Does neonatal imitation exist? Insights from a meta-analysis of 336 effect sizes

Jacqueline Davis[1], Jonathan Redshaw[2], Thomas Suddendorf[2], Mark Nielsen[2,3], Siobhan Kennedy-Costantini[2,4], Janine Oostenbroek[2], & Virginia Slaughter[2]*

MANUSCRIPT IN PRESS: *Perspectives on Psychological Science*

1. Department of Psychology, University of Cambridge, UK
2. School of Psychology, University of Queensland, Australia
3. Faculty of Humanities, University of Johannesburg, South Africa
4. School of Psychology, University of Auckland, New Zealand

* Correspondence: vps@psy.uq.edu.au

Author contributions: JD designed the project, collected effect sizes, ran all analyses and wrote the manuscript. JR contributed to project design and wrote the manuscript. MN, TS, SK-C, JO contributed to writing the manuscript. VS contributed to project design and wrote the manuscript. All authors approved the final version of the manuscript for submission.

Acknowledgements: We sincerely thank the researchers who provided unpublished data for this study. We also acknowledge Alan Fridlund, Bill von Hippel, Cecilia Heyes, Susan Jones and an anonymous reviewer who provided insightful comments on an earlier version of this manuscript.

Abstract

Neonatal imitation is a cornerstone in many theoretical accounts of human development and social behavior, yet its existence has been debated for the last 40 years. To examine possible explanations for the inconsistent findings in this literature, we conducted a multilevel meta-analysis synthesizing 336 effect sizes from 33 independent samples of human newborns, reported in 26 papers. The meta-analysis found significant evidence for neonatal imitation ($d = 0.68$, 95% CI = 0.39 to 0.96, $p < .001$), but substantial heterogeneity between study estimates. This heterogeneity was not explained by any of thirteen methodological moderators identified by previous reviews, but it was associated with researcher affiliation, $QM(15) = 57.09$, $p < .001$. There are at least two possible explanations for these results: (1) neonatal imitation exists and its detection varies as a function of uncaptured methodological factors common to a limited set of studies, and (2) neonatal imitation does not exist and the overall positive result is an artefact of high researcher degrees of freedom.

Does neonatal imitation exist? Insights from a meta-analysis of 336 effect sizes

Neonatal imitation, or the capacity of newborns to flexibly copy others' actions (Meltzoff & Moore, 1977), is one of the most controversial phenomena in all of psychological science. Despite its status as a cornerstone in prominent theories of social cognitive development (e.g., Meltzoff, 2007; Meltzoff & Decety, 2003; Nadel & Butterworth, 1999; Trevarthen & Aitken, 2001), many alternative accounts deny its existence altogether (e.g., Anisfeld, 1996; Heyes, 2016a; Jones, 2009). The controversy shows no signs of abating: Following a recent high-profile target article arguing that neonatal imitation of tongue protrusion is physiologically unlikely (Keven & Akins, 2017), 10 open peer commentaries agreed that the phenomenon probably does not exist, whereas 6 commentaries remained committed to its existence and 5 were noncommittal.[1] At the heart of the disagreement is a highly heterogeneous literature, with some studies reporting large imitation effects in newborns and others reporting no effects at all (for previous reviews, see Anisfeld, 1991; Oostenbroek et al., 2013; Ray & Heyes, 2011; Simpson, Paukner, Suomi, & Ferrari, 2014). Devising means to empirically decide between possible reasons for this heterogeneity is, as in many fields of psychology, a matter of utmost importance (see Zwaan, Etz, Lucas, & Donnellan, 2017).

Debates about inconsistencies in the infant psychology literature may seem particularly intractable. On one hand, infants' perceptual-cognitive limitations and poor state regulation may heighten sensitivity to minor differences in experimental procedures, meaning that failed replications of true effects may have methodological sources (Coyne, 2016). On the other hand, infant behavior can be highly ambiguous, meaning that researchers may selectively analyze and report behaviors that produce interpretable results – and journal editors and reviewers may also look more favorably upon such findings. Notably, however, underlying these problems are testable hypotheses about the pattern of results in the overall literature. If methodological

---

[1] *For neonatal imitation*: Aitken; Buck; Desseilles; Meltzoff; Murray et al.; Simpson et al.
Against neonatal imitation: Beisart et al.; Campos et al.; Fitch; Jones; Kennedy-Costantini et al.; Leisman; Libertus et al.; O'Sullivan et al.; Provine; Zapperttini.
Noncommittal: Booth; Casartelli et al.; Choi et al.; Mayer et al.; Vincini et al.

variation is driving inconsistency of results, then effects should vary systematically as a function of the critical methodological factors. Alternatively, if publication bias is driving inconsistency, then the literature should show indicators of suppression of null and negative effects.

Here we conduct a systematic review and meta-analysis of neonatal imitation, with the primary aim of objectively assessing its degree of empirical support. We also conduct tests assessing (i) a comprehensive set of potential methodological moderators of the neonatal imitation effect (as identified by previous authors), and (ii) publication bias favoring positive results.

Neonatal Imitation: History and Controversy

The capacity to imitate is often placed at the foundation of human social interaction (e.g., Meltzoff, 2007; Nadel & Butterworth, 1999). We like people who imitate us and we imitate people whom we like (Chartrand & Bargh, 1999; Lakin et al., 2003). Children's tendency to faithfully copy other people's object-directed actions underpins cultural learning (Nielsen & Tomaselli, 2010; Whiten, McGuigan, Marshall-Pescini, & Hopper, 2009) and contributes to the enormous diversity of behavioral traditions observed across human groups (Legare & Nielsen, 2015). This tendency is not observed in humans' closest living great ape relatives, whose social learning typically involves emulating others' instrumental outcomes rather than copying specific action sequences, especially those with limited functional utility (Clay & Tennie, 2017; Horner & Whiten, 2005). A capacity for imitation, therefore, is often put forward as a leading candidate in explaining the uniqueness of human social behavior and the extraordinary achievements of our cumulative culture (e.g., Boyd & Richerson, 1996; Henrich & McElreath, 2003; Legare & Nielsen, 2015; Tomasello, Kruger, & Ratner, 1993).

The ubiquity and significance of imitation in human social life prompts the question of whether the capacity is innate in our species. If newborns can imitate, it would imply that humans are born with a neurological solution to the 'correspondence problem' of matching and coordinating one's own motor behaviors with those of another (Brass & Heyes, 2005). Some have suggested that mirror neurons provide us with this inborn solution (Lepage & Theoret, 2007; Meltzoff & Decety, 2003), although that account has not gone unchallenged (Hickok, 2014). Alternatively, absence of neonatal imitation would imply that the capacity emerges later, perhaps as a product of both innate and environmental factors (Ray & Heyes, 2011; Heyes, 2016b; Jones, 2017).

Forty years ago, a striking finding appeared to have settled this question. Directly contradicting dominant theories of the time (e.g., Piaget, 1962; Uzgiris & Hunt, 1975), Meltzoff and Moore (1977) reported that human infants could copy adults' facial and manual gestures in the initial weeks of life. This finding was followed by a similar high-impact finding five years later, suggesting that neonates could reliably imitate facial expressions even in their first few days (Field et al., 1982). Consequently, it was generally concluded that human infants do possess an inbuilt solution to the correspondence problem - namely, a pre-wired representational system that links their own and others' actions (Meltzoff, 2007).

The conclusion that neonates can imitate has been incorporated into influential theories of human development, social psychology and neuroscience. In developmental psychology, neonatal imitation is argued to form the basis of a lifelong capacity to reproduce others' actions with high fidelity, which in turn supports the grand scale of human social learning and cultural evolution (Meltzoff & Moore, 1997; Nadel & Butterworth, 1999). In social psychology, neonatal imitation has been put forward as a precursor to 'nonconscious mimicry' (Cheng & Chartrand, 2003; Lakin et al., 2003), whereby adults imitate the actions of others without awareness of having done so (e.g., Chartrand & Bargh, 1999; van Baaren, Holland, Kawakami, & Van Knippenberg, 2004). In the neurosciences, the phenomenon has become intricately associated with the existence and function of a mirror neuron system that has been proposed to provide the neurological substrate for representing self-other equivalences (e.g., Gallese, 2001; Nagy & Molnar, 2004; Simpson et al., 2014). Comparative psychological research has suggested that non-human primate neonates may also imitate facial actions (e.g., Ferrari et al., 2006, 2009; Myowa-Yamakoshi, Tomonaga, Tanaka, & Matsuzawa, 2004) – and that this reflects the same neuronal and cognitive mechanisms as the human phenomenon (Simpson et al., 2014; Iacobini, 2009) – even given the striking differences between the styles of humans and other primates in object-directed imitation and cultural learning (see Tennie, Call & Tomasello, 2009). Indeed, neonatal imitation has been included as an essential component of several influential integrative theories of social cognition and behavior (e.g., Meltzoff & Decety, 2003; Trevarthen & Aitken, 2001).

