Abstract

Among infant researchers there is growing concern regarding the widespread practice of undertaking studies that have small sample sizes and employ tests with low statistical power (to detect a wide range of possible effects). For many researchers, issues of confidence may be partially resolved by relying on replications. Here, we bring further evidence that the classical logic of confirmation, according to which the result of a replication study confirms the original finding when it reaches statistical significance, could be usefully abandoned. With real examples taken from the infant literature and Monte Carlo simulations, we show that a very wide range of possible replication results would in a formal statistical sense constitute confirmation as they can be explained simply due to sampling error. Thus, often no useful conclusion can be derived from a single or small number of replication studies. We suggest that, in order to accumulate and generate new knowledge, the dichotomous view of replication as confirmatory/disconfirmatory can be replaced by an approach that emphasizes the estimation of effect sizes via meta-analysis. Moreover, we discuss possible solutions for reducing problems affecting the validity of conclusions drawn from meta-analyses in infant research.

Keywords: Replication; Meta-analysis; Sampling error; Prediction Interval; Infancy

1. Introduction

Over the past 50 years, research on infant cognition has made considerable progresses. From the introduction of important methodological innovations such as the violation-of-expectation method to the rapidly growing literature on various aspects of reasoning, infant research is perceived to be in great shape nowadays, and researchers are certainly hoping for many more important discoveries to come in the next 50 years.

Take for instance the domain of *social cognition*. After having unveiled complex psychological expectations in infants (e.g., the understanding that others hold beliefs that sometimes can be false), researchers began to investigate whether the infant mind is equipped with early-emerging sociomoral reasoning abilities (e.g., Hamlin, Wynn, & Bloom, 2007; Margoni, Baillargeon, & Surian, 2018; Surian & Margoni, 2020; for a review see Ting, Dawkins, Stavans, & Baillargeon, 2019). A number of studies indeed reported that infants (a) expect people to approach helping rather than harming/hindering agents, and (b) prefer prosocial agents (those who help or distribute resources fairly) over antisocial agents (those who hinder or behave unfairly; Baillargeon et al., 2015). Finding evidence of psychological reasoning in infancy was surprising. Finding evidence of an early-emerging moral sense was, if possible, even more surprising.

However, before concluding that humans really develop a basic moral sense early in life, researchers—as it is common in science—need to gain a substantial degree of confidence in the reliability of the original results. To this end, with respect to the literature on infant socio-moral reasoning, a number of laboratories attempted to replicate the initial findings of Hamlin et al. (2007) and Hamlin and Wynn (2011), which reported evidence that preverbal infants prefer prosocial agents to antisocial agents. Out of the 26 studies retrieved in a recent meta-analysis, six concluded that the original finding has not been confirmed (see Margoni & Surian, 2018). How should we interpret these disconfirmations? What are they due to? Likewise, what do we require of a replication such that it can be accepted as being confirmatory?

It has been argued that, as a whole, research in psychology suffers because of a *crisis of confidence* (Pashler & Wagenmakers, 2012). A symptom of this is often identified in the extremely low rates of replication: According to Makel, Plucker, and Hegarty (2012), a replication rate of scarcely 1% resulted from a random selection of psychological studies published in top-tier journals. Still worrying, the Open Science Collaboration (2015) selected 100 experimental and correlational studies to replicate and found a dramatic drop in the statistical significance rate from 97% to 36%. In addition, and more interestingly, they found that the effects of the replication studies were likely to be half the magnitude of the effects of the original studies. However, some commentators suggest that this should be understood in the context of many of the replications being inexact, although Anderson et al. (2016) disputed some of the analyses (see Gilbert, King, Pettigrew, & Wilson, 2016). A Bayesian re-analysis led Etz and Vandekerckhove (2016) to conclude that the majority of the studies (original and replication) failed to offer strong evidence for either the null or alternative hypothesis.

Questionable research practices such as not reporting all of the dependent measures of a study in the paper or deciding whether to exclude data only after looking at the impact of doing so on the results are also believed to have contributed to this so-called ‘crisis’ (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). But beyond these issues, particularly relevant to infant research is the widespread practice to run studies with small sample sizes (Oakes, 2017). In this context, in order to regain confidence, many suggest the community should conduct and publish more replication studies. However, whereas replications are potentially useful in any experimental research field, they are sometimes invoked uncritically as a remedy. Especially among researchers that are not necessarily statisticians, it is often unclear how to properly interpret the results of a replication study, or interpretations are based on approximate, wrong or no answers to fundamental questions such as:

"What does it mean to *confirm* a study?"

"How do we know when a study has been *confirmed*?"

"What can we conclude from the fact that a study has *not* been confirmed?"

"Is it useful to interpret replications as simply indicating the presence or absence of an effect?"

In this paper we explore the logic of replication and demonstrate that it can be extremely inefficient. We focus on one particular source of error, the sampling error, and show that this error alone, which is unavoidably present in any inferential analysis of empirical data, can account for a wide range of possible results in a given replication attempt (Patil, Peng, & Leek, 2016). This is true especially when the statistical test of the original study is under-powered (for a given effect size or range of effect sizes) and therefore the study possibly reports inaccurate estimates (although this does not imply that estimates will be perfectly accurate when the power is high). Consequently, a surprisingly broad range of results can *actually* constitute confirmation. In other words, it is easy to confirm a vague result. Or, in other words, it is surprisingly difficult to disconfirm a vague result, which is problematic given falsifiability is one the main pillars of any scientific enterprise (LeBel, Berger, Campbell, & Loving, 2017; Popper, 1934/1959). We consequently argue that single replication studies are of extremely limited practical value in evaluating the original findings and accumulating knowledge.

We make two main practical suggestions for infant researchers. First, as many have already recommended, original studies need to have statistical tests with power curves[1] that have high points for many of the theoretically interesting effects (e.g., via larger sample sizes, reduced measurement error or better experimental design; see Maxwell, 2004). Second, single replication studies should not be used to confirm or reject original studies in the traditional dichotomous fashion (Wilson, Harris, & Wixted, 2020). Instead, they can be usefully treated as data that, if pooled with original findings and other available findings, would translate into improved estimates of the true effect size. Experimental results should yield refinements of the effect along a continuous dimension, rather than on a dichotomous scale. Of course, in many contexts small effects may not be very interesting although this is not universally the case (Lakens & Evers, 2014). However, we see this as a different debate. We recommend the adoption of a meta-analytic mindset (Schmidt, 1992, 1996; Tsuji, Bergmann, & Cristia, 2014). At the same time, we address some specific problems arising when using meta-analysis (such as the problems related to publication bias; e.g., Kraemer, Gardner, Brooks, & Yesavage, 1998).

Thus, first we argue that no single or small number of replication studies should be used to confirm or disconfirm in a dichotomous fashion an original finding because sampling error can often account for their results. We illustrate this by calculating prediction intervals though we are not suggesting that researchers should use them to assess findings in replication studies. Indeed, as observed by Morey and Lakens (2016), interval reasoning is dichotomous too. Prediction intervals are employed here to reveal that often studies can be easy to replicate but

[1] A power curve is a way to represent the relationship between a specific inferential test in the context of a *specific* study and a range of possible effect sizes since even if we presume a particular effect size the truth of this presumption is unknown. Such curves graphically show that larger effects are easier to detect. Note that a different experimental design will have a different associated power curve (for a more detailed discussion see, e.g., Krzywinski & Altman, 2013; Morey & Lakens, 2016).

because this encompasses a wide range of effects, sometimes in both directions, this shows that a dichotomous reasoning would not be useful in the context of replication. Second, we argue that a meta-analytic approach can be fruitfully employed in order to accumulate knowledge. We suggest that, instead of trying to decide whether an effect exists or does not exist by looking at whether a replication study 'succeeded' or 'failed', it can be more effective to collect evidence to gain a better estimate of the true underlying effect size.

2. What does replication mean?

Though there is some confusion in terminology and basic terms are not yet standardized, we can state that the concept of replication is not isomorphic to the concept of reproduction (Goodman, Fanelli, & Ioannidis, 2016; Plesser, 2018; see also Schauer & Hedges, 2020). *Reproducibility* refers to the possibility that an independent researcher or analyst finds the same results reperforming the same analysis on a given dataset of a given experiment. The purpose of reproduction is to check for errors of omission or commission (e.g., errors in coding, performing the analyses or in reporting them). Note that the term was coined by Jon Claerbout, a computer scientist, to discuss issues of transparency in the computer sciences, and not issues of corroboration (Claerbout & Karrenbach, 1992). In fields such as computational science reproduction has value because a different analyst can reperform the analysis using the same code or data. In psychology reproduction (a) is seldom a possibility because the same sample cannot be studied twice without the influence of learning or maturation, or (b) it simply refers to a procedure with which researchers check whether the analyses were properly undertaken and reported. *Replication*, by contrast, refers to the possibility of finding the same or a qualitatively similar answer by conducting, on a new sample or dataset, the same analysis and relying on a procedure broadly equivalent to the procedure of the original study. Of course, reproducibility is a precondition for replication; there is little point in attempting to replicate a study which cannot even be reproduced.

So, replicability assesses whether similar results can be obtained by applying the same or similar procedures. In practice, replicability is used by many researchers to address issues of confidence around whether the original study actually detected the ‘truth’, often, however, without reaching a definitive solution to the problem. Following the definition given by a U.S. National Science Foundation advisory committee on replicability in science, replicability refers to “the ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected” (Bollen, Cacioppo, Kaplan, Krosnick, & Olds, 2015, p. 3). However, the definition does not clarify the criteria for considering the results to be ‘the same’ or qualitatively similar (Goodman et al., 2016). Moreover, how similar should the procedures be? Suppose there is an opportunity to refine and enhance them? Should this be rejected?

With respect to experimental design, method and procedure, a further distinction is made between *direct (or exact) replication* and *conceptual replication* (Fabrigar & Wegener, 2017; Simons, 2014). With a direct replication, researchers seek confirmation by running a study that employs the exact same procedures of the original study. With a conceptual replication, researchers aim instead to address the same questions of the original study, but procedures and methods might be modified to some extent. Although this distinction is commonly accepted as a clear-cut one, in practice it is somewhat blurred (Crandall & Sherman, 2016; LeBel et al., 2017). If we consider that two independent samples will always differ, it is unclear how to interpret the term ‘exact’. Please note, however, that here we are not interested in using the term ‘exact’ as a strawman, as sometimes it is used to conclude that a case of exact replication will never exist in practice. Rather, we focus on the fact that problems arise if we try to understand when the attempt to replicate ceases to be exact or direct. Even if the replication uses the same participants who were involved in the original study, the two samples would differ, as participants would be affected by learning. We of course recognize that questions assessed with inferential statistics address whether there is an effect in the population and not in the sample per se, as the idea is to test hypotheses by drawing random samples from the population. However, whereas it is perfectly appropriate to conduct replications on different samples, we point out that variations in the sample composition among studies and random error in the results sometimes play a greater role than one might imagine (Klein et al., 2014; see also Simonsohn, 2015). In particular, we will see that sampling error alone can predict and explain a wide range of replication results, and a consideration of this factor may thus influence the way we interpret and assess replication results. Please also note that this reasoning does not apply to exact or direct replication only but also to conceptual replication. In both cases, a full consideration of sampling error would help us to see the usefulness of shifting from hypothesis testing within a dichotomous fashion to a focus on effect size estimation.

