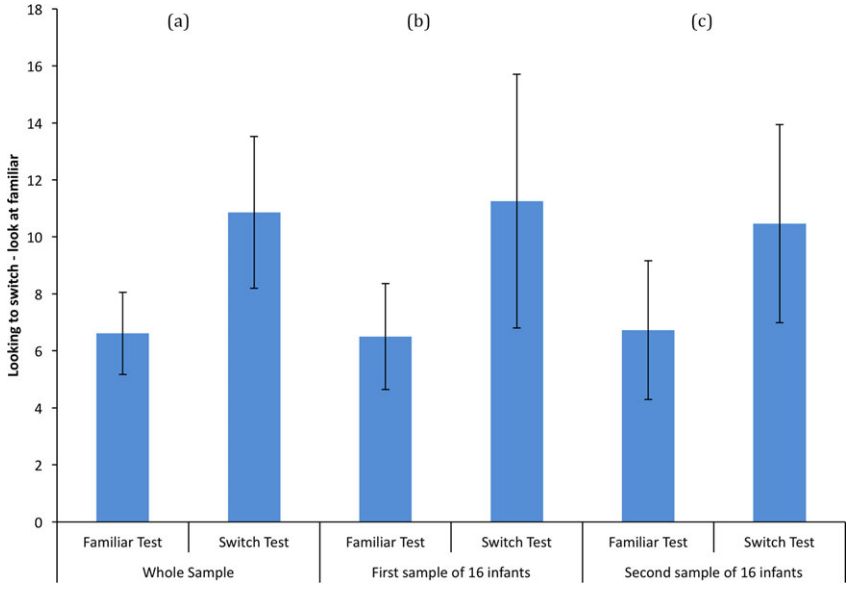

Figure 10 Mean looking time (in seconds) to the familiar and switch tests in Experiment Action for the whole sample (a), and two nonoverlapping samples of 16 infants (b and c). Error bars represent 95% confidence intervals around the mean.

of 16 infants showed ambiguous results; their increase to the novel event was marginally significant, $t(15) = 2.09$, $p = .054$, $d = .522$. The second subsample of 16 infants significantly increased their looking to the switch test, $t(15) = 3.19$, $p = .006$, $d = .797$.

The distributions of the mean difference scores (looking at switch – looking at familiar) obtained from the 1,000 simulated experiments with subsamples of 8, 12, 16, 20, and 24 infants are presented in Figure 11a. Again, these distributions differed as a function of sample size. The number of subsamples with means near the center increased as the number of infants in the subsamples increase. This is even clearer when looking at the distributions in Figure 11c. For the larger subsamples of infants, the mean difference between the switch and the test item is rarely <2 sec (only 1% of the subsamples of 20 infants), whereas this occurred with some frequency with smaller subsamples (e.g., 14% of the subsamples of eight infants had difference scores <2 sec). Moreover, over 99% of the subsamples of 24 infants and 97% of the subsamples of 20 infants had difference scores of >4 sec.

The results of the *t*-tests comparing infants' responding to the two tests confirmed the impression that larger sample sizes consistently more accurately represent the group of infants as a whole. The distribution of *p*-values from those *t*-tests is presented in two forms in Figure 12. In Figure 12a, it is clear that although the *p*-values for larger subsamples rarely were over $p = .05$, there were many instances of larger *p*-values with subsamples of fewer infants. This is confirmed in the alternative way of visualizing these distributions in Figure 12c.

Experiment Sound was similar to the Experiment Action, except that the particular combination of features used differed. In this case, as illustrated in Figure 13, the group as a whole failed to significantly dishabituate to the switch event relative to the

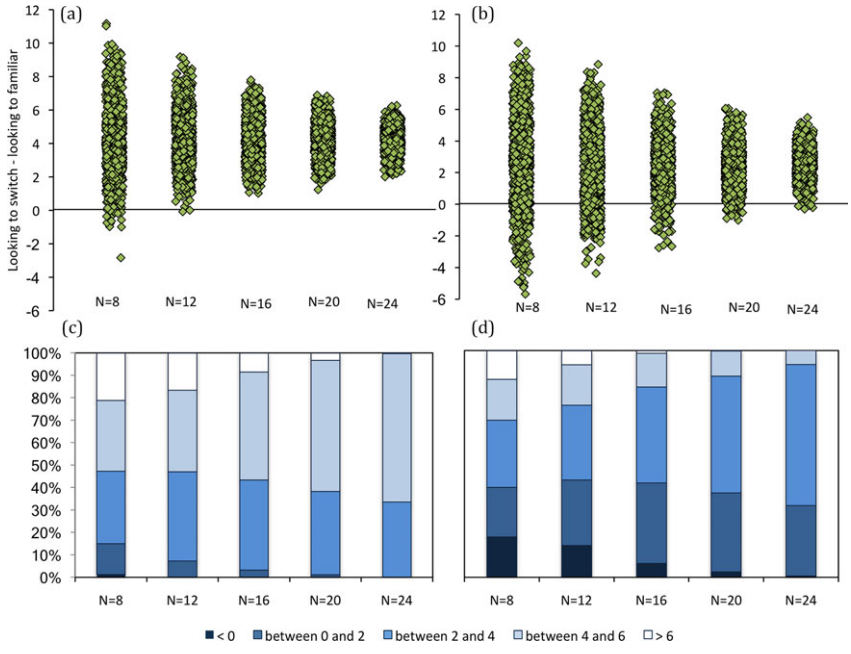

Figure 11 Distribution of difference scores in Experiment Action (a and c) and Experiment Sound (b and d). In panels (a) and (c), each diamond represents the difference score for one of the 1000 samples of each size. In panels (b) and (d), the different heights of the regions within the columns represent the proportion of the total number of samples with difference scores in a given range (e.g., <0 sec, between 0 and 2 sec, and so on).