Despite its heavy impact, the phenomenon of neonatal imitation has proven empirically unreliable, with many reported failures to replicate the basic finding (for a critical review, see Ray & Heyes, 2011). Recently, the largest-ever longitudinal study of human neonatal imitation (Oostenbroek et al., 2016) reported that 106 neonates failed to imitate any of nine gestures at any of four time points between one and nine weeks of age. Although this study was originally aimed at uncovering whether neonatal imitation predicted later social cognitive capacities (see Redshaw et al., 2019; Suddendorf, Oostenbroek, Nielsen, & Slaughter, 2013), the results have been interpreted as a compelling and perhaps fatal challenge to the existence of the phenomenon itself (Heyes, 2016b). In response, a group of 13 prominent neonatal imitation researchers published a detailed critique, arguing that the failure to replicate could be explained by specific methodological choices (Meltzoff et al., 2017; see Table 1). This piece echoed many points made previously about failures to replicate neonatal imitation effects (e.g., Meltzoff & Moore, 1983a; Vincini, Jhang, Buder, & Gallagher, 2017). Be that as it may, the credibility of post hoc methodological critiques of specific null results has recently come under criticism in the context of the broader replication crisis within psychology. As Zwaan et al. (2017) write: "uncritical acceptance of post hoc context-based explanations of failed replications ignores the possibility that false positives … ever exist and seems to irrationally privilege the chronological order of studies over the objective characteristics of those studies". A critical question now looming over neonatal imitation is, therefore: Can the replication failures be explained in terms of their methodological variations?

Here we use meta-analytic techniques to empirically evaluate this question. In a meta-analysis framework, inconsistencies between studies are captured as statistical heterogeneity, or variance, in study results. The sources of such heterogeneity, including methodological differences between studies, may subsequently be identified via tests of factors that moderate the overall effect size (Higgins & Green, 2011).

Possible Methodological Moderators of the Neonatal Imitation Effect

Although it is impossible to capture all of the factors that might lead some laboratories to obtain positive neonatal imitation effects, many potential methodological sources of heterogeneity have been proposed in the neonatal imitation literature. For instance, a reply by Meltzoff and Moore (1983a) to the failed replications of Koepke et al. (1983) and McKenzie and Over (1983) identifies a list of procedural items that did not appear in the foundational study (Meltzoff & Moore, 1977) but are potentially critical to finding the effect. Other potential factors are explored in narrative review papers by Oostenbroek et al. (2013) and Simpson et al. (2014). Most recent are Meltzoff et al.’s (2017) critique, which identifies several methodological features that may have biased Oostenbroek et al.’s (2016) study toward null results, and Vincini et al.’s (2017) review of the neonatal imitation literature and recommendations for optimal experimental design. These five sources offer a comprehensive list of possible methodological sources of heterogeneity previously identified in the neonatal imitation literature (summarized in Table 1).

The complete list of potential moderators is impractically long, and it is also unclear how important each of the variables is, to overall study quality. In general, not all study quality indicators are equal, and badly implemented studies are not necessarily less accurate than well-implemented ones (Glass, 1976). In the case of neonatal imitation, some of the “best practice” recommendations directly contradict each other in their explanations for why previous replication attempts failed. For example, in a response to Koepke’s et al.’s (1983) failure to replicate, Meltzoff and Moore (1983a) write that infants should not be exposed to the experimenter before the start of data collection, because they may not be as attentive to a face they have seen before (a point later reiterated by Meltzoff et al., 2017; and Simpson et al., 2014). On the other hand, Vincini et al (2017) write that infants should be given a preliminary familiarization phase where the experimenter can seek the optimal posture and conditions for that infant. As a compromise, we selected for analysis all criteria that were mentioned in *at least two* separate sources.

+++++++++++Insert Table 1 about here.+++++++++++++++

In addition to collating potential moderators from the neonatal imitation literature, we also examined relevant criteria from a more general source, the TREND statement on the quality of nonrandomized studies (Des Jarlais, Lyles, Crepaz, & TREND Group, 2004). The TREND statement was designed to support research synthesis efforts by providing a list of general study design elements that might moderate study results. Many of the moderators that we ultimately selected for the meta-analysis are also broadly mentioned in this statement, such as the duration and number of times exposed to the treatment, the setting of data collection, the identity of the person delivering the treatment, the unit of statistical analysis, and the number of participants included in the analysis.

Exploratory potential moderator: Researcher allegiance. The neonatal imitation literature is replete with contradictory findings. Studies of this phenomenon have been conducted by researchers from over a dozen institutions across several countries, and it may be that some research groups are systematically more likely to report positive results than others. To examine this possibility, we therefore included the institutional affiliation of the corresponding author as an additional moderator.

This analysis is in line with increasing scrutiny of the role of researcher allegiance to a particular method or to a particular finding, in psychology (Boccaccini, Marcus, & Murrie, 2017) and in related fields (Singh, Grann, & Fazel, 2013; Manea et al., 2016; Munder et al., 2013;

Dragioti, Dimoliatis, & Evangelou, 2015). The “researcher allegiance effect” refers to associations between researchers’ prior interest in finding a particular result, and their research outcomes. Researcher allegiance is most frequently documented for founders of a research field. For example, in psychotherapy, one review found zero studies published by a treatment founder, in which the results failed to support the founder’s treatment (Luborsky et al., 1999), and in criminology, risk assessment instruments were found to be twice as effective in papers published by the instrument’s author than in papers published by other authors (Singh et al., 2013). Other analyses have revealed correlations between the strength of researchers’ endorsement of a therapeutic technique, and the size of the effects they find in RCTs (see Leykin & DeRubeis, 2009). Increasingly, meta-analyses attempt to calculate an effect size for researcher allegiance, with several meta-analyses finding a substantial association between researcher allegiance and study outcome (Munder et al., 2013).

The mechanisms underlying researcher allegiance effects are debated (Boccaccini et al, 2017). In some instances, these effects may reflect researcher biases which could influence all aspects of the research from study design and reporting practices to intangible aspects of data collection such as enthusiasm, fidelity, and attention to detail. Such biases are ambiguous with respect to causality: researchers’ beliefs could be based on their observations of the efficacy of certain methods or the strength of certain effects, and/or their pre-existing beliefs might contribute to producing those outcomes (Leykin & DeRubeis, 2009). Alternatively, systematic differences in research groups’ tendencies to find significant effects may be attributable to undocumented variance in methodological practices across laboratories, without those research groups being biased toward particular outcomes. In our meta-analysis, we examine the exploratory moderator of researcher allegiance, operationalised as institutional affiliation, without making assumptions about underlying mechanisms. The procedure for coding this moderator variable is presented in the Supplementary Material.

Sample Size and Publication Bias

Sample size has been identified as a specific problem for neonatal imitation, and is also important for meta-analysis more generally. Meltzoff and Moore (1983a), for example, suggest that Koepke et al. (1983) should have increased their sample size of 6 infants (which was chosen to match Meltzoff & Moore, 1977, Study 1) in order to make their overall null effects more interpretable (also see Simpson et al., 2014). Vincini et al. (2017), on the other hand, suggest that studies should aim for a sample size of no more than 26-30 infants, as when samples are larger than this, experimenters might neglect to pay proper attention to the testing conditions for each infant. Part of the reason that neonatal imitation researchers focus on sample size is that the presence or absence of neonatal imitation is usually decided by statistical significance. If a result is significant, then neonatal imitation is said to be present; but if a result is not significant, then researchers may be encouraged to increase their sample size to maximize the chances of finding a positive result (Gelman & Loken, 2013). At its extreme, this tendency may result in the suppression of smaller studies with non-significant or negative results (Rosenthal, 1979; Rothstein et al., 2006).