Another crucial distinction is the one between *confirmation* and *disconfirmation*. This is far from being a trivial distinction, especially in the context of replication (e.g., Cumming, 2014; Goodman et al., 2016; Nickerson, 2000; see also Badenes-Ribera, Frias-Navarro, Iotti, Bonilla-Campos, & Longobardi, 2018; Bakker, Hartgerink, Wicherts, & van der Maas, 2016). Simonsohn (2015) presents an approach which seeks to combine the traditional ideas of dichotomous hypothesis testing with notions of statistical power and effect size. His suggestion is that a replication study needs to show that the effect of interest is detectably different from zero *before* it is appropriate to consider either whether the effect magnitude is significantly different from the original study or to estimate the actual effect size. We agree that studies employing under-powered tests are fraught with problems and may well not be worth replicating (a point we shall return to), but the problems of thresholds and dichotomous decision-making regarding hypotheses are widely reported (e.g., Amrhein, Greenland, & McShane, 2019; Colquhoun, 2014). Also, one can envisage a situation where a small low-powered experiment could effectively block further enquiry in what might be a promising field.

In the present article we provide an example of why there is much confusion regarding the nature of confirmation and disconfirmation in the context of replication. Replication attempts by infancy researchers are low (Oakes, 2017; see also Asendorpf et al., 2013; Fanelli, 2012). Researchers face the problem of recruiting and testing a particular population, and they are well aware of how these processes can be challenging. Likely because of this problem, many infant studies have small sample sizes. The risk of reporting both false positive and false negative results is thus non-negligible (Oakes, 2017). In addition, effect sizes are likely to be over-estimated especially when viewed through the prism of statistical significance (Ioannidis, 2008). Starting from these observations, the *ManyBabies* project has been proposed as a partial solution of confidence issues (Frank et al., 2017). The idea is to assess and promote replicability by conducting large-scale, multi-laboratory replications of findings of particular interest according to the research community. ManyBabies collaborators are now seeking to replicate Surian and Geraci (2012), a study on infants' ability to represent true and false beliefs, and, during 2017-2018, they have started to actively plan to replicate Hamlin et al. (2007), a study suggesting that humans develop a moral sense very early in life.

3. A recent example of replication in infant research

In 2007, Hamlin and her colleagues published research reporting evidence of a preference for helping or 'morally good' agents over hindering or 'morally bad' agents in a group of preverbal infants. This initial finding attracted the attention of many researchers who further investigated the ontogenetic origins of human socio-moral understanding and preferences. Among them, Salvadori and others (2015) attempted to replicate one of the main Hamlin's studies, namely Hamlin and Wynn (2011). In this latter (original) study, infants within their first year of life were presented with a puppet show in which a protagonist tried to open a box and was either helped or thwarted by a second puppet. Infants were then encouraged to choose whether to pick up the opener, helping agent, or the closer, hindering agent. Nine-month-olds preferred the helper: 12 infants out of 16 chose the opener over the closer, and the result reached statistical significance at $p < 0.05$ (one-tailed test).

Salvadori et al. (2015) had a bigger sample size, $n = 24$, but reported that only 15 infants chose the helper, a result that did not reach statistical significance. Authors then conducted a second study, in which they followed even more closely the procedure used in the original study. However, they found an even stronger null effect: 12/24 infants chose the helper. Authors thus concluded that they did not confirm the original study ("We made two attempts to replicate the result of Hamlin and Wynn (2011), without success", Salvadori et al., 2015, p. 7). In their discussion, Salvadori and colleagues speculated that having found a lower preference compared to the original study and results that did not reach statistical significance could be somewhat explained by (a) the true effect size being smaller than what originally suggested, and/or (b) methodological dissimilarities, for example in the population tested or in the sample used (see also Schlingloff, Csibra, & Tatone, 2020). These are valid points, and in the current paper we show the importance of considering the fact that different samples have been used in different studies (sampling error) when interpreting replication results.

Is the pattern of replication results found by Salvadori et al. (2015) surprising? Could have we foreseen it? To answer these questions, we need to briefly discuss the confirmatory logic of replication currently accepted (at least in practice) by most infant researchers. But first let us consider a factor that across many scientific disciplines has been found useful in predicting the outcome of a replication study: the distinction between internal and external replication (Duncan, Engel, Claessens, & Dowsett, 2014). Do researchers who attempted to replicate belong to the laboratory from which the original finding was obtained, or are they close co-authors though they work in a different laboratory? If the answer is no, we have an *external* or independent replication. If the answer is yes, we have an *internal* replication. The two classes, however, are not equivalent, as there is evidence that internal replications have a much higher probability of being interpreted as confirmatory than external replications (e.g., Coyne, 2016). This phenomenon is also known as the *same team science effect*, and perhaps can be in part due to the fact that the original laboratory has better expertise or ability to conduct closer replications (Ioannidis, 2012). But there is also the risk of confirmation bias. With respect to research on infant socio-moral preferences, Margoni and Surian (2018) reported that studies conducted in the laboratory where the original finding was discovered had significantly larger effect sizes than studies conducted by independent laboratories, and this increment of the effect size was equal to 6.29 units in terms of the average proportion of choosing the helper.

4. The confirmatory logic of replication

Null Hypothesis Significance Testing (NHST) is still the dominant approach to hypothesis testing in psychology (e.g., Bakker & Wicherts, 2011; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Under this approach, to count as a confirmation, the replication study should report a statistically significant result in the same direction of the original finding. When this does not occur, authors most often conclude that the original finding has not been confirmed, and the credibility of the research proposition is lowered. On a side note, we observe that, within the NHST approach, the possibility of replicating non-significant results most often does not seem to arise, although researchers have developed interesting tools for testing for the absence of a meaningful effect (e.g., the smallest effect size of interest, SESOI) and judging whether values are statistically equivalent (Lakens, Scheel, & Isager, 2018). Tools such as equivalence testing, however, are seldom employed in psychological research perhaps because they are bound to a process of identifying and justifying the chosen SESOI that sometimes can be difficult to go through. In general, researchers do sometimes focus on the size of the effects when interpreting the relevance and implications of their results, but generally this is done only after that *p*-values are carefully considered.

According to the NHST approach, the null hypothesis typically states that there is no effect, specifically it is precisely zero (although we recognize that the null hypothesis can assume any numerical value). The alternative hypothesis states instead that there is some non-zero effect, but the magnitude is often not specified. Within this approach, researchers rely on the value of *p* to reach their conclusions. However, finding that the value of *p* is below the significance level says nothing about the actual size of the effect (Cumming, 2014; see also Sellke, Bayarri, & Berger, 2001). The difference between significant and non-significant results may thus, in itself, not be very relevant if the focus—in the context of replication—is on analyzing the differences between effect sizes (Gelman & Stern, 2006; Goodman et al., 2016; Stanley & Spence; see also Hedges & Schauer, 2019).

According to Amrhein et al. (2017), if we take two studies, say a replication study and its original, each employing a test with a good statistical power of 0.80, and assume that there is an underlying true non-zero effect, we can calculate that they will be conflicting (i.e. one is significant but the other is not) in one third of the cases. Here one third of the cases refers to both significant original studies with non-significant replications and non-significant original studies with significant replications. Note, however, that due to publication bias, in practice original studies that are then replicated are hardly ever of the second kind. On the one hand, then, the percentage calculated by Amrhein and colleagues may be inflated. On the other hand, however, because it is also true that in practice, and especially in infant research, tests have low statistical power (for given effects), these assumptions may also be considered optimistic (Maxwell, 2014).

There is a growing consensus that the focus on *p*-value does not help to systematically accumulate and integrate knowledge (Szucs & Ioannidis, 2017). Many statisticians have been making this point for a long time now (e.g., Carver, 1978; Krantz, 1999; for an accessible summary see Colquhoun, 2014). Frustratingly, however, psychology researchers still heavily rely on the NHST approach and, in the context of replication, often tightly link confirmation to a *p*-value below the alpha level. Relying on *p*-values to decide whether a replication confirmed the original study is particularly problematic if the original study employed a statistical test that has low power (for a range of possible underlying true effect sizes). Low power increases the likelihood that a finding reaching statistical significance is a false positive or, in the case the effect is true, increases the likelihood that the magnitude of the estimate is inflated (Gelman & Carlin, 2014). So, when researchers interpret replication results by relying on statistical significance, it is unsurprising―for original studies that employ under-powered tests—that they often conclude that the original study is not confirmed or reported much bigger effects compared to the replication attempt.

The problems with replication studies that we raise and discuss here also apply to original studies. Please note, however, that in the present paper we focus specifically on replication, and for this reason we do not offer a detailed discussion of the traditional dichotomous approach to theory testing for original studies (for such a discussion see Cumming, 2014). Still, it is important to clarify that by inviting researchers to rely on effect sizes rather than *p*-values when evaluating replication and original studies, we are not dismissing theory testing and advocating a purely descriptive approach to statistics. Focusing on effect size estimation is not incompatible with inferential statistics and theory testing and building. We indeed believe that generating predictions, not just on the presence/absence of an effect, but on the size of it, and evaluating the meaning of the size of the effect for the theory under assessment would actually represent an improvement with respect to the current approach to theory testing.

5. Measurement and sampling errors

Two notions are crucial for understanding why confirmation under the NHST approach to replication is problematic: *measurement error* and *sampling error* (Spence & Stanley, 2016; Stanley & Spence, 2014). When researchers infer the presence of an effect in a population from a sample of that population, this inevitably involves some variability (e.g., Klein et al., 2014). This variability may explain why two researchers that independently investigate the same phenomenon are likely to obtain different results. However, by the same token, it is unclear how much these results can differ before being deemed mutually inconsistent. Researchers who accept the NHST approach simplify the task by grouping the results accordingly to whether they fall below or above the 0.05 *p*-value threshold. Going beyond this approach implies considering how crucial measurement and sampling errors are in determining the variability between studies.

Because each study potentially contains multiple sources of error, even two studies with identical procedures will produce different results. But how large is the impact of these sources of error?

As shown by Stanley and Spence (2014), measurement error alone can explain a great deal of variability and, thus, may help make sense of the reported 'low rates of confirmation' in replication studies. Even assuming ideal replication conditions (which is when the original study is replicated on the very same group of participants, who have no memory of the original testing session), the range of results that can be expected is very large. Say that the original study reported a correlation of $r = .30$, and we know (assuming omniscience) that the true correlation is indeed $r = .30$. Stanley and Spence (2014) calculate that, if there were 100,000 replication studies with $n = 40$ and reliability = .70, the results would range between $r = .03$ and $r = .46$. This range of possible results is very wide, and is normally distributed. We can thus appreciate the amount of variability that can be present only because of measurement error. And note that this is under otherwise ideal conditions, which are never met in practice.