familiar event, $t(31) = 1.46$, $p = .15$, $d = .26$. The Bayes factor analysis favored the null hypothesis by a factor of 2.0, so it is likely that either the null hypothesis is true or the actual effect is quite small.

Two simulated experiments with nonoverlapping subsamples of 16 infants showed different patterns. The first subsamples of 16 infants significantly dishabituated to the novel item, $t(15) = 2.21$, $p = .018$, $d = .666$. The second subsample of 16 infants as a group failed to respond differently to the familiar and switch tests, $t(15) = .42$, $p = .68$, $d = .10$, and actually looked slightly less to the switch than to the novel test event. Once again, we would have drawn very different conclusions from these two subsamples. If we had tested only the 16 infants in the first subsample, we would have concluded that infants do dishabituate to the switch event, and they learned the association embodied by the habituation events. The null finding in the second sample of 16 infants casts doubt on this conclusion. The question is which of these findings more closely resembles the "true" effect?

The distribution of dishabituation scores to the switch from the simulations is revealing. Figure 11b shows how the distribution of difference scores changes with sample size. Note the differences in the distributions for Experiment Action and Experiment Sound. Although the two nonoverlapping samples depicted in Figure 13 suggest that there may be some significant dishabituation, the distributions of the simulations presented in Figure 11d reveal that this significant dishabituation was rare. Moreover, large differences in looking to the switch and familiar test trials were more frequent in the subsamples of fewer infants than in the subsamples of more infants.

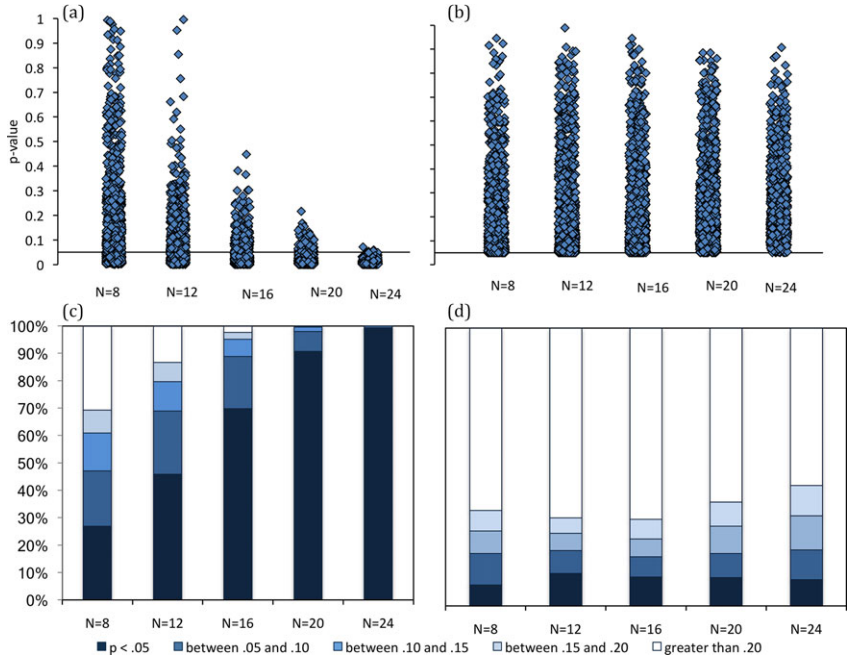

Figure 12 The distribution of *p*-values for the *t*-tests comparing mean looking to the familiar and switch test in each subsample of 8, 12, 16, 20, and 24 infants for Experiment Action and Experiment Sound. In panels (a) and (b), each diamond represents the *p*-value for a single sample. In panels (c) and (d), the height of the regions within the column represents the proportion of the 1,000 subsamples with *p*-values within a particular range (e.g., <.05, between .05 and .10, and so on).

This observation is corroborated by the distributions of *t*-values and the *p*-values. The distribution of *p*-values is presented in Figure 12. The effect of the size of the subsample is much subtler here than in the previous experiment. Across all subsample sizes, few subsamples yielded significant results (8% of subsamples of eight infants, 12% of subsamples of 12 infants, and 10% of subsamples of 16, 20, and 24 infants. Regardless of subsample size, more than 55% of the subsamples yielded *p*-values >.20.