Meta-analysis techniques such as funnel plots can detect whether there is a relationship between effect size and sample size (Egger, Smith, Schneider, & Minder, 1997). Such a relationship could be unproblematic if sample size was correlated with the presence of a true methodological moderator of the effect. For instance, if Meltzoff et al. (2017) are correct that unfamiliar experimenters are the "key" to finding neonatal imitation effects (also see Meltzoff & Moore, 1983a), and if unfamiliar experimenters are more common in either smaller or larger studies, then one would expect to see a spurious relationship between effect size and sample size. Another, more problematic reason for such a relationship, however, is publication bias (Sullivan & Feinn, 2012; Nakagawa & Cuthill, 2007). Publication bias traditionally refers to a general preference for publishing positive results over null and negative results (e.g., Rosenthal, 1979), but has become an umbrella term for the systematic tendency for positive results to be reported by researchers, accepted for publication by journals, and cited and disseminated by readers

(Rothstein et al., 2006). Publication bias has recently been detected in many fields of research, including in psychological science (Ferguson & Brannick, 2012), but has not yet been tested in the neonatal imitation literature. As part of our meta-analysis, therefore, we conducted funnel plot analyses to test for significant heterogeneity of neonatal imitation effects as a specific function of sample size (Egger et al., 1997).

Method

Our systematic review and meta-analysis was conducted according to the best-practice guidelines of the Cochrane Collaboration (Higgins & Green, 2011), including an *a priori* systematic review protocol and analysis plan. Our review protocol and details of our statistical analyses are available in full in the Supplementary Materials, and we summarize them here.

Systematic Search Method and Coding Protocol

We undertook a systematic online search, an on-search of references to relevant papers, and made direct solicitations for unpublished data. We discontinued our literature search in September 2019.

Systematic search strategy

Keyword formulation. The search keywords included terms relevant to the predictor, the outcome, and the population. Both general outcome terms (e.g. "imitation") and specific gestures (e.g. "tongue protrusion") were used. Thus, each search query contained elements of both (1) "imitation" or "tongue protrusion" or "mouth opening" or "lip smacking", *and* (2) "neonates" or "infants" or "children". We also searched the reference sections of previous reviews on neonatal imitation and the reference sections of included studies, and we requested unpublished datasets via a developmental psychology email list.

Search field. All searches were limited to abstracts where this option was available.

Database selection. We selected electronic databases that included relevant journals in the field of child development, infant psychology, and related fields; and on the basis of source overlap and usability. Additionally, we searched sources of unpublished literature, such as dissertations and theses, that may include null or negative results. The data sources selected were: ProQuest (including ProQuest Dissertations and Theses), Scopus, PsycInfo (including PsycArticles), and Cambridge University Library and Dependent Libraries Catalogue. Additionally, we searched Google Scholar using a more stringent set of search keywords.

Criteria for including and excluding studies.

Types of study designs. Repeated measures and crossed design studies were included. Cross-sectional and longitudinal designs were included, with appropriate statistical controls for dependency.

Types of participants. Following previous reviews, the current review only included data from human participants equal to or younger than 6 weeks at the time of testing. Data from children older than 6 weeks and non-human primates of any age were excluded. Only data from typically developing children were included; clinical groups and those selected on the basis of extreme values were excluded. Data from non-human primates were also excluded, although note that a robust statistical analysis of the entire published data set ($N = 163$) of macaque neonatal imitation studies found no significant evidence for neonatal imitation (Redshaw, 2019).

Types of imitation scores. The review included imitation scores, as measured using the following methods: (1) comparing gestures in response to a modelled behavior to gestures in response to a control behavior (active control), and (2) comparing gestures in response to a modelled behavior to gestures in a baseline condition (baseline control). A few studies where coders watched blinded videos of infant behavior and guessed the modelled action from a list including the target and one or more controls (best guess) were also included.

Infant responses. All types of measured infant responses were included, such as oral gestures (e.g., tongue protrusion and mouth opening) manual gestures (e.g., finger pointing), facial gestures (e.g., happy expressions), and verbalizations (e.g., *mmm* sounds). No types of measured responses were excluded.

Modelled actions. The review included infants’ responses to all types of modelled actions, including those that infants could in principle imitate (i.e., oral gestures, facial gestures, manual gestures, and verbalizations) and those they could not in principle imitate (e.g., object manipulations such as box opening, where the infant had no access to the box). Some studies included data for more than one statistical comparison per measured infant action – for example, infants' levels of tongue protrusions in response to the tongue protrusion model may have been compared to their levels of tongue protrusions in response to multiple non-matching control models. In these cases, effect sizes from *each* of the multiple comparisons were included in the meta-analysis, with statistical dependency controlled for using multilevel methods.

Types of settings. The review included data from all research settings, including home settings, hospitals, or research laboratories.

Data reporting. Studies were included if they reported sufficient information to allow an effect size and its variance to be calculated. Where sufficient data were not provided, we used next-best methods to estimate the likely effect size (e.g., back calculating from *p* values and *N*s). Studies reporting only "not significant" results, with no statistics, were not included.

Extracting and coding research for the review.

Identification of potentially relevant material for meta-analysis. Results of the systematic search were downloaded into a Zotero dataset (a bibliography management software program) and exported to Microsoft Excel for initial filtering. These results were assessed on the basis of title and abstract for relevance to the subject of the systematic review and for whether or not the article contained empirical data.

Coding of literature. The remaining articles had the full-text PDF downloaded and attached to the corresponding Zotero entry. These were then read in full and coded according to a detailed spreadsheet in Microsoft Excel. First, studies were assessed for eligibility according to the criteria detailed above. Studies deemed ineligible according to these criteria were not coded further. Eligible studies were coded fully, including all relevant information requested by the spreadsheet, as follows:

- Eligibility checklist
- Search information
- Reference information (e.g. authors, publication type)

And where the document was eligible for inclusion in the systematic review:

- Infant gesture
- Modelled gesture
- Participant age
- Data for effect size calculation
- Research design notes
- Imitation scoring method (control method or baseline method)
- Infant posture (held by mother, held by experimenter, not held)

Treatment of qualitative research. Qualitative studies were not included in the meta-analysis but were used to inform the literature review and interpretation of results.

Test of inter-rater reliability. All studies were coded by JD. However, to avoid potential bias, a subset of studies was given to a second rater (JO) who assessed them for eligibility according to the above criteria, and coded them according to the spreadsheet. The two coders had 100% agreement on overall eligibility for the review, and 96% agreement on all criteria.

Analyses

Average overall effect for neonatal imitation. The average overall effect for neonatal imitation was calculated using multilevel meta-analysis of the imitation effect sizes. Many studies reported effects for more than one gesture (for example, tongue protrusion and mouth opening) or more than one independent sample (for example, infants at 1 week and 3 weeks of age). Correlations between these study results were accounted for using a multilevel meta-analysis model (Higgins, Deeks, & Altman, 2011), implemented using the *metafor* package (Viechtbauer, 2010) in *R* statistical software. The multilevel meta-analysis provided an overall effect size estimate (*d*) for neonatal imitation across all samples and gestures; a 95% confidence interval for *d*; an overall *p* value for *d*; and parameters for heterogeneity at the within-study ($\sigma^2{}_1$) and between-study ($\sigma^2{}_2$) levels. As part of the meta-analysis procedure, summary estimates for the study-level effect sizes were also calculated, with corresponding variance estimates.

Explanations for inconsistencies in study results.

Methodology. We tested fourteen possible methodological moderators of the neonatal imitation effect, thirteen of which are variations of the factors identified by multiple reviews as being potentially important (see Table 1 in the introduction): (1) length of model presentation time per burst, (2) length of infant response time per burst, (3) total model presentation time (per model), (4) total infant response time (per model), (5) total number of actions modelled, (6) total active experiment time, (7) experimental setting, (8) modeler identity, (9) level of pre-experimental exposure to the modeler's face, (10) infant testing position, (11) whether or not the infant was secured in a padded seat during testing, (12) measure of infant alertness, and (13) statistical analysis method. We also examined the additional moderator of (14) researcher affiliation (i.e., the listed institution of the corresponding author). Two coders (JD and JR) independently scored each study on all fourteen dimensions, and disagreements (15.1% of cases) were resolved by discussion.

All methodological variables were tested as statistical moderators of the imitation effect size, using meta-regression (Higgins & Green, 2011). A separate meta-regression model was tested for each variable, since including all the variables in a single model would have over-specified the meta-regression relative to the number of available effect sizes. The meta-regression provided coefficients for the impact of each methodological variable on the overall effect size (b_{meta}), standard errors and *p* values for the coefficients, and for categorical moderators, an estimate of the relative impact of each category within the moderator.

Publication bias. Meta-analyses are vulnerable to skewed results if studies that report large effects and/or statistically significant results are published, and therefore accessed, at a higher rate than studies that report smaller effects or null results (Sutton et al., 2000). Funnel plots can be used to diagnose publication bias where other sources of heterogeneity have already been considered (Ioannidis, 2008). Funnel plots compare study effect sizes to variance, where variance is calculated from study sample size and study error, which can arise from measurement error or statistical mis-specification. An unbiased literature should show no relation between study effect size and study variance. A series of inverse variance funnel plots and corresponding regression tests (Egger et al., 1997) was conducted to assess whether the effect sizes differed for small and large studies. Full technical details of the method for detecting publication bias are included in the Supplementary Material.