Under real conditions, it is not feasible to have participants with no memory of previous experimental sessions. Therefore, an important source of error is related to *sampling*. By taking into account both measurement and sampling error, Stanley and Spence (2014) further calculated that a replication could legitimately find also a negative correlation of $r = -.33$ in a simulation where the original result reported a positive correlation of $r = .30$ with a true known correlation of $r = .10$. This already gives us a sense of how unhelpful it can be to draw any conclusion from a single replication study. Indeed, any pair of original and replication studies has two different samples. For this reason, results will be inevitably influenced by sampling error.

We can appreciate how crucial this is in the context of replication by calculating, from a given original study, the prediction interval (Spence & Stanley, 2016). The *Prediction Interval* is the range of results that can be expected in a replication study due to chance or sampling error alone. Imagine one randomly selects a set of samples from a given population and, for each of them, computes an effect size. First, observe that, because of the Central Limit Theorem, differences between effect sizes follow a sampling normal distribution. Second, two factors will determine those differences: sample size and population variance (or sampling error). From here, it is possible to estimate the range of effect sizes which are due only to sampling error. In order to compute the prediction interval, we only need to know (a) the sample size of the original study, (b) the observed effect size from the original study, and (c) the intended sample size of the replication study. This interval provides a measure of the extent to which the replication study could differ only because of sampling error. A prediction interval is not a confidence interval. Whereas the confidence interval estimates the uncertainty of a population parameter, the prediction interval refers to a future sample statistic. With the R package *predictionInterval*, developed by Spence and Stanley (2016), it is possible to calculate prediction intervals for commonly used indexes of effect size, such as correlations or Cohen's *d*s. As we will see by means of examples taken from the infant literature, prediction intervals can be surprisingly broad and include non-significant results.

Here, to illustrate how replication findings may vary, we took as an example the context of the infant socio-moral preferences literature and ran a Monte Carlo simulation based on 10,000 random samples from a known binomial distribution. We used one sample of proportions as in the real studies on socio-moral preferences and set the sample size of each replication n = 16 because this is the median study size in this literature (see Margoni & Surian, 2018). Further, since this was a simulation, we could build in the ground truth of the, in reality unknown, population parameter π, that we chose to set to $\pi = 0.64$, based on the unbiased estimate calculated by Margoni and Surian (2018).

Figure 1 displays the histogram of the simulated experimental results, where the dashed vertical line denotes the ‘true’ proportion. First, we can observe that results are quite coarse because if $n = 16$ then, by definition, there can only be $n+1 = 17$ possible outcomes. As a result, the simulation modal class slightly overestimated the true value (0.69 rather than 0.64). Please note that this occurs *even* when we build in the ground truth and there are no nuisance factors whatsoever. Furthermore, the closest possible result (0.63) will be a slight underestimate. Therefore, our first observation is that we need to be aware that small sample experiments are a very blunt instrument to accurately estimate proportions. Table 1 sets out the same simulation results in terms of ascending proportion P = (X/n), where X is the observed count of infants who chose the prosocial puppet, n is the sample size (here, $n = 16$), and p is the statistical significance from an exact binomial test (assuming it is 1-tailed as per Hamlin et al., 2007).

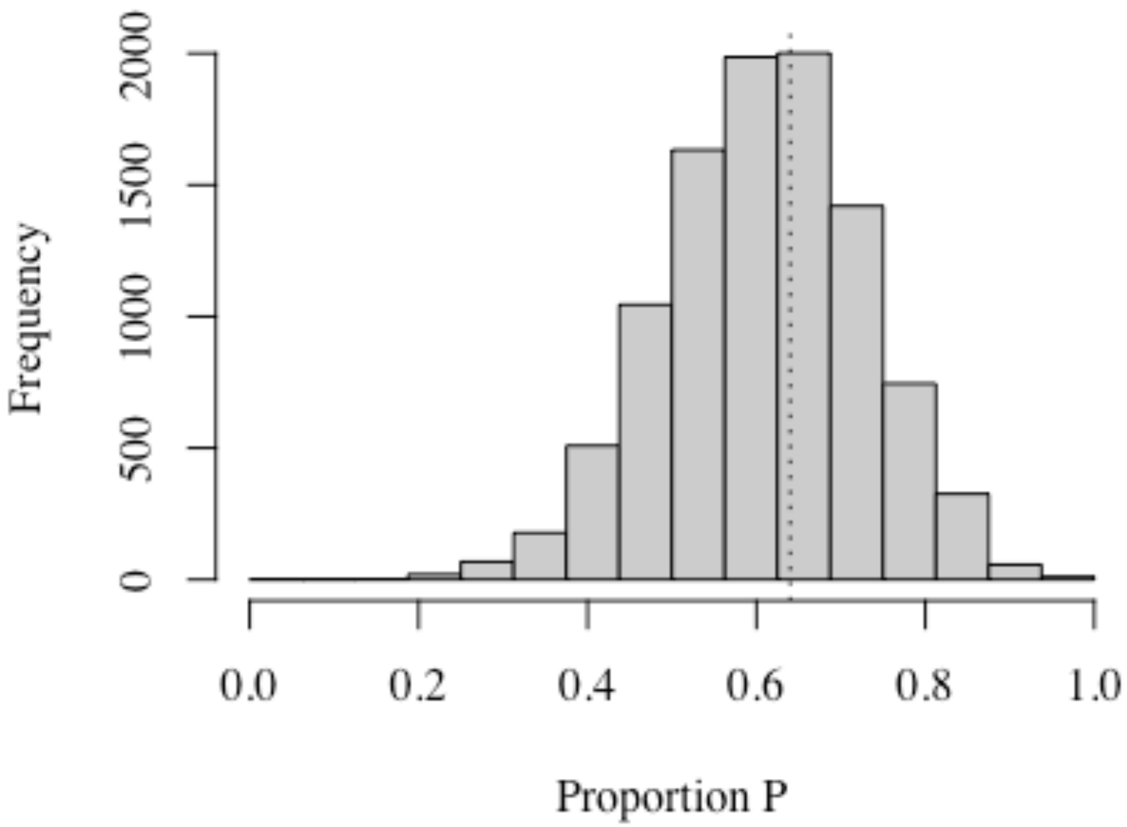

Figure 1. Histogram of the distribution of proportions P, found from a series of 10,000 random experiments with n = 16 when the true value is $\pi = 0.64$ (denoted by the dashed vertical line and based on Margoni and Surian, 2018).

Table 1

Simulated Experimental Results Ordered by Proportion (X/n where n = 16).

X/16	Proportion	Experiment count	% of total	Cumulative %	*p* (NHST 1-tailed)
0	0.00	0	0.00	0.00	1.0000
1	0.06	0	0.00	0.00	1.0000
2	0.13	0	0.00	0.00	0.9997
3	0.19	7	0.07	0.07	0.9979
4	0.25	16	0.16	0.23	0.9894
5	0.31	60	0.60	0.83	0.9616
6	0.38	194	1.94	2.77	0.8949
7	0.44	539	5.39	8.16	0.7728
8	0.50	1033	10.33	18.49	0.5982
9	0.56	1572	15.72	34.21	0.4018
10	0.63	1932	19.32	53.53	0.2272
11	0.69	2053	20.53	74.06	0.1051
12	0.75	1429	14.29	88.35	0.0384
13	0.81	794	7.94	96.29	0.0106
14	0.88	282	2.82	99.11	0.0021
15	0.94	82	0.82	99.93	0.0003
16	1.00	7	0.07	100.00	0.0000

Note. Experiment count indicates the number of experiments simulated from the known binomial distribution that, out of 10,000 simulations, reported a certain proportion.

By looking at the results of this simulation, we also note that there were no observations for X = 0-2. However, we found that the middle 98% of the 10,000 simulations fall between P = 0.38 and P = 0.88, which is a strikingly wide range of expected results to find from a study[2]. Any result between these two values, when the true effect is 0.64, can be considered as expected simply due to sampling error, assuming we accept a confidence interval at the 98% level (which of course is arbitrary; Morey & Lakens, 2016). We can further appreciate that ~18% of the

[2] Strictly speaking the 0.01 and 0.099 quantiles enclose more than 98% of the data due to the coarse nature of X (as only n+1 = 17 distinct observations are possible).

experiments reported P ≤ 0.50 (i.e. an effect in the opposite direction, which for our example is a preference for the antisocial agent), and ~75% would not attain statistical significance assuming a NHST perspective and $\alpha = 0.05$. These results show how problematic it would be to rely on *p*-values when considering whether a single study is 'true' or not. And it is especially problematic if (a) the single replication study has a small sample size and (b) there is great uncertainty about the true effect size, which was the case of Salvadori et al. (2015) who tried to replicate an original result of P = 0.75 when the true effect size is instead likely P = 0.64 (based on the pooled estimate of Margoni and Surian, 2018).

Because a wide range of differences between the original and the replication findings can be explained by sampling error alone, and anticipated by a prediction interval, results of the original studies will be confirmed (in a formal statistical sense) by replication attempts in almost all cases. This makes a single confirmation of the original result, or even a few of them, relatively uninformative. Moreover, it makes the possibility of gaining new knowledge or concluding anything useful quite improbable. Should we, for example, tell educators or parents that the original result that young children prefer prosocial agents has been confirmed when this is true in a formal statistical sense, yet the original study reports a large effect and the replication reports a medium effect but in the opposite direction?

6. Prediction interval in practice

We now focus on prediction intervals more closely by further discussing the case of Salvadori et al. (2015) and computing the prediction intervals from the original study that the authors attempted to replicate. Hamlin and Wynn (2011), the original study, had a sample size of 16 9-month-old infants, and reported that 12 of them chose the prosocial agent. First, we observe that the sample is small, although in line with current standards in infant research. Next, to calculate how large the range of results that can be obtained in a replication attempt we make use of the two attempted replications of Salvadori et al. (2015). Both had $n = 24$ and neither reached statistical significance. In the first attempt, 15 infants chose the helper (62.5%), whereas in the second attempt, 12 of the infants chose the helper (50%).

Because the tool developed by Spence and Stanley (2016) does not allow calculation of prediction intervals for binary measures, we developed a *R* code function which is available in the Supplementary Materials (please also see the on-line OSF project where we uploaded the data and *R* code associated with the current work: https://osf.io/swck7/). For a study with $n = 24$ which seeks to replicate a study with 12/16 successes (0.75), the prediction interval can be calculated to be PI = (0.43, 0.96). This is a wide range of possible results, corresponding to a lower bound of 10/24 and an upper bound of 23/24 successes. Note that this comfortably includes both results reported by Salvadori et al. (2015). The outcome does not change if we combine the samples of the two replication attempts, obtaining a replication study with 27/48 successes. The result falls again within the predicted interval, PI = (22, 45). Thus, although the results were interpreted by Salvadori and colleagues as disconfirmatory, they actually confirm the original result because the differences are compatible with sampling error alone.