CONCLUSIONS

This paper has described the state of the field with respect to sample size in studies of infant looking time. The conclusions should be considered in light of several facts. First, the target sample sizes were based on an evaluation of all the published studies recording infants' looking time in the top journals that publish much of this work. Second, the conclusions about specific sample sizes are only directly applicable to infant looking-time studies—work that involves more trials, different dependent measures, automatic observation, etc., may yield different amounts of variability and effect sizes. Thus, although the conclusions about the importance of power and sample size are generally true, the specific sample and effect sizes reported here are representative only of this subset of the literature. Third, as is true for many areas of science, samples sizes appear to have been determined in large part by rule of thumb and convention.

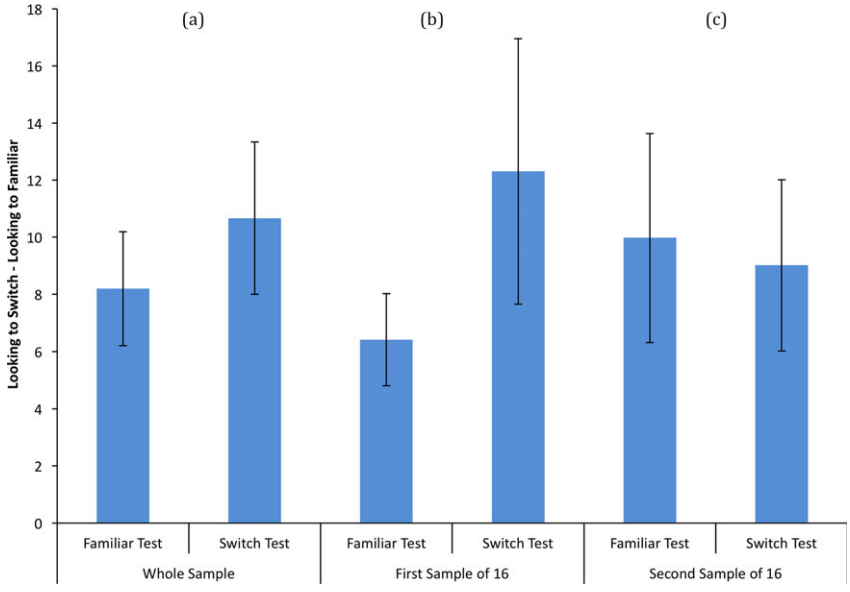

Figure 13 Mean looking time (in seconds) to the familiar and switch tests of the whole sample of Experiment Sound (a), and two nonoverlapping samples of 16 infants (b and c) drawn from the whole sample. Error bars represent 95% confidence intervals around the mean.

Unfortunately, because data collection with infants is difficult, expensive, and time-consuming, this has meant that many reported studies have relatively low power.

The problem of low power and relatively small sample sizes is not unique to this area of research. There has been significant debate about the effect of sample size on research, whether increasing sample sizes without making other changes will be effective, and about how best to calculate power. However, the simulations shown here are revealing. They show how the studies run in my laboratory would have differed if we had collected data from subsamples of different sizes. This is an important demonstration. For each of these studies, we could have had a target of only 16 infants per cell, and we would have stopped collecting data after just the first 16 infants tested. Note, moreover, that in the two positive cases, the effect sizes were medium—between .5 and .6. Thus, when smaller subsamples showed significant differences, the effect sizes observed clearly overestimated the true effect size. Thus, using smaller sample sizes, the outcomes in many cases would have been different from the ones we ultimately observed. To be clear, the subsamples did yield the same results as was observed from the full sample more frequently than any other outcome—and most outcomes that differed only differed by a small amount. However, in these examples smaller samples would have increased the likelihood of observing ambiguous results—or even results that led to a very clear but different conclusion than the conclusion drawn from the full sample. Thus, by adopting the convention of using relatively small sample sizes in our work, we as a field are increasing the chances that the outcome of any single study is difficult to interpret or not representative of the most likely outcome from the study.

So what do we do? One obvious solution is to increase sample sizes. Of course, given how difficult it is to identify, recruit, and test infant research participants, it is unlikely that all infant looking-time studies will involve very large samples. Moreover,

it is not clear that all studies should have very large samples Indeed, it has been argued that for some areas of science, sample sizes should be determined by considering cost efficiency in addition to power (Bacchetti, Simon, Mcculloch, & Segal, 2009; Miller & Ulrich, 2016). It is not clear, however, that radical changes need to be made to the standard sample size in infant looking-time studies to address the power problem, in general. Figure 14 illustrates the change in the power to observe effects of different sizes (these are Cohen's *d* for paired comparison two-tailed *t*-tests, *alpha* = .05). Note that although samples sizes of 8 provide insufficient power to detect any of the effect sizes depicted, samples of 20 or 24 infants will provide sufficient power for effect sizes of approximately .60 and higher. Thus, it is not clear that the field needs to adopt a convention of testing hundreds of infants per cell, but we may have more consistent results if we relied on sample sizes of 20–32 per cell, rather than 12–24 infants per cell.

Moreover, the cost of consistently relying on small sample sizes and low-powered studies may be too high. For researchers who have difficulty collecting data from infants, spending the time to collect the data for a single study with 24 or 32 infants is likely to consistently yield more interpretable, replicable results than spending that same time collecting the data for two studies each with 12 or 16 infants—and collecting the data from 48 infants in one study will clearly yield more precise results than four samples of 12. Given the data presented here, it seems likely that by having larger target sample sizes, researchers may actually end up with fewer file drawer studies, and their efforts may yield more published products in the long run.