Results

Results of the Systematic Search

The systematic search identified 100 provisionally eligible sources according to the title, abstract and keywords of the article. Of these, 76 studies had no comparative data, did not report on imitation on closer inspection of the full text (e.g., Jones 1996), did not include infants under 6 weeks of age (e.g., Kuhl & Meltzoff, 1982; Jones, 2007), did not include any comparison condition (e.g., Nagy & Molnar, 2004), or did not report sufficient statistics to calculate an effect size (e.g., Jacobson, 1979). These were excluded. The final set of 26 papers eligible for the meta-analysis contained 33 independent samples and 336 effect sizes for all gestures. Three of these final 26 papers included references to "preliminary work" or "pilot studies" on neonatal imitation that were conducted in addition to the formally reported studies (i.e., Meltzoff & Moore, 1977; 1983b; 1992). Because no quantitative data were reported in these cases, however, they could not be included in the meta-analysis. Figure 1 summarizes the results of the systematic search.

+++++++++++++++++Insert Figure 1 about here.+++++++++++++++

Description of Studies included in the Analysis

Characteristics of the 26 included studies are summarized in Table 2.

+++++++++++++++Insert Tables 2 and 3 about here.+++++++++++++

Coding of potential moderators for each included study. Table 3 reports the coding for every consensus moderator (i.e., identified by at least two sources) for each study included in the meta-analysis.

Meta-Analysis Results

Average overall effect for neonatal imitation. The multilevel meta-analysis of neonatal imitation across all gestures revealed that, when all 336 effect estimates were combined, infants produced the target gesture more frequently in the imitation condition than in the control condition, and this effect was statistically significant ($d = 0.68$, 95% CI = 0.39 to 0.96, $p < .001$, $\sigma^2_1 = 0.38$, $\sigma^2_2 = 0.16$, $I^2_{\text{multi-level}} = 65.17\%$). Figure 2 presents both the study-level effect sizes (panel A) and the full multi-level data with multiple estimates per study (panel B).

Analysis of heterogeneity.

Heterogeneity in study results was assessed using regression tests for funnel plot asymmetry. The tests revealed significant heterogeneity, $QE(335) = 961.79$, $p < .001$, and significant asymmetry, such that studies with larger variances reported larger effect sizes, $b_{meta} = 2.46$, $se = 0.48$, $QM(1) = 26.50$, $p < .001$. One possible explanation for this heterogeneity, and funnel plot asymmetry, is publication bias. However, a plausible alternative explanation is the existence of methodological moderators that explain this pattern of results. We therefore first examine whether variations in study methods can explain heterogeneity in study results.

Methodology. Our moderator analyses revealed that, across the included studies, none of the thirteen categories of methodological variations identified by at least two previous reviews had a significant impact on the size of the overall neonatal imitation effect (see Table 4).

One specific methodological variation, the use of videos to model actions to infants, produced a significantly larger imitation effect than studies using live (or unspecified) models, $b_{meta} = 2.07$, $se = .89$, $p = .020$. The reliability of this restricted effect remains questionable, however, given that it is based on only two studies (Coulon et al., 2013; Soussignan et al., 2011). Furthermore, other literature suggests that infants do not recognize video images as depicting their 3-D referents until around 6 months of age (Pempek et al., 2010; Shinskey & Jachens, 2014), which makes infants' responses in these two experiments difficult to interpret. Importantly, the overall moderator of 'model identity' failed to account for a significant amount of variance in the imitation effect.

Studies that used one specific statistical analysis, the *Q*-test, also detected significantly larger imitation effects than studies using other analyses, $b_{meta} = 2.08$, $se = 1.08$, $p = .044$. This effect is also questionable, however, as only Meltzoff and Moore's (1977) original study and Koepke et al.'s (1983) failed replication used this method, which has since been superseded by alternative analytical techniques. Again, the overall moderator of 'statistical methods' failed to account for a significant amount of variance in the effect.

++++++++++++++++Insert Table 4 about here.++++++++++++++

Moderating effect of researcher affiliation. Unlike the other moderators, the effect of researcher affiliation on the overall size of the neonatal imitation effect was significant and substantial, QM(16) = 112.80, p < .001. As seen in Figure 3, six of the sixteen research groups consistently found neonatal imitation: CSGA, b = 3.14, se = .70, 95% CI [1.77, 4.48] p < .001; UO, b = 1.58, se = .58, 95% CI [0.45, 2.71] p = .006; UW, b = 1.44, se = .21, 95% CI [1.03, 1.85], p < .001; UPD, b = 1.41, se = .38, 95% CI [0.67, 2.15], p < .001; UD, b = 1.09, se = .27, 95% CI [0.57, 1.61], p < .001; and UM, b = 0.69, se = .32, 95% CI [0.05, 1.32], p = .035. In the discussion we will raise two possible interpretations of this pattern of results.

+++++++++++++++Insert Figures 3 and 4 about here.+++++++++++++++

Publication bias. Our combined tests revealed patterns matching those that would be expected in a literature that has been affected by publication bias. The funnel plot (Figure 4) shows imitation effect sizes (Cohen's *d*) plotted against their corresponding standard errors. Each point on the plot represents a single effect size, with larger effects to the right of the plot, and more precise effects (with smaller standard errors) at the top. The vertical line represents the overall multilevel meta-analysis result. Larger standard errors may reflect low sample sizes, measurement errors, or statistical errors. In an unbiased literature, effect sizes should be independent of standard error, by design (Egger et al., 1997). Effect sizes with small standard errors should cluster around the overall effect estimate at the top of the plot, and effect sizes with larger standard errors should be symmetrically spread around the overall effect estimate, forming a funnel shape. Points in the lower right corner of the funnel, without matching points in the lower left corner of the funnel, might indicate publication bias favoring a positive result, especially where these points fall outside the confidence intervals of the meta-analysis result (shown on the plot with dotted diagonal lines at 95% and 99% confidence).

A visual inspection of the funnel plot (Figure 4) reveals the expected pattern of variance compared to precision for the majority of effect estimates, with a relatively symmetrical distribution. Notably, however, the plot shows several studies with large standard errors that exhibit large positive effect sizes, and no corresponding studies with large standard errors that show null or negative effect sizes. This pattern is characteristic of a literature in which smaller studies have been conducted, found no evidence of neonatal imitation, and then were not published. The funnel plot is significantly asymmetrical according to a multilevel regression test (see Supplementary Material for methods), $b = 2.46$, $se = 0.48$, $p < .001$.

Impact of Individual Studies (Outliers). Studies with very high, or very low, estimates for neonatal imitation may change the results of the overall meta-analysis. The meta-analysis estimate is calculated as a weighted average, and therefore, study results that are very far from the average result may pull the meta-analysis estimate up (if they are larger than the average) or down (if they are smaller than the average). Although study estimates are weighted to address this, the problem may be magnified in meta-analyses where each study contributes several estimates to the meta-analysis, as is the case for neonatal imitation. We therefore tested the effect of each study on the multilevel meta-analysis, using meta-regression for the impact of the individual study ID on the overall meta-analysis result. As seen in Table 5, two of the studies (Meltzoff & Moore, 1994; Soussignan et al., 2011) contained effect sizes that were sufficiently positive to significantly increase the size of the overall meta-analysis estimate. No studies contained effect sizes that were sufficiently negative to significantly decrease the size of the overall estimate. Notably, including Oostenbroek et al.'s (2016) large-scale study, which contributed 225 of the 336 total effects, did not significantly change the size of the estimate (see Supplementary Material for full details of a meta-analysis excluding this study).

+++++++++++++++++Insert Table 5 about here.++++++++++++++

Supplementary Analyses

In addition to the main analyses reported here, we also conducted several supplementary analyses to more fully explore potential variations in the neonatal imitation effect. These analyses included (1) a restricted meta-analysis and assessment of publication bias that included only one result per study, (2) additional analyses of potential moderators that did not meet our final inclusion criterion, and (3) a separate meta-analysis and assessment of publication bias for the tongue protrusion action, which is the most commonly used gesture in neonatal imitation experiments (Ray & Heyes, 2011) and has been hypothesized to function differently from other actions (Anisfeld, 1996; Jones, 1996; Keven & Akins, 2017). None of these analyses produced results that contrasted with the overall pattern reported here, and so we have included them in the Supplementary Material only.