At least two lessons can be learned from this example. First, *p*-values are not a reliable way to determine whether the original findings have been confirmed or not. A focus on the effect sizes (and associated confidence intervals) as a measure of the relevance of the findings would be more useful. Second, because variability accounted by sampling error can explain why two or more similar studies report very different results, we cannot infer much from a few low-powered attempts to replicate. It is necessary to run more studies and, we suggest, adopt a meta-analytic mindset.

We mentioned that part of the problem with replication attempts resides in the fact that statistical tests in the original studies have power curves that are low for a considerable number of theoretically relevant deviations from the null hypothesis. We next consider what insight simulation can shed on the relationship between statistical power and prediction intervals. First, in the current context, we acknowledge that most of the experiments covered by Margoni and Surian (2018) employed tests that had very low statistical power (especially to detect the assumed true underlying effect size P = 0.64, which is the best guess we currently have, based on the meta-analysis). Figure 2 shows the relationship between the sample size and the power to detect an effect of P = 0.64, which corresponds to a small effect size of Cohen's *h* = ~0.284[3]. The statistical power of the test in the original experiment of Hamlin et al. (2007), where *n* = 16, is indicated in the figure with a dashed line (note, however, that they used 1-tailed test), and we can clearly see how low it is[4]. Moreover, we can notice that to achieve a power = 0.80, which is what psychological researchers customarily aspire to, we need approximately 100 participants, and this is without any nuisance factors being added (thus, in a scenario that likely overestimates the statistical power of the test).

[3] Cohen's *h*, which is interpreted analogously to Cohen's *d* (see Cohen, 1992), is the difference between the root arcsine transformations of the two proportions (here P = 0.50, the arbitrary no effect value, and P = 0.64).

[4] Indeed, the power is almost certainly over-estimated because the calculation uses a post hoc estimate of the effect size which will tend to be inflated due to the small sample size and use of a NHST filter which ensures only very extreme values of P can be 'significant' (Szucs & Ioannidis, 2017).

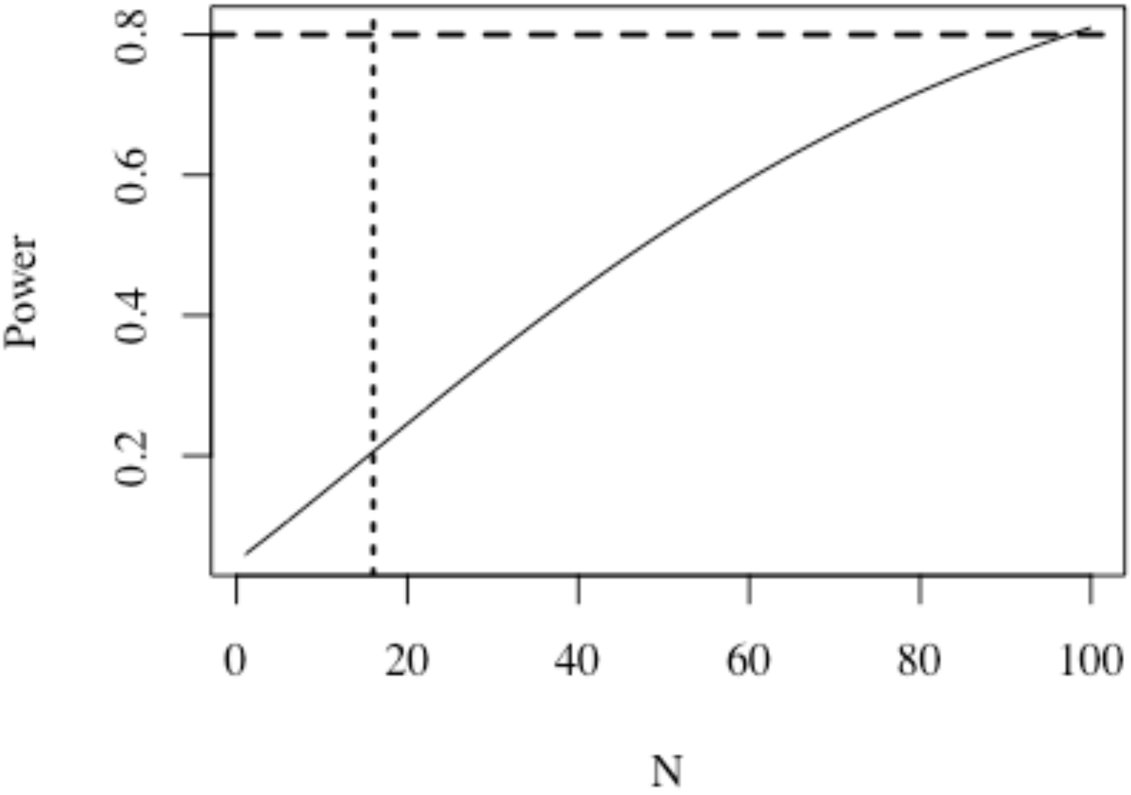

Figure 2. Plot of the relationship between statistical power and sample size when the true effect is π = 0.64. The vertical, dashed line indicates the power of the test in the original experiment in Hamlin et al. (2007), where *n* = 16, and the horizontal, dashed line indicates power = 0.80.

Second, in Table 2 we can see how if the statistical test of the original study is under-powered (i.e. *n* = 16) in the sense that the interval due to sampling is very wide, this is not something that can be remedied simply by powering up the replication. The reason is not hard to appreciate: confirming an imprecise result can be accomplished by a wide spectrum of possible replication results.

Table 2

Relationship Between Prediction Interval (PI) and Replication Sample Size for an Original Study With n = 16.

Size of replication	PI Lower bound	PI Upper bound
16	0.473	1.000
20	0.498	0.997
40	0.531	0.978
60	0.533	0.965
80	0.543	0.961
100	0.549	0.959
200	0.559	0.953

Note. The assumed true effect size is $\pi = 0.64$.

Table 3 reveals the impact on prediction interval if the original study had a larger sample size, whilst keeping the replication sample size small at $n = 16$ or setting it at $n = 32$. What is striking is that the prediction interval will still be broad even when the original study is relatively well powered with $n = 200$. The reason for this is that a replication study with a small sample, even if we have a high level of confidence in our estimate of the proportion π from the original study, might reasonably produce a wide range of outcomes simply due to sampling error. This, again, illustrates how uninformative the traditional dichotomous replication paradigm can be, particularly in the context of under-powered studies. In contrast, pooling estimates would be more informative. As Goodman and colleagues wrote “a preferred way to assess the evidential meaning of two or more results with substantive stochastic variability is to evaluate the cumulative evidence they provide vis-á-vis a hypothesis of interest and not whether one contradicts or discredits the other through the lens of statistical significance." (Goodman et al., 2016, p. 3).

Table 3

Relationship Between Prediction Interval (PI) and Original Sample Size for a Replication Study With n = {16, 32}.

Size of original study	PI Lower bound n=16	PI Upper bound n=16	PI Lower bound n=32	PI Upper bound n=32
16	0.272	0.921	0.319	0.879
20	0.333	0.938	0.361	0.881
40	0.381	0.921	0.411	0.857
60	0.366	0.892	0.426	0.851
80	0.384	0.894	0.449	0.835
100	0.395	0.895	0.457	0.835
200	0.407	0.890	0.472	0.827

Note. The assumed true effect size is $\pi = 0.64$. This, however, was not exact due to the lack of whole number solutions for the smaller samples.

7. Monte Carlo simulations with nuisance factors

We now compare several Monte Carlo simulations of experiments still with $n = 16$ and $\pi = 0.64$ but introduce a nuisance factor since up until now we have assumed a perfect binomial distribution and no complicating factors. In less ideal scenarios, a common problem with the binomial distribution is overdispersion (excess variance over what might be expected given the data generator), leading to outliers and heavy tails. Overdispersion implies that summary statistics may have larger variance than expected by the simple model, and this latent heterogeneity may cause loss of efficiency when using statistics appropriate for the simple model and indicate the presence of different subpopulations. For instance, Johnson et al. (2015) showed the harmful effects of overdispersion upon statistical power in their Monte Carlo simulations. We believed such a situation was likely to pertain also for the results of Margoni and Surian (2018). First of all, we tested the possibility that overdispersion was strongly present in the results inserted in Margoni and Surian (2018) by fitting a general linear mixed model (GLMM) of the binomial family based on age (mean age of participants in each experiment), year of the study, and type of scenario (whether infants saw helping vs. hindering events, fair vs. unfair distributions, or events of giving vs. taking an object from others). We selected two variables of potential theoretical interest (age and scenario) and one variable (year) that can be considered a nuisance factor as it is of no potential theoretical interest (for details see Margoni & Surian, 2018)[5].

response ~ (1 | year) + age_group + scenario

To assess overdispersion, we examined the ratio of the residual scaled deviance which should be roughly equal to the residual degrees of freedom (Hinde & Demétrio, 1998). In our case, the ratio was approximately 1.97 indicative of problems and the fact that the model does not adequately account for all the variance[6]. Consequently, we consider overdispersion to be a realistic and non-trivial nuisance factor. Next, we added a normally distributed error term to a

[5] The model was chosen to (i) minimize AIC, (ii) maximize the variance associated with the grouping variable (year) of the hierarchical model (i.e. there is no point in having a mixed model with the random components near to zero variance), (iii) keep it parsimonious (there are only 61 observations) and avoid a model that is (or is close to) singular (Bates, Kliegl, Vasishth, & Baayen, 2015).

[6] We fitted the generalized linear mixed model using the Laplace Approximation for maximum likelihood as implemented by the R function ‘glmer’.

simple model where the intercept was the estimated p to generate excess variance compared with what we would expect from the binomial model[7].

$$p' = \hat{p} + N(m, \sigma)$$

In the model, p' is the updated value of the estimated proportion p-hat by a random normal variate with a mean of zero. We then took three different cases of the standard deviation σ, with values of 0 (i.e. no nuisance), 0.169 (the SD of the proportions p in the studies identified by the meta-analysis of Margoni and Surian, 2018) and 0.338 (extreme overdispersion with a SD twice that observed in the meta-analysis), and simulated 10,000 experiments with $n = 16$ under each of these three conditions. Table 4 shows the impact of increasing the number of outliers. Interestingly, having more outliers means having a higher likelihood to obtain a statistically significant result but increasingly in the wrong direction. Recall that we have set the true effect to $\pi = 0.64$ which corresponds to a small positive effect of Cohen's $h = {\sim}0.28$. In the last column of the table, we see the effect of the *winner's curse* (which predicts that the original findings systematically overestimate the true effect size; see Gelman & Loken, 2014) in that, if we only accept statistically significant results, when attempting to estimate the proportion by pooling, we obtain highly inflated estimates (Ioannidis, 2008; Young et al., 2008).

[7] The base model of estimated p as the intercept implies a homogenous effect.

Table 4

Monte Carlo Simulations Results of 10,000 Experiments, n = 16, π = 0.64 (True Effect Size h = ~0.28), NHST (α = 0.05, 2-tailed Using Binomial Exact Test).