Increasing sample sizes is not the only solution, however, and other approaches to this general problem may be fruitful. A compromise might be found in the use of careful sequential hypothesis analyses (Lakens, 2014). In this approach, researchers identify a target final sample size (e.g., 48 infants), but conduct a test of their hypothesis at

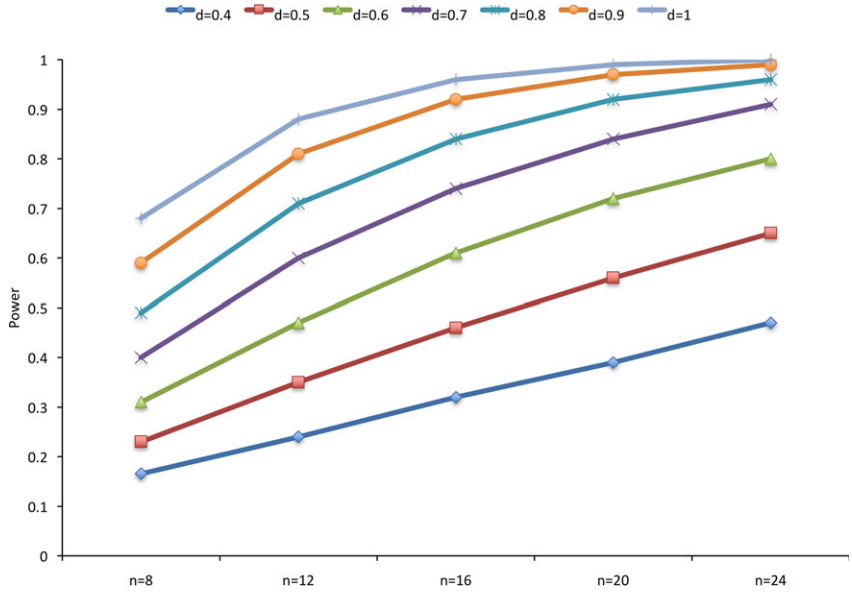

Figure 14 Demonstration of the effect of sample size on the power to detect a range of effect sizes. The dashed line corresponds to the .80 power level.

some interim point in data collection (e.g., 24 infants). By adopting a clear stopping rule and an approach that adjusts for the increase in type 1 error, this method may allow researchers to efficiently conduct high-powered studies even with difficult samples such as infant subjects. Importantly, these designs depend on *planned* interim data analysis, as flexible, undisclosed interim data analysis and stopping rules may lead to an increase in the publication of false positive results (see Simmons, Nelson, & Simonsohn, 2011). The point is that by adopting ethical, transparent means of "peeking" at the data (see Sagarin, Ambler, & Lee, 2014, for a suggestion), researchers may more effectively and efficiently obtain the samples sizes required to have sufficient power to draw strong conclusions from their results.

There are other solution to the problem of small sample sizes, as described by Tressoldi and Giofré (2015). The use of Bayes factors instead of *p*-values has been increasingly described as one solution. The Bayes factor can be used to indicate the relative likelihood of the observed results arising from the null versus alternative hypotheses (Jarosz & Wiley, 2014; Rouder et al., 2009; Wagenmakers et al., 2015). It may be possible to combine this approach with sequential hypothesis testing (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015). Others have argued abandoning traditional testing altogether, with some suggesting that estimation statistics may be a reasonable alternative to traditional *t*-tests (Claridge-Chang & Assam, 2016). Others have explored the value of *p*-curving, or the evaluation of the distribution of *p*-values across a set of experiments (Lakens & Evers, 2014; Simonsohn, Nelson, & Simmons, 2014a,b). This may be one way researchers can look at the data across a collection of relatively small sample studies to assess their value. Other alternatives are to conduct resampling analyses to evaluate the replicability of an observed result, for example, using a Jackknife approach (Ang, 1998), or permutation analyses (Berry, Johnston, & Mielke, 2011; Huo, Heyvaert, Van Den Noortgate, & Onghena, 2014). For such approaches to be successful, authors, editors, and reviewers need to be open to other ways of evaluating data, and using other approaches as the basis of our conclusions about those data. However, adopting new approaches may be critical for increasing our confidence about the conclusions we can draw from any particular finding, and as a result what conclusions we draw in general about our work. Although I have focused here on a narrow slice of infant research, these issues are important across methods, measures and questions.

Future work will make specific recommendations about how to address the issue of small sample sizes in infant looking-time studies, or any subarea of infant development. Here I have illustrated an issue that we as a field need to address. In addition, by describing some of the approaches that are being considered in other areas of research, the hope is that infant researchers will expand the set of tools they use to evaluate their research to help draw the strongest conclusions from whatever sample sizes programs of research can support.

ACKNOWLEDGMENTS

Preparation of the manuscript and the research reported here were made possible by support from grant R01EY022525 awarded by the National Institutes of Health and grant BCS 0921634 awarded by the National Science Institute. I thank Katharine Graf Estes, Steve Luck, and Simine Vazire for helpful comments on drafts of this

manuscript and discussions of these issues. I also thank the students and staff of the UC Davis Infant Cognition Lab for catching typos and helping to clarify this manuscript.