Discussion

Despite being incorporated into a wide range of influential theories across various psychological disciplines (e.g., Cheng & Chartrand, 2003; Lakin et al., 2003; Gallese, 2001; Ferrari et al., 2006; Meltzoff & Decety, 2003; Nadel & Butterworth, 1999; Trevarthen & Aitken, 2001), the existence of neonatal imitation continues to be hotly debated (see Keven & Akins, 2017, and commentaries). Here we used meta-analytic techniques to combine results from 26 separate studies of neonatal imitation. These studies included 336 data points on various infant behaviors, such as tongue protrusion, mouth opening, and finger pointing. The studies reported inconsistent findings for the same behaviors. For example, some studies found that infants only imitated tongue protrusion (e.g., Anisfeld et al., 2001; Heimann et al., 1989), whereas other studies found that infants imitated other gestures (e.g., Nagy et al., 2007; 2014) or found no imitation at all (e.g., Fontaine, 1984; Koepke et al., 1983; McKenzie & Over, 1983; Oostenbroek et al., 2016).

Overall, the multi-level meta-analysis indicated a significant and positive result for neonatal imitation but also significant heterogeneity in the literature. We found no evidence that the substantial heterogeneity in study results could be accounted for by methodological variables identified as potentially important in previous reviews (Meltzoff & Moore, 1983a; Meltzoff et al., 2017; Oostenbroek et al., 2013; Simpson et al., 2014; Vincini et al., 2017). Rather, an exploratory moderator analysis indicated that the overall neonatal imitation effect varies significantly by research institution.

Does Neonatal Imitation Exist?

At first glance, it may be tempting to conclude that the overall results of the meta-analysis provide robust evidence for the existence of neonatal imitation. After all, the significance test provided a *p* value of less than .001, and the overall *d* value of 0.68 is situated between Cohen's (1991) conventional cut-offs for 'medium' and 'large' effect sizes. However this overall result must be interpreted with caution. The only significant predictor of neonatal imitation was researcher affiliation, with some research groups consistently finding neonatal imitation and others consistently not finding it. As Figure 3 indicates, the positive effects from the subset of six research groups reporting non-zero effects on average, had wide confidence intervals that overlapped with the intervals of many other affiliations. Similarly the funnel plot (Figure 4) indicates that a large majority of the effect sizes cluster in a roughly normal distribution centered on zero. The overall positive effect therefore appears to be driven by some disproportionately large positive effects from particular research institutions.

There are at least two possible interpretations of this pattern of findings. We now go through each interpretation in turn.

Possible interpretation 1: Neonatal imitation does exist and varies systematically as a function of uncaptured methodological factors. Our planned moderator analyses found that the overall neonatal imitation effect was not significantly affected by any of the thirteen methodological factors that met our inclusion criteria. Therefore, the argument that those specific differences in methodology are critical to the effect (see Meltzoff & Moore, 1983a; Meltzoff et al., 2017; Oostenbroek et al., 2013; Simpson et al., 2014; Vincini et al., 2017) is not supported by the meta-analysis.

Yet, these analyses were not without limitations, and it could be that they simply failed to capture systematic variance critical to the effect. There may be methodological variations that are crucial to observing neonatal imitation, but they were not coded in appropriate detail or even not coded at all. For instance, over half of the studies failed to describe whether infants' pre-exposure to the experimenter's face was controlled prior to testing, and only three studies explicitly described preventing infants from seeing the experimenter's face in all conditions (Reissland, 1989; Meltzoff & Moore, 1983b; 1989). The null result for this moderator should therefore be interpreted with caution (see Meltzoff & Moore, 1983a; Meltzoff et al., 2017; Simpson et al., 2014), and we recommend that future studies directly compare infants' responses across conditions with and without pre-experimental exposure. Another limitation was that, because the analysis required us to narrowly categorize study methods, we necessarily failed to capture some of the complexity and context of certain methodological practices. For example, we classified the studies into various categories of modeler identity, testing location, and infant testing position, but one might expect systematic variation *within* these categories as well as between them (see Coyne, 2016). Some studies where testing took place in the infants' home, for instance, may have better controlled for noise interference than others.

It is also possible that methodological factors critical to the effect have yet to be identified. Presumably these variations are common to the research groups that report larger effects than others, which would explain why our moderator analyses found a significant effect for researcher affiliation. It may therefore be fruitful to closely examine the method sections of papers reporting large neonatal imitation effects to identify common procedures that may have positively influenced the outcome variable. Although, it could also be that methodological practices *not* reported in these papers are critical. It is unreasonable to expect infancy researchers to report every minute procedural detail in a method section, and yet it may turn out that certain overlooked and unreported details have been responsible for eliciting imitation in newborns.

Even if future research were to establish that neonatal imitation can be reliably elicited under certain methodological conditions, the demonstrated difficulty in identifying these conditions does not align well with the proposal that the phenomenon is a "foundation for social cognition" (e.g., Meltzoff, 2007). A foundational skill surely should be evident in a range of conditions both in the lab and in the home. Future theorising about the significance of neonatal imitation to human development, social cognition and interpersonal behaviour, should take this into account.

Possible interpretation 2: Neonatal imitation does not exist and positive findings are an artefact of researcher allegiance. If uncaptured common methodological variance does not explain the heterogeneity in the neonatal imitation literature, then we must accept the alternative possibility that neonates do not imitate. On this view, the overall positive meta-analytic effect may be an artefact of researcher allegiance. This could be attributable to researcher bias, whereby researchers' beliefs about the veracity of neonatal imitation, either positive or negative, influence environmental and procedural factors that contribute to the overall result in their laboratories. Currently there is no evidence of such biases, so further research would be required to investigate this possibility. Alternatively, idiosyncratic practices in particular research labs might erroneously identify some infant behaviours as imitation, even in the absence of a bias toward one or another outcome.

One important consideration is the inordinately high researcher degrees of freedom (see Gelman & Loken, 2013; Ioannidis, 2005) available in neonatal imitation studies.

Operationalizations of dependent variables fluctuate widely (see Oostenbroek et al., 2013), with different studies reporting raw frequencies of the gesture (e.g., Fontaine, 1984), mean frequencies (e.g., Oostenbroek et al., 2016; Ullstadius, 1998), mean rates per minute (e.g., Legerstee, 1991), duration (e.g., McKenzie & Over, 1983), or number of matches and mismatches (e.g., Abranavel & Sigafoos, 1984). Recent advances in our understanding of the susceptibility of psychological studies to Type I error under conditions of statistical ambiguity suggest that these sorts of variations may be just as problematic as variations in experimental methods (Simonsohn, Nelson, & Simmons, 2014). Indeed, case studies of classic and recent neonatal imitation papers suggest that issues arising from researcher degrees of freedom may be widespread:

Case Study 1. In Meltzoff and Moore's (1992) paper, the authors report collecting pilot data from 48 infants, for whom a qualitative pattern of results is described in the absence of formal statistics. Importantly, this pattern of results is portrayed as being inconsistent with a neonatal imitation effect: "Infants saw one person perform the mouth-opening demonstration and then the second person perform the tongue-protrusion demonstration. Infants acted in an interesting and surprising way when the first person disappeared and the other person appeared and began producing the new gesture. Infants often stared at the second person and then intently produced what the *first* model had demonstrated" (p. 483). This unexpected finding then prompted the authors to design a follow-up study. The critical issue is that data from the pilot study were not presented in any quantitative detail, and so these data could not be included in our meta-analysis. The follow-up study, on the other hand, produced positive results and was written up as the main quantitative finding. Such practices fail to meet modern standards of research quality control (Simmons, Nelson, & Simonsohn, 2011), and it is difficult to estimate how many other null or negative findings from early studies were similarly discounted.

Case Study 2. In Heimann and Tjus' (2019) recent paper, the authors describe an initial analysis plan in which they would split their 120-second imitation data into two 60-second time windows, reasoning that infants needed some time to organize their imitative responses after seeing the model. When reporting their results, however, the authors wrote that this plan "was changed before carrying out the final analysis based on the observation that few infants (<40%) imitated during the final 20 s of the experiment" (pp. 681). Instead, the authors analyzed responses across three time windows (1-60 seconds, 61-100 seconds, and 101-120 seconds). However, inspection of Figures 1 and 2 of this paper suggests that infants' overall response frequencies in the last 20 seconds were not abnormal, and that splitting the time windows in this post-hoc manner may have been responsible for a significant finding of mouth-opening imitation between 61 and 100 seconds. Substantiating this possibility, our meta-analysis indicated no overall imitation effect for Heimann and Tjus' (2019) study.