Nuisance factor	Significant results	Wrong direction significant results	Pooled significant estimated effect size (*h*)
σ = 0 (none)	11.52%	0.02%	0.74
σ = 0.169	26.54%	2.14%	0.96
σ = 0.338	36.53%	12.30%	1.22

Note. The pooled effect size estimate is derived from a weighted mean of significant results in the expected direction, i.e. P > 0.5.

Finally, in Figure 3 we can see graphically the difficulties of under-powered tests in the face of adding substantial numbers of outliers where the distribution is essentially uniform except for the tails where the observations are elevated due to clamping the distribution [0, 1].

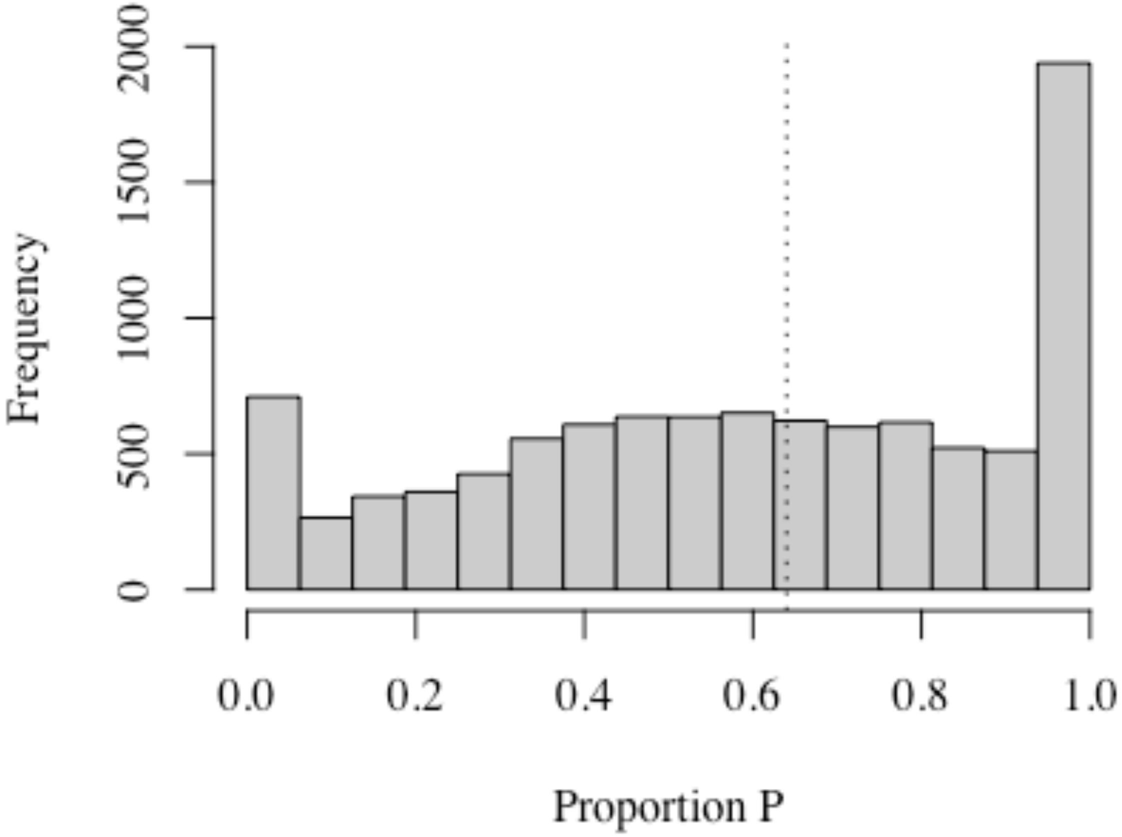

Figure 3. Histogram of the simulated experiments when the true effect is π = 0.64 and severe overdispersion is injected (SD = 0.338).

8. Moving towards a meta-analytic mindset

Inasmuch as, in the context of infant research, statistical tests are likely frequently under-powered and sampling error is a real but unavoidable concern, what can researchers conclude from single replication studies? Because the range of results that can be predicted from an original result is often very broad, it would be better not to conclude anything. Attention could be shifted from addressing whether the original result was confirmed to accumulating knowledge and providing better estimates. In this regard, a useful approach is meta-analysis (for those unfamiliar with this tool, a concise introduction can be found in Del Re, 2015; see also Borenstein, Hedges, Higgins, & Rothstein, 2009; Viechtbauer, 2010).

As now suggested by many, we may benefit from shifting the focus from *p*-values to estimating effect sizes or, in other words, from abandoning the logic of confirmation under the NHST approach (Cumming, 2014; Fidler & Loftus, 2009; Finch et al., 2004; Kirk, 2003; Kline, 2004; Velicer, Cumming, Fava, Rossi, Prochaska, & Johnson, 2008). Diverting the attention from the comparison of results, meta-analysis would permit to reach a conclusion about the size of the effect by pooling together the findings. In this way, using meta-analysis would make it easy to resist the temptation of drawing conclusions from a single study.

In the literature on infant socio-moral preferences, the meta-analysis by Margoni and Surian (2018), besides estimating the average effect size and its confidence interval, proved useful in assessing the heterogeneity among the studies. The Higgins' I^2 index expresses how much of the total variability in the effect size estimate can be attributed to heterogeneity among the selected studies rather than to chance (Higgins, Thompson, Decks, & Altman, 2003; see also Lin, 2019). With respect to the studies reported in the meta-analysis, the estimated amount of total heterogeneity was $\tau^2 = 0.27$, 95% CI (0.13, 0.63), and Higgins' I^2 was 0.52, 95% CI (0.35, 0.71), suggesting moderate heterogeneity (Borenstein, Higgins, Hedges, & Rothstein, 2017). Finding evidence of heterogeneity usually motivates researchers to test which factors help explain the differences between studies outcomes.

A crucial source of difference between studies is often whether they are published or not (Coyne, 2016; Ioannidis, 2012). Out of the 61 effect sizes included in Margoni and Surian

(2018), 17 were unpublished. Likely because of publication bias, these unpublished studies, compared to the published ones, reported a smaller average effect size. Because many journals still tend to reject manuscripts that report only null results, unpublished studies, compared to published studies, likely contain more small effects and statistically non-significant results (Rosenthal, 1979; Sterling, 1959; see also Schmidt & Hunter, 2014).

A way to assess whether there is publication bias is to select the effect sizes that were published and use the trim-and-fill method (Duval & Tweedie, 2000; but see also Moreno et al., 2009, for regression-based methods to adjust for publication bias that can outperform classical methods). Trim-and-fill estimates how many studies are missing to produce a symmetric funnel plot distribution. In Margoni and Surian (2018), nine studies were missing on the left side of the graph to produce the symmetric funnel plot, suggesting the presence of a publication bias. Moreover, it was possible to estimate the average effect size and its confidence interval adjusted for the publication bias. A random-effects meta-analysis using both published and unpublished effect sizes which did not account for publication bias estimated that an average proportion of 0.68, 95% CI (0.64, 0.72), of infants would choose the prosocial agent. In contrast, a meta-analysis which accounted for the publication bias estimated an average proportion of 0.64, 95% CI (0.60, 0.69), that is the estimate we used throughout our article.

It is important to note, however, that the methods currently available to adjust meta-analytic estimates for publication bias are far from being perfect. Trim-and-fill in particular has limitations and many suggested to exercise caution in its interpretation (especially in the presence of substantial heterogeneity between studies) or to use alternative methods (e.g., Carter, Schönbrodt, Gervais, & Hilgard, 2019; Macaskill, Walter, & Irwig, 2001; Terrin, Schmid, Lau, & Olkin, 2003). An interesting alternative solution to detect publication bias (and *p*-hacking) is *p*-curve (Simonsohn, Nelson, & Simmons, 2014). On balance, we believe that meta-analysis can be fully effective only if the research community adopts a number of measures aimed at reducing or eliminating publication bias (we will return to this point and briefly discuss these desirable measures).

In Margoni and Surian (2018), out of 26 studies reporting a total of 61 effect sizes, seven studies (reporting 11 effect sizes) claimed to have disconfirmed the original result. Three were unpublished and four were published. What factors could account for these outcomes? First, it is possible that methodological shortcomings influenced the results (Holvoet, Scola, Arciszewski, & Picard, 2016; Margoni & Surian, 2018). Second, across several disciplines, one of the best predictors of finding a confirmation is the fact that the replication was internal, as opposed to external. And, as we have already seen, this was true also for the research on infant socio-moral preferences. Without entering too much into the discussion, we observe that, in infant literature, as well as in other disciplines, small procedural changes are sometimes crucial to find significant results (see, as an example with respect to the literature on infant moral sense, Hamlin, 2015; see also Johnson & Zamuner, 2010). Although it is undeniably important to pay attention to small details when studying little children who are indeed very sensitive to minor variations in the environment (Oakes, 2017), we may also ask what can we conclude from a set of results replicating an original finding only under extremely specific and detailed procedural conditions.

We have already shown that the findings reported by Salvadori et al. (2015), which were interpreted as disconfirmations, can instead be explained by sampling error. But could have other studies that reported a disconfirmation and included in Margoni and Surian (2018) been explained by taking into account sampling error? Table 5 lists all the replication attempts that were interpreted by their authors as disconfirmations (on the basis that the results did not reach statistical significance, did not differ too much from chance level, or were in the opposite direction than what expected), their relative prediction intervals, and whether they, based on the prediction interval calculation, did indeed falsify rather than confirm the original result. Out of 11 attempts, only five reported a result which indeed falls outside the prediction interval. Yet for all 11 effects authors claimed to have failed to confirm the original result. This was because researchers usually rely on *p*-values to decide whether a study verified or falsified a previous finding. So, in more than half of the cases a result that was interpreted as a failure to replicate the original finding can instead be simply explained by taking into account sampling error.

Table 5

Prediction Intervals for Replication Studies Interpreted as Disconfirmatory and Inserted in Margoni and Surian (2018) Meta-analysis.