REFERENCES

Althaus, N., & Plunkett, K. (2015). Timing matters: The impact of label synchrony on infant categorisation. *Cognition*, *139*, 1–9.

Ang, R. P. (1998). Use of the jackknife statistic to evaluate result replicability. *The Journal of General Psychology*, *125*, 218–228.

Bacchetti, P. (2013). Small sample size is not the real problem: Letter. *Nature Reviews. Neuroscience*, *14*, 585.

Bacchetti, P., Simon, R., Mcculloch, C. E., & Segal, M. R. (2009). Simple, defensible sample sizes based on cost efficiency – With discussion and rejoinder simple, defensible sample sizes based on cost efficiency – With discussion and rejoinder. *Biometrics*, *64*, 577–594. doi:10.1111/j.1541-0420.2008.01004.x.

Bahrick, L. E., Lickliter, R., & Castellanos, I. (2013). The development of face perception in infancy: Intersensory interference and unimodal visual facilitation. *Developmental Psychology*, *49*, 1919–1930.

Bahrick, L. E., Lickliter, R., Castellanos, I., & Todd, J. T. (2015). Intrasensory redundancy facilitates infant detection of tempo: Extending predictions of the intersensory redundancy hypothesis. *Infancy*, *20*, 1–28.

Baillargeon, R. (1987). Object permanence in 3 1/2- and 4 1/2-month-old infants. *Developmental Psychology*, *23*, 655–664.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognitive Psychology*, *20*, 191–208.

Baker, J. M., Mahamane, S. P., & Jordan, K. E. (2014). Multiple visual quantitative cues enhance discrimination of dynamic stimuli during infancy. *Journal of Experimental Child Psychology*, *122*, 21–32.

Baker, R. K., Pettigrew, T. L., & Poulin-Dubois, D. (2014). Infants' ability to associate motion paths with object kinds. *Infant Behavior and Development*, *37*, 119–129.

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*, 1069–1077.

Bakker, M., & Wicherts, J. M. (2014a). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples *t* tests: The power of alternatives and recommendations. *Psychological Methods*, *19*, 409–427.

Bakker, M., & Wicherts, J. M. (2014b). Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS One*, *9*, 1–9.

Bardi, L., Regolin, L., & Simion, F. (2014). The first time ever I saw your feet: Inversion effect in newborns' sensitivity to biological motion. *Developmental Psychology*, *50*, 986–993.

Benavides-Varela, S., & Mehler, J. (2015). Verbal positional memory in 7-month-olds. *Child Development*, *86*, 209–223.

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2011). Analysis of a trend: A permutation alternative to the F test. *Perceptual and Motor Skills*, *112*, 247–257.

Bidet-Ildei, C., Kitromilides, E., Orliaguet, J.-P., Pavlova, M., & Gentaz, E. (2014). Preference for point-light human biological motion in newborns: contribution of translational displacement. *Developmental Psychology*, *50*, 113–120.

Biro, S., Verschoor, S., Coalter, E., & Leslie, A. M. (2014). Outcome producing potential influences twelve-month-olds' interpretation of a novel action as goal-directed. *Infant Behavior and Development*, *37*, 729–738.

Bramwell, A., Bittnerjr, A., & Morrissey, S. (1992). Repeated-measures analysis: Issues and options. *International Journal of Industrial Ergonomics*, *10*, 185–197.

Brand, A., Bradley, M. T., Best, L. A., & Stoica, G. (2008). Accuracy of effect size estimates from published psychological research. *Perceptual and Motor Skills*, *106*, 645–649.

Bremner, J. G., Slater, A. M., Mason, U. C., Spring, J., & Johnson, S. P. (2013). Trajectory perception and object continuity: Effects of shape and color change on 4-month-olds' perception of object identity. *Developmental Psychology*, *49*, 1021–1026.

Brower, T. R., & Wilcox, T. (2013). Priming infants to use color in an individuation task: Does social context matter? *Infant Behavior and Development*, *36*, 349–358.

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

Cantrell, L., Boyer, T. W., Cordes, S., & Smith, L. B. (2015). Signal clarity: An account of the variability in infant quantity discrimination tasks. *Developmental Science*, *18*, 877–893.

Casasola, M., & Park, Y. (2013). Developmental changes in infant spatial categorization: When more is best and when less is enough. *Child Development*, *84*, 1004–1019.

Cashon, C. H., Ha, O. R., Allen, C. L., & Barna, A. C. (2013). A U-shaped relation between sitting ability and upright face processing in infants. *Child Development*, *84*, 802–809.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior and Organization*, *81*, 1–8.

Claridge-Chang, A., & Assam, P. N. (2016). Estimation statistics should replace significance testing. *Nature Methods*, *13*, 108–109.

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

Cohen, L. B., & Marks, K. S. (2002). How infants process addition and subtraction events. *Developmental Science*, *5*, 186–201.

Coubart, A., Izard, V., Spelke, E. S., Marie, J., & Streri, A. (2014). Dissociation between small and large numerosities in newborn infants. *Developmental Science*, *17*, 11–22.