Case Study 3. Nagy, Pilling, Blake and Orvos' (2019) recent report of neonatal imitation may have also been influenced by researcher degrees of freedom. In particular, the method included no formal stopping rule for the modelling of actions to infants, such that some modelling periods lasted nearly twice as long as others (i.e., the average varied from 192 seconds to 359 seconds across different actions). The lack of a formal stopping rule invites potential unconscious bias, whereby the experimenter only stops modelling each action after perceiving that the infant has produced a matching response (instead of stopping when the infant has produced a non-matching response). Such a bias would artificially increase the likelihood of detecting significant neonatal imitation effects. Furthermore, this study did not include exact statistics for non-significant results. This not only makes it difficult to obtain a clear picture of the overall findings, but it also prevented us from including the study in the meta-analysis (as including only positive results would have biased the overall estimate).

Presence of Publication Bias

Our findings suggest that the obtained average effect size was affected by publication bias. In particular, the funnel plot was significantly and substantially asymmetrical, implying that publication bias may at least partially explain the inconsistencies in neonatal imitation studies.

We acknowledge that post-hoc statistical methods for assessing publication bias in meta-analysis can only indicate correlational evidence and not causality. We do, however, propose that publication bias may be a natural outcome of conditions like those observed in the study of neonatal imitation.

Logistical constraints on studies with infants mean that the samples are necessarily small. Adding to that, a certain number of false positives would be expected in an experimental paradigm such as the one typically used in these studies, where infants' gestures are often counted and sorted into two categories: match and mismatch. Given the pressures within academic publishing to find and report only "interesting" research, it is expected that chance positive findings will take precedence in the published literature while null or negative findings are sidelined (Rosenthal, 1979; Rothstein et al., 2006).

Conclusion and Recommendation for Future Research

Overall, our meta-analysis reveals a significant neonatal imitation effect that is significantly moderated by researcher affiliation, but not moderated by any of 13 methodological factors previously identified as important to demonstrating the effect. These findings suggest several ways forward.

Neonatal imitation researchers might adopt practices from the clinical psychology literature to evaluate associations between researchers' beliefs about the veracity of neonatal imitation, and the effects reported from their labs. Via reprint analysis and interviews with key researchers and their colleagues, it is possible to code the direction and strength of beliefs, and then correlate those with neonatal imitation outcomes. However it should be noted that these procedures for deriving an "allegiance score" were not well correlated in the one study that utilised all of them (Luborsky et al, 1999). Furthermore a positive association would have to be interpreted carefully—as noted above, the causal arrow could go both ways. Despite the drawbacks, this type of analysis would be worthwhile if it uncovered suggestions of researcher bias in this research field.

If neonatal imitation effects vary as a function of uncaptured systematic methodological variation, then it should be possible to consistently detect positive results by replicating procedures of the laboratories that previously reported large effects. This might involve cross-institutional training prior to designing any new studies on neonatal imitation, ideally with researchers visiting more than one of the institutions where the effect has been consistently observed. Additionally, neonatal imitation researchers from different laboratories could share not only detailed protocols but also videos of their stimuli and the infants' behaviour. Currently the largest open-source video data repository for developmental scientists (databrary.org) contains no footage of neonatal imitation.

Based on these shared resources, researchers should implement large-scale, multi-lab, pre-registered replication studies (see Frank et al., 2017). Ideally, the lead researchers and participating labs would be independent, with no previous experience conducting or reporting on neonatal imitation studies and no strong commitment one way or the other to the existence of the effect.

It must be acknowledged that research with neonates is difficult and expensive to execute, and replication studies, no matter how important, may not be a top priority for labs with limited resources. Given this, another approach would be an "adversarial collaboration" between laboratories currently engaged in neonatal imitation research. This would involve two or more research groups who have reported differing outcomes, performing a joint study under an agreed protocol with a neutral "arbiter" overseeing the planning, implementation and publication (as outlined in Mellers, Hertwig & Kahneman, 2001).

Sharing of detailed protocols and video footage of neonatal imitation studies would also facilitate methodological analyses, ideally by scientists who have not themselves implemented neonatal imitation studies. While it is essential that studies of neonatal imitation account for newborns' unique sensori-motor limitations, it possible that the field's almost uniform adherence to some variant of Meltzoff and Moore's (1977; 1983b) testing protocol, has limited innovation.

Some “clear-eyed” methodological reviews might reveal systematic factors associated with both negative and positive findings, that may not be readily evident to developmental researchers who are steeped in knowledge and assumptions about newborn behaviour.

Utilising one or more of these strategies, it should be possible to find positive evidence of neonatal imitation if the phenomenon exists. If positive outcomes are not found in studies designed to mitigate the effects of researcher allegiance, then the most parsimonious conclusion would be that neonatal imitation does not exist and previous positive findings are artefacts of researcher bias and/or idiosyncratic research practices.

References

Studies used in the meta-analysis are marked with *

Abranavel, E., & Sigafoos, A. D. (1984). Exploring the presence of imitation during early infancy. *Child Development*, *55*(2), 381-392.

Anisfeld, M. (1979). Interpreting "imitative" responses in early infancy. *Science*, *205*(4402), 214-215.

Anisfeld, M. (1991). Neonatal imitation. *Developmental Review*, *11*(1), 60-97.

Anisfeld, M. (1996). Only tongue protrusion modelling is matched by neonates. *Developmental Review*, *16*(2), 149-161.

* Anisfeld, M., Turkewitz, G., & Rose, S. A. (2001). No compelling evidence that newborns imitate oral gestures. *Infancy*, *2*(1), 111-122.

Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R., & Livingstone, M. S. (2017). Seeing faces is necessary for face-domain formation. *Nature Neuroscience*, *20*(10), 1404-1412.

* Barbosa, P. G. (2017). *Are you like me? Maybe, but I will not imitate you! A longitudinal study on newborns and infants' imitation and conspecific identification skills*. Doctoral dissertation. University of Alberta.

Bjorklund, D. F. (2018). A metatheory for cognitive development (or "Piaget is dead" revisited). *Child Development*. doi: 10.1111/cdev.13019

Boccaccini, M. T., Marcus, D., & Murrie, D. C. (2017). Allegiance effects in clinical psychology research and practice. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed remedies* (pp. 323-339). New York: John Wiley & Sons.

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

Boyd, R., & Richerson, P. J. Why culture is common, but cultural evolution is rare. *Proceedings of the British Academy*, *88*(1), 77-93.

Brass, M., & Heyes, C. (2005). Imitation: Is cognitive neuroscience solving the correspondence problem?. *Trends in Cognitive Sciences*, *9*(10), 489-495.

Bushneil, I. W. R., Sai, F., & Mullin, J. T. (1989). Neonatal recognition of the mother's face. *British Journal of Developmental Psychology*, *7*(1), 3-15.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893-910.

Cheng, C. M., & Chartrand, T. L. (2003). Self-monitoring without awareness: Using mimicry as a nonconscious allegiance strategy. *Journal of Personality and Social Psychology*, *85*(6), 1170-1179.

Clay, Z., & Tennie, C. (2017). Is overimitation a uniquely human phenomenon? Insights from human children as compared to bonobos. *Child Development*. doi: 10.1111/cdev.12857

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

* Coulon, M., Hemimou, C., & Streri, A. (2013). Effects of seeing and hearing vowels on neonatal facial imitation. *Infancy*, *18*(5), 782-796.

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology*, *4*(1), 28.

Cracco, E., Bardi, L., Desmet, C., Genschow, O., Rigoni, D., De Coster, L., ... & Brass, M. (2018). Automatic imitation: A meta-analysis. *Psychological Bulletin*, *144*(5), 453-500.

Des Jarlais, D. C., Lyles, C., Crepaz, N., & TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. *American Journal of Public Health*, *94*(3), 361-366.

Dragioti E, Dimoliatis I, & Evangelou E. (2015). Disclosure of researcher allegiance in meta-analyses and randomised controlled trials of psychotherapy: a systematic appraisal. BMJ Open, 5, e007206, doi:10.1136/bmjopen-2014-007206

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629-634.

Egger, M., Smith, G. D., & Sterne, J. A. C. (2001). Uses and abuses of meta-analysis. *Clinical Medicine*, *1*(6), 478-484.

Fanelli, D., Costas, R., & Ioannidis, J. P. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences*, *114*(14), 3714-3719.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120-128.

Ferrari, P. F., Visalberghi, E., Paukner, A., Fogassi, L., Ruggiero, A., et al., (2006). Neonatal imitation in rhesus macaques. *PLoS Biology*, *4*(9), e302.