Replication study	Original study	Original results	Replication results	Prediction interval	Confirm?
Scarf et al. (2012)	Hamlin et al. (2007)	14/16 (0.88) *p* = .003	8/16 (0.5) *p* = ns.	[0.56, 1]	✗
Cowell & Decety (2015)	Hamlin et al. (2007)	14/16 (0.88) *p* = .003	27/54 (0.5) *p* = ns.	[0.61, 0.99]	✗
Raggio et al. (2015)	Hamlin et al. (2007)	14/16 (0.88) *p* = .003	3/10 (0.3) *p* = .206	[0.49, 1]	✗
Salvadori et al. (2015) A	Hamlin & Wynn (2011)	12/16 (0.75) *p* = .045	15/24 (0.63) *p* = .307	[0.43, 0.96]	✓
Salvadori et al. (2015) B	Hamlin & Wynn (2011)	12/16 (0.75) *p* = .045	12/24 (0.5) *p* = ns.	[0.43, 0.96]	✓
Abramson et al. (unpublished) – A	Hamlin & Wynn (2011)	12/16 (0.75) *p* = .045	17/31 (0.55) *p* = .593	[0.45, 0.96]	✓
Abramson et al. (unpublished) – B	Hamlin & Wynn (2011)	12/16 (0.75) *p* = .045	28/62 (0.45) *p* = .449	[0.47, 0.94]	✗
Hamlin (unpublished) – A	Hamlin & Wynn (2011)	12/16 (0.75) *p* = .045	11/16 (0.69) *p* = .135	[0.40, 0.98]	✓

Hamlin (unpublished) – B	Hamlin & Wynn (2011)	12/16 (0.75) $p = .045$	7/14 (0.5) p = ns.	[0.42, 1]	✓
Woo & Hamlin (unpublished) – A	Hamlin et al. (2007)	14/16 (0.88) $p = .003$	21/32 (0.66) $p = .078$	[0.59, 1]	✓
Woo & Hamlin (unpublished) – B	Hamlin et al. (2007)	14/16 (0.88) $p = .003$	15/32 (0.47) $p = .717$	[0.59, 1]	✗

What can we conclude from this revised finding (5 failed attempts instead of 11)? As argued by Morey and Lakens (2016), the success rate calculated from a prediction interval perspective is difficult to interpret because prediction intervals can encompass a very broad range of different outcomes. Also, we have a further problem: is 5 out of 11 high or low? Again, it is not possible to draw any useful conclusion from these results if we reason in a dichotomous fashion. We may instead reach a meaningful conclusion if we change the logic of replication, and focus on combining the evidence to gain estimates that come closer and closer to the true underlying effect size. All 11 studies could be combined with the original findings and the reported successful attempts to replicate in a meta-analysis to gain a better estimate of the effect we are interested in.

But take again Salvadori et al. (2015) and, this time, imagine that their study is the first attempt to replicate and the second study on infant socio-moral preferences after the original result of Hamlin and Wynn (2011). From their two null results, one may be tempted to be skeptical about the existence of a socio-moral preference in infancy and, perhaps, be persuaded to conclude that this research program does not deserve much attention. However, we have seen that sampling error alone can account for these null results. Figure 4 displays what happens when the original result of Hamlin and Wynn (2011) and the two replications reported in Salvadori et al. (2015) are pooled together. A random-effects meta-analysis on these three effects, with a restricted maximum-likelihood estimation for τ^2 (total heterogeneity between studies), estimated an average logit transformed proportion of 0.44, 95% CI (-0.12, 0.99), that back-transformed to a raw proportion was equal to 0.61, 95% CI (0.47, 0.73)[8]. Moreover, the estimated τ^2 was 0.04 and the Higgins' I^2 was 14%, CI (0%, 98%), which suggests low heterogeneity, that is, the studies are quite comparable. Most importantly, however, this meta-analysis returned an estimate that provides better information than (inaccurately) observing that the original result was not verified. Note also that, somewhat curiously, this estimate was close to the one obtained when including all the 61 effect sizes retrieved in Margoni and Surian (2018) and adjusting for the publication bias (P = 0.64). Now, the big questions that are left to researchers in this field would be how to make sense of this proportion, what factors contribute to its explanation, and what is its significance for existing theories of moral sense development.

[8] In the calculation, we used the logit-transformed proportions rather than the raw proportions to respect the normality assumption of the parametric tests.

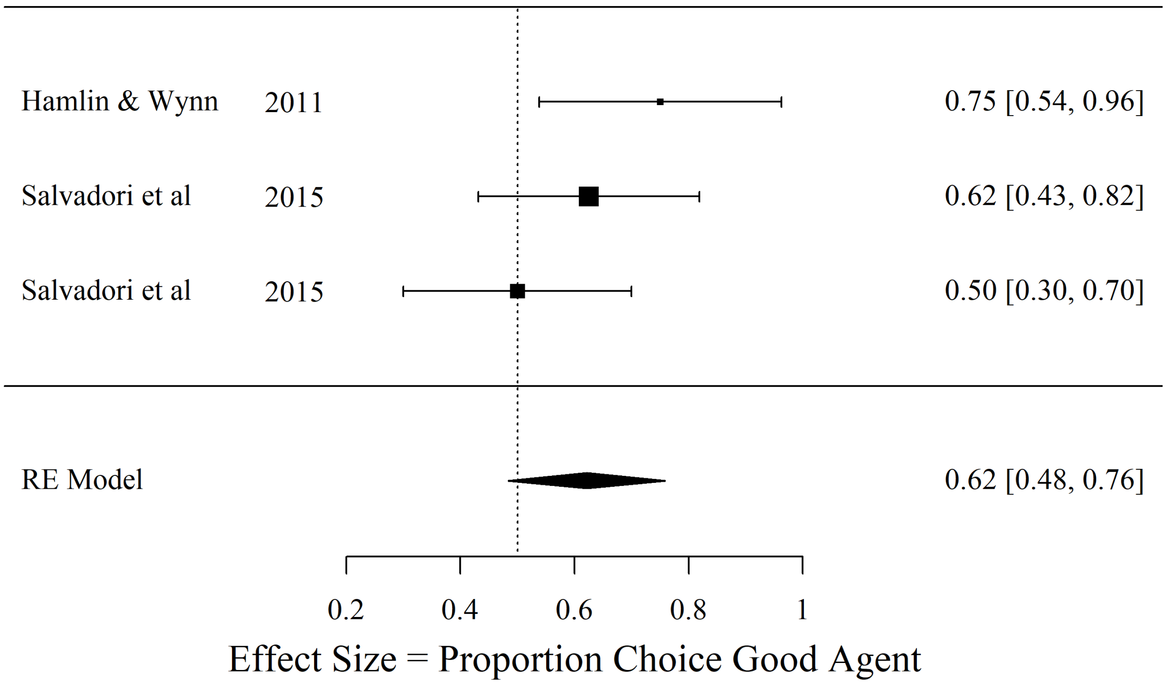

Figure 4. Forest plot showing the raw proportions and the 95% CIs for the original study of Hamlin and Wynn (2011) and the two attempts to replicate reported in Salvadori et al. (2015). The vertical dotted line at 0.50 represents the reference point indicating no effect.

9. Suggestions for (infant) researchers

We have seen that substantial variability in the findings can be expected simply because of sampling error. Whereas researchers cannot eliminate sampling error, they can adopt strategies to reduce it and to better approximate their estimates. First, it is not a meaningful practice to confirm or disconfirm a certain result by looking at the statistical significance of the outcomes of a small number of replications. Second, effect sizes can be effectively combined in a meta-analysis and a more reliable estimate of the true but unknown effect size in the population calculated. Third, studies need to be conducted with appropriate statistical power. We now discuss some of the problems that affect infant research and some of the solutions that infant researchers can adopt, a discussion that is attracting more and more attention, as evidenced by the special issue on “Replication, Collaboration, and Best Practices in Infancy Research” recently published in this journal (Frank, 2019).

One way to increase statistical power is to increase sample size. When planning a study and its statistical analyses, researchers can run a priori sample size calculations or, when they are conducting the study, assess whether the statistical test has adequate power (Asendorpf et al., 2013; Button et al., 2013; Perugini, Gallucci, & Costantini, 2014). Small samples and low statistical power (for a given effect) increase both false positive and false negative results (Fraley & Vazire, 2014; Vadillo, Konstantinidis, & Shanks, 2016). In general, the tendency of most journals to selectively accept for publication manuscripts which report positive results leads to inflated effect size estimates (Maner, 2014; Sterling, Rosenbaum, & Weinkam, 1995). If, however, we add to this practice the further tendency of undertaking experiments and tests with low statistical power, we will have an even worse problem of over-estimation of the true effect size.

Infant researchers are in special need of studies with bigger samples and tests with higher power (to detect the effects of interest). In this subdiscipline, small samples are often motivated by the fact that recruitment of participants is challenging. Infant researchers, however, may consider that increasing the power in their studies will result in more interpretable, accurate and reliable estimations. This can be a benefit for the researcher who targets real effects, as they will file fewer drafts in the drawer and perhaps, in the long run, publish more than their colleagues who instead conduct studies employing tests with low statistical power that cannot detect small but real and potentially interesting effects (see Oakes, 2017). If, instead, the researcher is targeting truly null effects, conducting tests with high statistical power will increase the likelihood of not finding an effect. But this is exactly what it is required: a way to better learn about and estimate effects, whether they are large, small, or null. Indeed, null results are, no less than significant results, an important source of information for theory testing and building.

Being fully transparent by reporting results that did not reach statistical significance will help obtain better estimates. Perhaps, we still need to internalize the idea that a single researcher's contribution is interpretable only when combined with the outcomes of other (independent) laboratories. In order to build a cumulative science and gain an always better approximation to the truth, researchers should act transparently and rely on others' results. Multi-laboratory replication projects like *Many Labs* or *ManyBabies* seek to go exactly in this direction by overcoming problems related to laboratory and sample specificity and statistical power (Frank et al., 2017; Uhlmann et al., 2019). They are extremely useful to help guarantee the quality and accuracy of the data been collected. However, the generalizability of the data in these projects is often scarce. Although being well powered to detect a very wide range of effects, these studies are still single points, as they employ a specific methodology and procedure, thus failing to systematically take into account methodological variability (Cristia, Tsuji, & Bergmann, 2020). Without dismissing multi-lab replications, the need to take into account methodological variability (among other sources of variability) leads to meta-analysis as our recommended tool for both theory testing and building and effect size estimating, where recently an interesting approach has been put forward that combines the traditional virtues of meta-analysis with concerns for research being publicly and openly available, and the possibility for the research community to constantly and easily suggests updates and improvements in the meta-analysis dataset (for details see Tsuji et al., 2014).

Meta-analysis is powerful. However, it has some important limitations too. First, the reliability closely depends on the methodological quality of the research that has been conducted (the *garbage in, garbage out* problem). If individual studies are of poor quality (e.g., severely under-powered or with a poor methodological quality), no meta-analysis can return a reliable estimate. Second, a number of biases can lead to inflated estimates. Selective reporting (Williamson, Gamble, Altman, & Hutton, 2005), file-drawer problems (Rosenthal, 1979), and other questionable research practices (John, Loewenstein, & Prelec, 2012) can bias the estimate of the true effect size. Although meta-analysis has some capacity to adjust for those biases, these problems still lead to difficulties in calculating a reliable estimate (Carter et al., 2019; Schmidt & Hunter, 2014). A problem like *p*-hacking, for example, cannot be resolved purely by relying on meta-analysis.