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99.

Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods*, *118*, 115–128.

Epskamp, S., & Nuijten, M. B. (2016). statcheck: Extract statistics from articles and recompute p values. (R package version 1.2.2).

Esteve-Gibert, N., Prieto, P., & Pons, F. (2015). Nine-month-old infants are sensitive to the temporal alignment of prosodic and gesture prominences. *Infant Behavior and Development*, *38*, 126–129.

Fantz, R. L. (1958). Pattern vision in young infants. *The Psychological Record*, *8*, 43–47.

Fantz, R. L. (1963). Pattern vision in newborn infants. *Science*, *140*, 296–297. doi:10.1126/science.140.3564.296.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention familiar patterns relative to novel ones. *Science*, *146*, 668–670.

Fantz, R. L., & Nevis, S. (1967). Pattern preferences and perceptual-cognitive development in early infancy. *Merrill-Palmer Quarterly of Behavior and Development*, *13*, 77–108.

Ferry, A. L., Hespos, S. J., & Gentner, D. (2015). Prelinguistic relational concepts: Investigating analogical processing in infants. *Child Development*, *86*, 1386–1405.

Flom, R., Janis, R. B., Garcia, D. J., & Kirwan, C. B. (2014). The effects of exposure to dynamic expressions of affect on 5-month-olds' memory. *Infant Behavior and Development*, *37*, 752–759.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One*, *9*. doi:10.1371/journal.pone.0109019.

Frick, A., & Möhring, W. (2013). Mental object rotation and motor development in 8- and 10-month-old infants. *Journal of Experimental Child Psychology*, *115*, 708–720.

Frick, A., & Wang, S. H. (2014). Mental spatial transformations in 14- and 16-month-old infants: Effects of action and observational experience. *Child Development*, *85*, 278–293.

Friston, K. (2012). Ten ironic rules for non-statistical reviewers. *NeuroImage*, *61*, 1300–1310.

Gazes, R. P., Hampton, R. R., & Lourenco, S. F. (2017). Transitive inference of social dominance by human infants. *Developmental Science*, *20*, e12367. doi:10.1111/desc.12367.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words?: Statistical segmentation and word learning. *Psychological Science*, *18*, 254–260.

Graf Estes, K., & Hay, J. F. (2015). Flexibility in bilingual infants' word learning. *Child Development*, *86*, 1371–1385.

Gustafsson, E., Brisson, J., Beaulieu, C., Mainville, M., Mailloux, D., & Sirois, S. (2015). How do infants recognize joint attention? *Infant Behavior and Development*, *40*, 64–72.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, *12*, 179–185.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*, 557–559.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*, 1–15.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*, *21*, 299–332.

Henderson, A. M. E., & Scott, J. C. (2015). She called that thing a mido, but should you call it a mido too? Linguistic experience influences infants' expectations of conventionality. *Frontiers in Psychology*, *6*, 1–11.

Hernik, M., & Csibra, G. (2015). Infants learn enduring functions of novel tools from action demonstrations. *Journal of Experimental Child Psychology*, *130*, 176–192.

Heron-Delaney, M., Quinn, P. C., Lee, K., Slater, A. M., & Pascalis, O. (2013). Nine-month-old infants prefer unattractive bodies over attractive bodies. *Journal of Experimental Child Psychology*, *115*, 30–41.

Hillairet de Boisferon, A., Dupierrix, E., Quinn, P. C., Lœvenbruck, H., Lewkowicz, D. J., Lee, K., & Pascalis, O. (2015). Perception of multisensory gender coherence in 6- and 9-month-old infants. *Infancy*, *20*, 661–674.

Hillairet de Boisferon, A., Uttley, L., Quinn, P. C., Lee, K., & Pascalis, O. (2014). Female face preference in 4-month-olds: The importance of hairline. *Infant Behavior and Development*, *37*, 676–681.

Hock, A., Kangas, A., Zieber, N., & Bhatt, R. S. (2015). The development of sex category representation in infancy: Matching of faces and bodies. *Developmental Psychology*, *51*, 346–352.

Huo, M., Heyvaert, M., Van Den Noortgate, W., & Onghena, P. (2014). Permutation tests in the educational and behavioral sciences: A systematic review. *Methodology*, *10*, 43–59.

Imura, T., Masuda, T., Shirai, N., & Wada, Y. (2015). Eleven-month-old infants infer differences in the hardness of object surfaces from observation of penetration events. *Frontiers in Psychology*, *6*, 1–7.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*, 2–9.

Kampis, D., Somogyi, E., Itakura, S., & Király, I. (2013). Do infants bind mental states to agents? *Cognition*, *129*, 232–240.

Kavsek, M., & Marks, E. (2015). Infants perceive three-dimensional illusory contours as occluding surfaces. *Child Development*, *86*, 1865–1876.

Krzywinski, M., & Altman, N. (2013). Points of significance: Power and sample size. *Nature Methods*, *10*, 1139–1140.

Kwon, M. K., Luck, S. J., & Oakes, L. M. (2014). Visual short-term memory for complex objects in 6- and 8-month-old infants. *Child Development*, *85*, 564–577.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12.