Ferrari, P. F., Paukner, A., Ruggiero, A., Darcey, L., Unbehagen, S., & Suomi, S. J. (2009). Interindividual differences in neonatal imitation and the development of action chains in rhesus macaques. *Child Development*, *80*(4), 1057-1068.

* Field, T. M., Woodson, R., Cohen, D., Greenberg, R., Garcia, R., & Collins, K. (1983). Discrimination and imitation of facial expressions by term and preterm neonates. *Infant Behavior and Development*, *6*(4), 485-489.

* Field, T. M., Woodson, R., Greenberg, R., & Cohen, D. (1982). Discrimination and imitation of facial expressions neonates. *Science*, *218*(4568), 179-181.

* Fontaine, R. (1984). Imitative skills between birth and six months. *Infant Behavior and Development*, *7*(3), 323-333.

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*(4), 421-435.

Gallese, V. (2001). The ‘shared manifold’ hypothesis: From mirror neurons to empathy. *Journal of Consciousness Studies*, *8*(5-6), 33-50.

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

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly?. *Infant Behavior and Development*, *21*(2), 167-179.

* Heimann, M. (1998). The story of neonatal imitation: New facts and old conclusions. *Infant Behavior and Development*, *21*(Supplement), 454.

* Heimann, M., Nelson, K. E., & Schaller, J. (1989). Neonatal imitation of tongue protrusion and mouth opening: Methodological aspects and evidence of early individual differences. *Scandinavian Journal of Psychology*, *30*(2), 90-101.

* Heimann, M., & Schaller, J. (1985). Imitative reactions among 14-21 day old infants. *Infant Mental Health Journal*, *6*(1), 31-39.

* Heimann, M., & Tjus, T. (2019). Neonatal imitation: Temporal characteristics in imitative response patterns. *Infancy*, *24*(5), 674-692.

Henrich, J. (2015). Culture and social behavior. *Current Opinion in Behavioral Sciences, 3*, 84–89.

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *12*(3), 123-135.

Heyes, C. (2016a). Homo imitans? Seven reasons why imitation couldn't possibly be associative. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150069.

Heyes, C. (2016b). Imitation: Not in our genes. *Current Biology*, *26*(10), R412-R414.

Hickok, G. (2014). *The Myth of Mirror Neurons: The Real Neuroscience of Communication and Cognition*. W. W. Norton & Company, New York.

Higgins, J. P. T., Deeks, J., & Altman, D. G. (2011). Chapter 16: Special topics in statistics. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (Vol. 5). The Cochrane Collaboration.

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 5). The Cochrane Collaboration.

Higgins, J. P. T., Altman, D. G., Gotzsche, P. C., Juni, P., Moher, D., Oxman, A. D., Savovic, J., Schulz, K. F., Weeks, L., & Sterne, J. A. C. (2011b). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, *343*, d5928.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Animal Cognition*, *8*(3), 164-181.

Iacobini, M. (2009). Imitation, empathy, and mirror neurons. *Annual Review of Psychology*, *60*, 653-670.

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

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640-648.

Jacobson, S. W., & Kagan, J. (1979). Interpreting "imitative" responses in early infancy. *Science*, *205*(4402), 215-217.

Jones, S. S. (1996). Imitation or exploration? Young infants' matching of adults' oral gestures. *Child Development*, *67*(5), 1952-1969.

Jones, S. S. (2007). Imitation in infancy: The development of mimicry. *Psychological Science*, *18*(7), 593-599.

Jones, S. S. (2009). The development of imitation in infancy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2325-2335.

Jones, S. S. (2017). Can newborn infants imitate? *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(1-2), e1410.

* Kennedy-Costantini, S. (unpublished). *Very early social learning within the context of the mother-baby relationship*. University of Queensland: Unpublished PhD dissertation.

Keven, N., & Akins, K. (2017). Neonatal imitation in context : Sensory-motor development in the perinatal period. *Behavioral and Brain Sciences*, *Behavioral and Brain Sciences*, *40*, e381.

* Koepke, J. E., Hamm, M., & Legerstee, M. (1983). Neonatal imitation: Two failures to replicate. *Infant Behavior and Development*, *6*(1), 97-102.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138-1141.

Lakin, J. L., Jefferis, V. E., Cheng, C. M., & Chartrand, T. L. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior*, *27*(3), 145-162.

Legare, C. H., & Nielsen, M. (2015). Imitation and innovation: The dual engines of cultural learning. *Trends in Cognitive Sciences*, *19*(11), 688-699.

* Legerstee, M. (1991). The role of person and object in eliciting early imitation. *Journal of Experimental Child Psychology*, *51*(3), 423-433.

Lepage, J. F., & Théoret, H. (2007). The mirror neuron system: Grasping others' actions from birth?. *Developmental Science*, *10*(5), 513-523.

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analyses and subsequent large randomized, controlled trials. *New England Journal of Medicine*, *337*(8), 536-542.

Leykin, Y., & DeRubeis, R. J. (2009). Allegiance in psychotherapy outcome research: Separating association from bias. *Clinical Psychology: Science and Practice, 16*(1), 54–65.

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., Halperin, G., Bishop, M., Berman, J. S., & Schweizer, E. (1999). The researcher's own therapy allegiances: A "wild card" in comparisons of treatment efficacy. Clinical Psychology: Science and Practice, 6(1), 95-106.

Manea, L., Boehnke, J. R., Gilbody, S., Moriarty, A. S., & McMillan, D. (2017). Are there researcher allegiance effects in diagnostic validation studies of the PHQ-9? A systematic review and meta-analysis. BMJ Open, 7(9), e015247.

Masters, J. C. (1979). Interpreting "imitative" responses in early infancy. *Science*, *205*(4402), 215.

* McKenzie, B., & Over, R. (1983). Young infants fail to imitate facial and manual gestures. *Infant Behavior and Development*, *6*(1), 85-95.

McKyton, A., Ben-Zion, I., & Zohary, E. (2018). Lack of automatic imitation in newly sighted individuals. *Psychological Science*, *29*(2), 304-310.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12*(4), 269–275.

Meltzoff, A. N. (2002). Elements of a developmental theory of imitation. In A. Meltzoff & W. Prinz (Eds.), *The imitative mind: Development, evolution, and brain bases* (pp. 19–41). Cambridge: Cambridge University Press.

Meltzoff, A.N. (2005). Imitation and other minds: the 'like me' hypothesis. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation: From neuroscience to social science (Vol. 2: Imitation, human development, and culture*, pp. 55 – 77). Cambridge, MA: MIT Press.

Meltzoff, A. N. (2007). 'Like me': A foundation for social cognition. *Developmental Science*, *10*(1), 126-134.

Meltzoff, A. N., & Decety, J. (2003). What imitation tells us about social cognition: A rapprochement between developmental psychology and cognitive neuroscience. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *358*(1431), 491-500.

* Meltzoff, A. N. & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*(4312), 75-78.

Meltzoff, A. N., & Moore, M. K. (1983a). Methodological issues in studies of imitation: Comments on McKenzie & Over and Koepke et al. *Infant Behavior and Development*, *6*(1), 103-108.

* Meltzoff, A. N. & Moore, M. K. (1983b). Newborn infants imitate adult facial gestures. *Child Development*, *54*(3), 702-709.

* Meltzoff, A. N., & Moore, M. K. (1989). Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Developmental Psychology*, *25*(6), 954-962

* Meltzoff, A. N. & Moore, M. K. (1992). Early imitation within a functional framework: The importance of person identity, movement, and development. *Infant Behavior and Development*, *15*(4), 479-505.

* Meltzoff, A. N. & Moore, M. K. (1994). Imitation, memory, and the representation of persons. *Infant Behavior and Development*, 17(1), 83-99.

Meltzoff, A. N., Murray, L., Simpson, E., Heimann, M., Nagy, E., Nadel, J., ... & Subiaul, F. (2017). Re-examination of Oostenbroek et al. (2016): evidence for neonatal imitation of tongue protrusion. *Developmental Science*.

Munder, T., Brutsch, O., Leonhart, R., Gerger, H., & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: An overview of reviews. Clinical Psychology Review, 33(4), 501-511. doi: 10.1016/j.cpr.2013.02.002.

Myowa-Yamakoshi, M., Tomonaga, M., Tanaka, M., & Matsuzawa, T. (2004). Imitation in neonatal chimpanzees (Pan troglodytes). *Developmental Science*, *7*(4), 437-442.

Nadel, J., & Butterworth, G. (1999). *Imitation in infancy*. Cambridge: Cambridge University Press.

Nagy, E., & Molnar, P. (2004). Homo imitans or homo provocans? Human imprinting model of neonatal imitation. *Infant Behavior & Development*, *27*(1), 54-63.