If we are going to rely on meta-analysis to build cumulative evidence, we need to rid the research community of publication bias whilst ensuring that published studies are of quality. A way to do so would be, as many are now advocating, to implement measures such as preregistration (Bakker, van Dijk, & Wicherts, 2012; van't Veer & Giner-Sorolla, 2016; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) and registered reports (Chambers, 2013; Nosek & Lakens, 2014). Another possibility, that has been largely overlooked, is the result-blind review or result-masked review model (Grand, Rogelberg, Banks, Landis, & Tonidandel, 2018; see also Findley, Jensen, Malesky, & Pepinsky, 2016; Woznyj, Grenier, Ross, Banks, & Rogelberg, 2018). On the one hand, these measures may lead researchers to report their findings regardless of their statistical significance and, thus, make false positives less likely and reduce publication bias (Nelson, Simmons, & Simonsohn, 2018). On the other hand, they have the potential to focus the review and publication process on the quality of the work, that is, on whether the study employs tests with adequate statistical power curves (with high enough points to detect a range of possible theoretically interesting effects).

Another problem we may need to at least be aware of is accumulation bias (ter Schure & Grünwald, 2019; see also Ellis & Stewart, 2009; Kulinskaya, Huggins, & Dogo, 2016). This bias is inherent to the process of accumulating data, and it could be present in meta-analysis because of the inevitable dependency between the time point at which the meta-analysis is conducted and the specific results (positive or null) of the studies that are included in the meta-analysis. Indeed, the timing of the meta-analysis (if performed) could be guided, for instance, by the publication of a couple of overly-optimistic estimates. The problem arises because this affects the estimates and also, and importantly, the sampling distribution, resulting in inflated type-I errors in *p*-value-based tests. As ter Schure and Grünwald (2019) put it "no sampling distribution can be specified that fully represents the variety of processes in accumulating scientific knowledge and all decision made along the way." (ter Schure & Grünwald, 2019, p. 3). These authors, however, also suggest that tests based on likelihood ratios, unlike tests based on *p*-value, can withstand accumulation bias.

Lastly, as we have already briefly discussed, small procedural changes in infant studies can sometimes be detrimental to the discoveries of true effects, and a peer review process that occurs before data collection can thus be of help in reducing the risk of unwanted variability in the results. In general, and most importantly, the ability of a meta-analysis to perform a reliable calculation is dependent upon the quality of the studies that have been combined. For this reason, to move forward we do not simply need to adopt a meta-analytic mindset, but we also need to start combining unbiased results, e.g., results of registered reports that likely are not as biased as non-preregistered studies, or at least, in the meta-analytic process, weigh differently results that are likely biased and results that are likely not biased (Tsuji, Cristia, Frank, & Bergmann, 2020).

In light of these concerns, as a way to reduce publication selection bias and obtain better estimates, as well as to react to the current situation and publication practices in many disciplines, a final and more radical suggestion could be to select only the most precise 10% of all the reported effects. If applied to Margoni and Surian (2018), this strategy would lead us to select seven studies that had 32 or more participants (i.e. approximately the top 10% of most precise or largest studies). With these, a random-effect meta-analysis would estimate that 0.61, 95% CI (0.50, 0.70), of infants would choose the prosocial agent. This 'Top10' strategy, according to its proponents, has the potential to approximate the truth better than conventional summary estimators because it heavily relies on the estimate precision, which is the inverse of the standard error (Stanley, Jarrell, & Doucouliagos, 2010). We may therefore combine the suggestion to run single studies (either original or replication studies) with bigger sample sizes, as this will produce more precise estimates, with a strategy that, at a meta-analytical level, in the face of a series of studies with under-powered tests published mainly because they reported statistically significant effects, suggests to choose only the most precise estimates at disposal. The more we conduct studies with adequate statistical power, the less we will need measures such as the Top10.

10. Conclusion

We have argued that the classical logic of replication, which relies on the concept of confirmation under the NHST assumptions, could be fruitfully substituted by a meta-analytic mindset and a focus on estimating effect sizes with confidence intervals. A single replication attempt is not sufficiently informative. To debate whether it is the replication or the original study that reports a 'true' finding often misses the point. We need to view each study as yielding an estimate of the unknown population parameter. The confidence intervals surrounding these estimates can often be extremely wide (possibly covering effects in both directions). Thus, even a few studies can be scarcely more informative than the original study. Indeed, sampling error alone can explain a large portion of variability between the results even before we introduce measurement error and subtle differences in experimental design. Therefore, it is unsurprising that when statistical tests are under-powered, outcomes that are commonly taken as disconfirmations should have actually been anticipated from the original result. Replication attempts should then simply be analyzed in combination with all the other studies. Only by pooling together our findings we can hope to get more accurate estimates of the phenomena of interest.

Acknowledgments

We would like to thank Luca Surian, Denis Tatone, Charlie Ebersole and Anne Scheel for their helpful comments on a previous version of the manuscript. Special thanks go to Martha Arterberry, Editor in Chief, who provided useful suggestions and assistance throughout the review process.

References

Abramson, L., Dar, M., Te'eni, A., & Knafo-Noam, A. (2016). *Preferences for helpers and hinders in 9- and 18-month-old infants*. Unpublished raw data.

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567,* 305-307. https://doi.org/10.1038/d41586-019-00857-9

Amrhein, V., Korner-Nievergelt, F., & Roth, T. (2017). The earth is flat (p > 0:05): significance thresholds and the crisis of unreplicable research. *PeerJ, 5,* e3544. https://doi.org/10.7717/peerj.3544

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., . . . Zuni, K. (2016). Response to Comment on Estimating the reproducibility of psychological science. *Science, 351,* 1037-1039. https://doi.org/10.1126/science.aad9163

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108-119. https://doi.org/10.1002/per.1919

Badenes-Ribera, L., Frias-Navarro, D., Iotti, N.O., Bonilla-Campos, A., & Longobardi, C. (2018). Perceived statistical knowledge level and self-reported statistical practice among academic psychologists. *Frontiers in Psychology, 9,* 996. https://doi.org/10.3389/fpsyg.2018.00996

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27,* 1069-1077. https://doi.org/10.1177/0956797616647519

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543-554. https://doi.org/10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43,* 666-678. https://doi.org/10.3758/s13428-011-0089-5

Baillargeon, R., Scott, R., He, Z., Sloane, S., Setoh, P., Jin, K.-S., . . . Lin, B. (2015). Psychological and sociomoral reasoning in infancy. In M. Mikulincer & P. R. Shaver (Eds.), E. Borgida & J. A. Bargh (Assoc. Eds.), *APA handbook of personality and social psychology: Attitudes and social cognition* (Vol. 1, pp. 79–150). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/14341-003

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

Bollen, K., Cacioppo, J., Kaplan, R., Krosnick, J., and Olds, J. (2015). Social, behavioral, and economic sciences perspectives on robust and reliable science: Report of the subcommittee on replicability in science. *Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences*. Retrieved from the National Science Foundation, available at www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley.

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8,* 5-18. https://doi.org/10.1002/jrsm.1230

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. https://doi.org/10.1038/nrn3475

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2,* 115-144. https://doi.org/10.1177/2515245919847196

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399. https://doi.org/10.17763/haer.48.3.t490261645281841

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex, 49,* 609-610. https://doi.org/10.1016/j.cortex.2012.12.016

Claerbout, J., & Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. *Proceedings of the 62nd Annual International Meeting of the Society of Exploration Geophysics*, New Orleans, USA, 25-29 October.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science*, *1*, 140216. https://doi.org/10.1098/rsos.140216

Cowell, J. M., & Decety, J. (2015). Precursors to morality in development as a complex interplay between neural, socioenvironmental, and behavioral facets. *Proceedings of the National Academy of Sciences of the United States of America, 112,* 12657-12662. http://dx.doi.org/10.1073/pnas.1508832112

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology, 4,* 28. https://doi.org/10.1186/s40359-016-0134-3

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. https://doi.org/10.1016/j.jesp.2015.10.002

Cristia, A., Tsuji, S., & Bergmann, C. (2020). Theory evaluation in the age of cumulative science. Pre-print retrieved from https://doi.org/10.31219/osf.io/83kg2

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7-29. http://dx.doi.org/10.1177/0956797613504966

Del Re, A. C. (2015). A practical tutorial on conducting meta-analysis in R. *The Quantitative Methods for Psychology, 11,* 37-50. https://doi.org/10.20982/tqmp.11.1.p037

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology, 50,* 2417-2425. https://doi.org/10.1037/a0037996

Duval, S., & Tweedie, R. (2000). Trim-and-fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455–463. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

Ellis, S. P., & Stewart, J. W. (2009). Temporal dependence and bias in meta-analysis. *Communications in Statistics—Theory and Methods, 38,* 2453-2462. https://doi.org/10.1080/03610920802562772

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE, 11,* e0149794. https://doi.org/10.1371/journal.pone.0149794

Fabrigar, L., & Wegener, D. (2017). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology, 66,* 68-80. https://doi.org/10.1016/j.jesp.2016.09.003

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891-904. https://doi.org/10.1007/s11192-011-0494-7

Fidler, F., & Loftus, G. (2009). Why figures with error bars should replace p values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology, 217,* 27-37. https://doi.org/10.1027/0044-3409.217.1.27

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., . . . & Goodman, O. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments, & Computers, 36,* 312-324. https://doi.org/10.3758/BF03195577

Findley, M. G., Jensen, N. M., Malesky, E. J., & Pepinsky, T. B. (2016). Can results-free review reduce publication bias? The results and implications of a pilot study. *Comparative Political Studies, 49,* 1667–1703. https://doi.org/10.1177/0010414016655539

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,* e109019. http://dx.doi.org/10.1371/journal.pone.0109019

Frank, M. C. (2019). Towards a more robust and replicable science of infant development. *Infant Behavior and Development, 57,* 101349. https://doi.org/10.1016/j.infbeh.2019.101349

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

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspective in Psychological Science, 9,* 641-651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Loken, E. (2014). Ethics and statistics: The AAA tranche of subprime science. *Chance, 27,* 51-56. https://doi.org/10.1080/09332480.2014.890872

Gelman, A., & Stern, H. (2006). The difference between significant and not significant is not itself statistically significant. *The American Statistician, 60,* 328-331. https://doi.org/10.1198/000313006x152649

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science, 351,* 1037-1037. https://doi.org/10.1126/science.aad7243

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine, 8,* 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Grand, J. A., Rogelberg, S. G., Banks, G. C., Landis, R. S., & Tonidandel, S. (2018). From outcome to process focus: Fostering a more robust psychological science through registered reports and results-blind reviewing. *Perspectives on Psychological Science, 13,* 448-456. https://doi.org/10.1177/1745691618767883

Hamlin, J. K. (2015). The case for social evaluation in preverbal infants: gazing toward one's goal drives infants' preferences for Helpers over Hinderers in the hill paradigm. *Frontiers in Psychology, 5,* 1563. http://dx.doi.org/10.3389/fpsyg.2014.01563

Hamlin, J. K. (2016). *A box show replication attempt*. Unpublished raw data.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development, 26,* 30-39. http://dx.doi.org/10.1016/j.cogdev.2010.09.001

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450,* 557–559. http://dx.doi.org/10.1038/nature06288

Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods, 24,* 557-570. https://doi.org/10.1037/met0000189

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557-560. http://dx.doi.org/10.1136/bmj.327.7414.557

Hinde, J., & Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, *27*, 151-170. https://doi.org/10.1016/s0167-9473(98)00007-3

Holvoet, C., Scola, C., Arciszewski, T., & Picard, D. (2016). Infants' preference for prosocial behaviors: A literature review. *Infant Behavior and Development, 45,* 125-139. https://doi.org/10.1016/j.infbeh.2016.10.008

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640-648. https://doi.org/10.1097/ede.0b013e31818131e7

Ioannidis, J. P. (2012). Scientific inbreeding and same-team replication: type D personality as an example. *Journal of Psychosomatic Research, 73,* 408-410. https://doi.org/10.1016/j.jpsychores.2012.09.014

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

Johnson, E., & Zamuner, T. (2010). Using infant and toddler testing methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Language learning & language teaching: Vol. 27. Experimental methods in language acquisition research* (pp. 73-93). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.27.06joh

Johnson, P., Barry, S., Ferguson, H., & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution, 6*, 133-142. https://doi.org/10.1111/2041-210x.12306

Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology* (pp. 83-105). Malden, MA: Blackwell.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., . . . & Cemalcilar, Z. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45,* 142-152. https://doi.org/10.1027/1864-9335/a000178

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: APA Books.