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

Lakens, D. (2015). What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *The Quarterly Journal of Experimental Psychology*, *68*, 829–832.

Lakens, D., & Evers, E. R. K. (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*, 278–292.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107–112.

Lee, V., Cheal, J. L., & Rutherford, M. D. (2015). Categorical perception along the happy-angry and happy-sad continua in the first year of life. *Infant Behavior and Development*, *40*, 95–102.

Lewkowicz, D. J. (2013). Development of ordinal sequence perception in infancy. *Developmental Science*, *16*, 352–364.

Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: Its emergence and the role of experience. *Journal of Experimental Child Psychology*, *130*, 147–162.

Libertus, K., & Needham, A. (2014). Face preference in infancy and its relation to motor activity. *International Journal of Behavioral Development*, *38*, 529–538.

Liu, S., Xiao, W. S., Xiao, N. G., Quinn, P. C., Zhang, Y., Chen, H., … & Lee, K. (2015). Development of visual preference for own- versus other-race faces in infancy. *Developmental Psychology*, *51*, 500–511.

Longhi, E., Senna, I., Bolognini, N., Bulf, H., Tagliabue, P., Macchi Cassia, V., & Turati, C. (2015). Discrimination of biomechanically possible and impossible hand movements at birth. *Child Development*, *86*, 632–641.

Loucks, J., & Sommerville, J. A. (2013). Attending to what matters: Flexibility in adults' and infants' action perception. *Journal of Experimental Child Psychology*, *116*, 856–872.

Mackenzie, H. K., Graham, S. A., Curtin, S., & Archer, S. L. (2014). The flexibility of 12-month-olds' preferences for phonologically appropriate object labels. *Developmental Psychology*, *50*, 422–430.

Matatyaho-Bullaro, D. J., Gogate, L., Mason, Z., Cadavid, S., & Abdel-Mottaleb, M. (2014). Type of object motion facilitates word mapping by preverbal infants. *Journal of Experimental Child Psychology*, *118*, 27–40.

May, L., & Werker, J. F. (2014). Can a click be a word?: Infants' learning of non-native words. *Infancy*, *19*, 281–300.

McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science*, *15*, 776–781.

McShane, B. B., & Böckenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*, 612–625.

Miller, J., & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science*, *11*, 664–691.

Moher, M., & Feigenson, L. (2013). Factors influencing infants' ability to update object representations in memory. *Cognitive Development*, *28*, 272–289.

Motulsky, H. J. (2015). Common misconceptions about data analysis and statistics. *British Journal of Pharmacology*, *172*, 2126–2132.

Muentener, P., & Carey, S. (2010). Infants' causal representations of state change events. *Cognitive Psychology*, *61*, 63–86.

Muthén, L. K., & Muthén, B. O. (2009). How to use a monte carlo study to decide on sample size and determine how to use a monte carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 599–620.

Novack, M. A., Henderson, A. M. E., & Woodward, A. L. (2013). Twelve-month-old infants generalize novel signed labels, but not preferences across individuals. *Journal of Cognition and Development*, *8372*, 1–12.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226. doi:10.3758/s13428-015-0664-2.

Oakes, L. M., Baumgartner, H. A., Barrett, F. S., Messenger, I. M., & Luck, S. J. (2013). Developmental changes in visual short-term memory in infancy: Evidence from eye-tracking. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00697.

Oakes, L. M., & Kovack-Lesh, K. A. (2013). Infants' visual recognition memory for a series of categorically related items. *Journal of Cognition and Development*, *14*, 63–86.

Oakes, L. M., & Ribar, R. J. (2005). A comparison of infants' categorization in paired and successive presentation familiarization tasks. *Infancy*, *7*, 85–98.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255–258.

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

Otsuka, Y., Motoyoshi, I., Hill, H. C., Kobayashi, M., Kanazawa, S., & Yamaguchi, M. K. (2013). Eye contrast polarity is critical for face recognition by infants. *Journal of Experimental Child Psychology*, *115*, 598–606.

Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound-shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, *114*, 173–186.

Park, Y., & Casasola, M. (2015). Plain or decorated? Object visual features matter in infant spatial categorization. *Journal of Experimental Child Psychology*, *140*, 105–119.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*, 531–536.

Peltola, M. J., Leppänen, J. M., Palokangas, T., & Hietanen, J. K. (2008). Fearful faces modulate looking duration and attention disengagement in 7-month-old infants. *Developmental Science*, *11*, 60–68.

Perone, S., & Spencer, J. P. (2014). The co-development of looking dynamics and discrimination performance. *Developmental Psychology*, *50*, 837–852.

Pruden, S. M., Roseberry, S., Göksun, T., Hirsh-Pasek, K., & Golinkoff, R. M. (2013). Infant categorization of path relations during dynamic events. *Child Development*, *84*, 331–345.

Quinlan, P. T. (2013). Misuse of power: In defence of small-scale science. *Nature Reviews Neuroscience*, *14*, 585.