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23-31. https://doi.org/10.1037//1082-989x.3.1.23

Krantz, D. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *94*, 1372-1381. https://doi.org/10.2307/2669949

Krzywinski, M., & Altman, N. (2013). Points of significance: Power and sample size. *Nature Methods, 10,* 1139-1140. https://doi.org/10.1038/nmeth.2738

Kulinskaya, E., Huggins, R., & Dogo, S. H. (2016). Sequential biases in accumulating evidence. *Research Synthesis Methods, 7,* 294-305. https://doi.org/10.1002/jrsm.1185

Lakens, D., & Evers, E. (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. https://doi.org/10.1177/1745691614528520

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1,* 259-269. https://doi.org/10.31234/osf.io/v3zkt

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology, 113,* 254-261. http://dx.doi.org/10.1037/pspi0000106

Lin, L. (2019). Comparison of four heterogeneity measures for meta-analysis. *Journal of Evaluation in Clinical Practice,* 1-9. https://doi.org/10.1111/jep.13159

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine, 20,* 641-654. https://doi.org/10.1002/sim.698

Makel, M., Plucker, J., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537-542. https://doi.org/10.1177/1745691612460688

Maner, J. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science*, *9*, 343-351. https://doi.org/10.1177/1745691614528215

Margoni, F., Baillargeon, R., & Surian, L. (2018). Infants distinguish between leaders and bullies. *Proceedings of the National Academy of Sciences of the United States of America, 115,* E8835–E8843. https://doi.org/10.1073/pnas.1801677115

Margoni, F., & Surian, L. (2018). Infants' evaluation of prosocial and antisocial agents: A meta-analysis. *Developmental Psychology, 54,* 1445-1455. https://doi.org/10.1037/dev0000538

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods, 9,* 147-163. https://doi.org/10.1037/1082-989x.9.2.147

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology, 9,* 2. https://doi.org/10.1186/1471-2288-9-2

Morey, R. D., & Lakens, D. (2016). *Why most of psychology is statistically unfalsifiable*. Manuscript in preparation. Retrieved from https://raw.githubusercontent.com/richarddmorey/psychology_resolution/master/paper/response.pdf

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69,* 511-534. https://doi.org/10.1146/annurev-psych-122216-011836

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5,* 241-301. https://doi.org/10.1037/1082-989x.5.2.241

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

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

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy, 22,* 436–469. http://dx.doi.org/10.1111/infa.12186

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. https://doi.org/10.1126/science.aac4716

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

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science, 11,* 539-544. https://doi.org/10.1177/1745691616646366

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*, 319-332. https://doi.org/10.1177/1745691614528519

Plesser, H. E. (2018). Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in Neuroinformatics*, *11*, 76. https://doi.org/10.3389/fninf.2017.00076

Popper, K. R. (1934/1959). *The logic of scientific discovery*. London: Hutchinson.

Raggio, E., Hendi, S. F., Modesti, C., Presaghi, F., & Nicolais, G. (2015). Temperament and attachment in the development of moral precursors: Preliminary data. *Infanzia e Adolescenza, 14,* 197-217.

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin, 86,* 638-641. https://doi.org/10.1037//0033-2909.86.3.638

Salvadori, E., Blazsekova, T., Volein, A., Karap, Z., Tatone, D., Mascaro, O., & Csibra, G. (2015). Probing the strength of infants’ preference for helpers over hinderers: Two replication attempts of Hamlin and Wynn (2011). *PLoS ONE, 10,* e0140570. http://dx.doi.org/10.1371/journal.pone.0140570

Scarf, D., Imuta, K., Colombo, M., & Hayne, H. (2012). Social evaluation or simple association? Simple associations may explain moral reasoning in infants. *PLoS ONE, 7,* e42698. http://dx.doi.org/10.1371/journal.pone.0042698

Schauer, J. M., & Hedges, L. V. (2020). Reconsidering statistical methods for assessing replication. *Psychological Methods.* Advance online publication. http://dx.doi.org/10.1037/met0000302

Schlingloff, L., Csibra, G., & Tatone, D. (2020). Do 15-month-old infants prefer helpers? A replication of Hamlin et al. (2007). *Royal Society Open Science, 7,* 191795. https://doi.org/10.1098/rsos.191795

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181. https://doi.org/10.1037//0003-066x.47.10.1173

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129. https://doi.org/10.1037//1082-989x.1.2.115

Schmidt, F., & Hunter, J. (2014). *Methods of meta-analysis: Correcting error and bias in research findings.* Sage Publications.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician, 55,* 62-71. https://doi.org/10.1198/000313001300339950

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. https://doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9,* 76-80. https://doi.org/10.1177/1745691613514755

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26,* 559-569. https://doi.org/10.1177/0956797614567341

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General, 143,* 534-547. http://dx.doi.org/10.1037/a0033242

Spence, J., & Stanley, D. (2016). Prediction interval: What to expect when you're expecting . . . a replication. *PloS ONE, 11,* e0162874. https://doi.org/10.1371/journal.pone.0162874

Stanley, D., Jarrell, S. B., & Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician, 64,* 70-77. https://doi.org/10.1198/tast.2009.08205

Stanley, D., & Spence, J. (2014). Expectations for replications: are yours realistic? *Perspectives on Psychological Science, 9,* 305-318. https://doi.org/10.1177/1745691614528518

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance. *Journal of the American Statistical Association, 54,* 30-34. https://doi.org/10.2307/2282137

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 439,* 108-112. https://doi.org/10.1080/00031305.1995.10476125

Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false beliefs to a geometric shape at 17 months. *British Journal of Developmental Psychology, 30,* 30-44. https://doi.org/10.1111/j.2044-835x.2011.02046.x

Surian, L., & Margoni, F. (2020). First steps toward an understanding of procedural fairness. *Developmental Science,* e12939. https://doi.org/10.1111/desc.12939

Szucs, D., & Ioannidis, J. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15,* e2000797. https://doi.org/10.1371/journal.pbio.2000797

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22,* 2113-2126. https://doi.org/10.1002/sim.2049 ter Schure, J. A., & Grünwald, P. D. (2019). Accumulation Bias in meta-analysis: the need to consider time in error control. *F1000Research, 8,* 962. https://doi.org/10.12688/f1000research.19375.1

Ting, F., Dawkins, M. B., Stavans, M., & Baillargeon, R. (2019). Principles and concepts in early moral cognition. In J. Decety (Ed.), *The social brain: A developmental perspective*. Cambridge, MA: MIT Press.

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

Tsuji, S., Cristia, A., Frank, M. & Bergmann, C. (2020). Addressing publication bias in meta-analysis: Empirical findings from community-augmented meta-analyses of infant language development. *Zeitschrift für Psychologie, 228,* 50-61. https://doi.org/10.1027/2151-2604/a000393

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., ... & Nosek, B. A. (2019). Scientific utopia III: Crowdsourcing science. *Perspectives on Psychological Science, 14,* 711-733. https://doi.org/10.1177/1745691619850561

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

van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology, 67,* 2-12. https://doi.org/10.1016/j.jesp.2016.03.004

Velicer, W. F., Cumming, G., Fava, J. L., Rossi, J. S., Prochaska, J. O., & Johnson, J. (2008). Theory testing using quantitative predictions of effect size. *Applied Psychology: An International Review, 57,* 589-608. https://doi.org/10.1111/j.1464-0597.2008.00348.x

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36,* 1-48. https://doi.org/10.18637/jss.v036.i03

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

Williamson, P., Gamble, C., Altman, D., & Hutton, J. (2005). Outcome selection bias in meta-analysis. *Statistical Methods in Medical Research, 14,* 515-524. https://doi.org/10.1191/0962280205sm415oa

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences, 117,* 5559-5567. https://doi.org/10.1073/pnas.1914237117

Woo, B. M., & Hamlin, K. (2016). *Older, but not younger, babies prefer helpers in a cartoon version of the hill paradigm*. Manuscript in preparation.

Woznyj, H. M., Grenier, K., Ross, R., Banks, G. C., & Rogelberg, S. G. (2018). Results-blind review: a masked crusader for science. *European Journal of Work and Organizational Psychology, 27,* 561-576. https://doi.org/10.1080/1359432x.2018.1496081

Young, N. S., Ioannidis, J. P., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine, 5,* e201. https://doi.org/10.1371/journal.pmed.0050201

Supplementary Materials

The following R function computes 95% prediction intervals (PIs) for proportions. It needs as inputs: x = count of positive events or successes; n1 = number of observations from which the proportion was calculated; n2 = what n will be for the proposed replication. The calculation makes use of the Freeman-Tukey (double arcsine) transformation and Miller's back transform (Freeman & Tukey, 1950; Miller, 1978), built into escalc which is part of Viechtbauer's metafor package (Viechtbauer, 2010).

```
library(metafor)

pi.prop <- function(x, n1, n2 = NA)

{
if (is.na(n2)) {
n2 <- n1
}

pi <- x / n1 #compute proportion

es1 <- escalc(measure = "PFT", xi=x, ni=n1)

temp1.ci <- summary.escalc(es1)
lb1 <- temp1.ci$ci.lb
ub1 <- temp1.ci$ci.ub
y1 <- temp1.ci$yi

es2 <- escalc(measure = "PFT", xi=round((pi*n2),0), ni=n2)
temp2.ci <- summary.escalc(es2)
lb2 <- temp2.ci$ci.lb
ub2 <- temp2.ci$ci.ub

pi.low <- y1 - sqrt(((y1-lb1)^2) + ((ub2-y1)^2) )
pi.high <- y1 + sqrt(((y1-lb2)^2) + ((ub1-y1)^2) )

rep.lo <- transf.ipft(pi.low, n2) # Back transform
rep.hi <- transf.ipft(pi.high, n2)

return(c(rep.lo,rep.hi))
}
```