Quinn, P. C., & Liben, L. S. (2014). A sex difference in mental rotation in infants: Convergent evidence. *Infancy*, *19*, 103–116.
Rigney, J., & Wang, S. (2013). Delineating the boundaries of infants' spatial categories: The case of containment. *Journal of Cognition and Development*, *16*, 420–441. doi:10.1080/15248372.2013.848868.
Robson, S. J., Lee, V., Kuhlmeier, V. A., & Rutherford, M. D. (2014). Infants use contextual contingency to guide their interpretation of others' goal-directed behavior. *Cognitive Development*, *31*, 69–78.
Ross-Sheehy, S., Oakes, L. M., & Luck, S. J. (2003). The development of visual short-term memory capacity in infants. *Child Development*, *74*, 1807–1822.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.
Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician*, *40*, 313–315.
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928. doi:10.1126/science.274.5294.1926.
Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *9*, 293–304.
Sanefuji, W., Wada, K., Yamamoto, T., Mohri, I., & Taniike, M. (2014). Development of preference for conspecific faces in human infants. *Developmental Psychology*, *50*, 979–985.
Sato, K., Masuda, T., Wada, Y., Shirai, N., Kanazawa, S., & Yamaguchi, M. K. (2013). Infants' perception of curved illusory contour with motion. *Infant Behavior and Development*, *36*, 557–563.
Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*, 597–599.
Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*. doi:10.1037/met0000061.
Schweizer, G., & Furley, P. (2016). Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychology of Sport and Exercise*, *23*, 114–122.
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.
Simon, T. J., Hespos, S. J., & Rochat, P. (1995). Do infants understand simple arithmetic? A replication of Wynn (1992). *Cognitive Development*, *10*, 253–269.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547.
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). p-curve and effect size: Correcting for publication bias using only significant results. *Psychological Science*, *9*, 666–681.
Skerry, A. E., & Spelke, E. S. (2014). Preverbal infants identify emotional reactions that are incongruent with goal outcomes. *Cognition*, *130*, 204–216.
Slone, L. K., & Johnson, S. P. (2015). Infants' statistical learning: 2- and 5-month-olds' segmentation of continuous visual sequences. *Journal of Experimental Child Psychology*, *133*, 47–56.
Soley, G., & Sebastián-Gallés, N. (2015). Infants prefer tunes previously introduced by speakers of their native language. *Child Development*, *86*, 1685–1692.
Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*, 605–632.
Starr, A. B., Libertus, M. E., & Brannon, E. M. (2013). Infants show ratio-dependent number discrimination regardless of set size. *Infancy*, *18*, 927–941.
Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71.
Takashima, M., Kanazawa, S., Yamaguchi, M. K., & Shiina, K. (2014). The homogeneity effect on figure/ground perception in infancy. *Infant Behavior and Development*, *37*, 57–65.
Tham, D. S. Y., Bremner, J. G., & Hay, D. (2015). In infancy the timing of emergence of the other-race effect is dependent on face gender. *Infant Behavior and Development*, *40*, 131–138.
Träuble, B., & Bätz, J. (2014). Shared function knowledge: Infants' attention to function information in communicative contexts. *Journal of Experimental Child Psychology*, *124*, 67–77.
Tressoldi, P. E., & Giofré, D. (2015). The pervasive avoidance of prospective statistical power: Major consequences and practical solutions. *Frontiers in Psychology*, *6*, 1–4.

Tsuruhara, A., Corrow, S., Kanazawa, S., Yamaguchi, M. K., & Yonas, A. (2014). Infants' ability to respond to depth from the retinal size of human faces: Comparing monocular and binocular preferential-looking. *Infant Behavior and Development*, *37*, 562–570.

Turati, C., Gava, L., Valenza, E., & Ghirardi, V. (2013). Number versus extent in newborns' spontaneous preference for collections of dots. *Cognitive Development*, *28*, 10–20.

Ulrich, R., & Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General*, *144*, 1137–1145.

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, *23*, 87–102. doi:10.3758/s13423-015-0892-6.

Vankov, I., Bowers, J., & Munafò, M. R. (2014). On the persistence of low power in psychological science. *The Quarterly Journal of Experimental Psychology*, *67*, 1037–1040.

Vukatana, E., Graham, S. A., Curtin, S., & Zepeda, M. S. (2015). One is not enough: Multiple exemplars facilitate infants' generalizations of novel properties. *Infancy*, *20*, 548–575.

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M., Matzke, D., … & Morey, R. (2015). A power fallacy. *Behavior Research Methods*, *47*, 913–917.

Woods, R. J., & Wilcox, T. (2013). Posture support improves object individuation in infants. *Developmental Psychology*, *49*, 1413–1424.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*, 749–750.

Yamashita, W., Kanazawa, S., & Yamaguchi, M. K. (2014). Tolerance of geometric distortions in infant's face recognition. *Infant Behavior and Development*, *37*, 16–20.

Young-Browne, G., Rosenfeld, H. M., & Horowitz, F. D. (1977). Infant discrimination of facial expressions. *Child Development*, *48*, 555–562.

Zieber, N., Kangas, A., Hock, A., & Bhatt, R. S. (2014). The development of intermodal emotion perception from bodies and voices. *Journal of Experimental Child Psychology*, *126*, 68–79.

Zieber, N., Kangas, A., Hock, A., & Bhatt, R. S. (2015). Body structure perception in infancy. *Infancy*, *20*, 1–17.