* Nagy, E., Kompagne, H., Orvos, H., & Pal, A. (2007). Gender-related differences in neonatal imitation. *Infant and Child Development*, *16*(3), 267-276.

* Nagy, E., Pal., A., & Orvos, H. (2014). Learning to imitate individual finger movements by the human neonate. *Developmental Science*, *17*(6), 841-857.

Nagy, E., Pilling, K., Blake, V., & Orvos, H. (2019). Positive evidence for neonatal imitation: A general response, adaptive engagement. *Developmental Science*, e12894.

Nagy, E., Pilling, K., Orvos, H., & Molnar, P. (2013). Imitation of tongue protrusion in human neonates: Specificity of the response in a large sample. *Developmental Psychology*, *49*(9), 1628.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*(4), 591-605.

Nielsen, M., & Tomaselli, K. (2010). Overimitation in Kalahari Bushman children and the origins of human cultural cognition. *Psychological Science*, *21*(5), 729-736.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422-1425.

Oostenbroek, J., Slaughter, V., Nielsen, M., & Suddendorf, T. (2013). Why the confusion around neonatal imitation? A review. *Journal of Reproductive and Infant Psychology*, *31*(4), 328-341.

* Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J., Clark, S., & Slaughter, V. (2016). Comprehensive longitudinal study challenges the existence of neonatal imitation in humans. *Current Biology*, *26*(10), 1334-1338.

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

Pempek, T. A., Kirkorian, H. L., Richards, J. E., Anderson, D. R., Lund, A. F., & Stevens, M. (2010). Video comprehensibility and attention in very young children. *Developmental Psychology*, *46*(5), 1283.

Piaget, J. (1962). *Play, dreams and imitation in childhood*. New York: Norton.

Ray, E., & Heyes, C. (2011). Imitation in infancy: The wealth of the stimulus. *Developmental Science*, *14*(1), 92-105.

Redshaw, J. (2019). Re-analysis of data reveals no evidence for neonatal imitation in rhesus macaques. *Biology Letters*, *15*(7), 20190342.

Redshaw, J., Nielsen, M., Slaughter, V., Kennedy-Costantini, S., Oostenbroek, J., Crimston, J., & Suddendorf, T. (2019). Individual differences in neonatal “imitation” fail to predict early social cognitive behaviour. *Developmental Science*, e12891.

* Reissland, N. (1988). Neonatal imitation in the first hour of life: Observations in rural Nepal. *Developmental Psychology*, *24*(4), 464-489.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2006). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. Hoboken, NJ: John Wiley & Sons.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-641.

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review*, *17*(8), 881-901.

Shinskey, J. L., & Jachens, L. J. (2014). Picturing objects in infancy. *Child Development*, *85*(5), 1813-1820.

Singh, J. P., Grann, M., & Fazel, S. (2013). Authorship bias in violence risk assessment? A systematic review and meta-analysis. *PLoS ONE*, 8(9), e72484. doi:10.1371/journal.pone.0072484

* Soh, S-E., Tint, M. T., Gluckman, P. D., Godfrey, K. M., Rifkin-Graboi, A., Chan, Y. H., Stünkel, W., Holbrook, J. D., Kewk, K., Chong, Y-S., Saw, S. M., The GUSTO Study Group. (unpublished). *Growing up in Singapore towards healthy outcomes (GUSTO) birth cohort study*. The GUSTO Study Group: Unpublished data set.

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

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76-80.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534-547.

Simpson, E.A., Murray, L. Paukner, A., & Ferrari, P. F. (2014). The mirror neuron system as revealed through neonatal imitation: presence from birth, predictive power and evidence of plasticity *Philosophical Transactions of the Royal Society, B, Biological Sciences, 369*, 20130289.

* Soussignan, R., Courtial, A., Canet, P., Danon-Apter, G., & Nadel, J. (2011). Human newborns match tongue protrusion of disembodied human and robotic mouths. *Developmental Science*, *14*(2), 385-394.

Sullivan, G. M., & Feinn, R. (2012). Using effect size – or why the *p* value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279-282.

Suddendorf, T., Oostenbroek, J., Nielsen, M., & Slaughter, V.P. (2013). Is newborn imitation developmentally homologous to later social-cognitive developments? *Developmental Psychobiology, 55*, 52-58.

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *BMJ*, *320*(7249), 1574-1577.

Tennie, C., Call, J., & Tomasello, M. (2009) Ratcheting Up the ratchet: On the evolution of cumulative culture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364*, 2405–2415.

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences*, *16*(3), 495-552.

Trevarthen, C., & Aitken, K. J. (2001). Infant intersubjectivity: Research, theory, and clinical applications. *The Journal of Child Psychology and Psychiatry*, *42*(1), 3-48.

* Ullstadius, E. (1998). Neonatal imitation in a mother-infant setting. *Early Development and Parenting*, *7*(1), 1-8.

Uzgiris, I. C., & Hunt, J. M. (1975). *Assessment in infancy: Ordinal scales of psychological development*. Urbana, IL: University of Illinois Press.

van Baaren, R. B., Holland, R. W., Kawakami, K., & Van Knippenberg, A. (2004). Mimicry and prosocial behavior. *Psychological Science*, *15*(1), 71-74.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1-48.

Vincini, S., Jhang, Y., Buder, E. H., & Gallagher, S. (2017). Neonatal imitation: Theory, experimental design, and significance for the field of social cognition. *Frontiers in Psychology*, *8*, 1323.

Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1528), 2417-2428.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *Behavioral and Brain Sciences*, 1-50. doi: 10.1017/S0140525X17001972

Table 1. Potential Moderators of the Neonatal Imitation Effect.

Criterion	**Source**
Sample size[a]	Meltzoff & Moore (1983a) Simpson et al. (2014) Vincini et al. (2017)
Length of modelling and response periods[1]	Simpson et al. (2014) Vincini et al. (2017) Meltzoff et al. (2017)
Number of models shown to infant[2]	Simpson et al. (2014) Vincini et al. (2017) Meltzoff et al. (2017)
Setting of data collection[3]	Oostenbroek et al. (2013) Vincini et al. (2017)
Pre-experimental exposure to experimenter[4]	Meltzoff & Moore (1983a) Oostenbroek et al. (2013) Simpson et al. (2014) Vincini et al. (2017) Meltzoff et al. (2017)
Whether or not infant is tested in a padded seat[5]	Oostenbroek et al. (2013) Meltzoff et al. (2017)
Infant alertness or state during testing[6]	Meltzoff & Moore (1983a) Oostenbroek et al. (2013) Simpson et al. (2014) Vincini et al. (2017) Meltzoff et al. (2017)
Suitability of statistical tests[7]	Meltzoff & Moore (1983a) Vincini et al. (2017)
Size of the video for coding	Meltzoff & Moore (1983a)
Scoring only well-defined movements	Meltzoff & Moore (1983a)
Reports gender distribution of participants	Oostenbroek et al. (2013)
Whether or not infant was tested on multiple occasions	Oostenbroek et al. (2013)
Dependent variable operationalization	Oostenbroek et al. (2013)

Categorization of infants into "imitators" and "non-imitators"	Simpson et al. (2014)
Baseline and non-social control conditions	Simpson et al. (2014)
Only one gesture modelled in each session	Vincini et al. (2017)
Specific behaviors modelled	Meltzoff et al. (2017)
Infant response coding criteria	Meltzoff et al. (2017)
Presence of distracting visual stimuli	Meltzoff et al. (2017)
Subject selection for statistical analyses	Meltzoff et al. (2017)
Deviations from modelling protocol	Meltzoff et al. (2017)
Counterbalancing of modelled actions across infants	Meltzoff et al. (2017)

Note. Criteria in **bold** are broadly mentioned in more than one source and were therefore included as moderators. Matching superscripts indicate overlap in sources for each criterion: Number superscripts (1-7) indicate the matching criteria that were analyzed as part of the moderator analysis, whereas the letter superscript (a) indicates the sample size criterion, which was analyzed as part of the publication bias analysis.

Table 2. Summary of Studies Included in the Neonatal Imitation Meta-analysis.

Study *Subgroup (if applicable)*	**N**[a]	**Mean age (days)**	**Infant gestures studied**
Anisfeld, Turkewitz, & Rose (2001)	83	2	Tongue protrusion Mouth opening
Barbosa (2017)			Tongue protrusion (forward and lateral)
Time 1	51	4	
Time 2	74	31	
Coulon, Hemimou, & Streri (2013)			Mouth opening Lip spreading
Visual-only condition	12	2	
Congruent condition	12	2	
Incongruent condition (no data)	12	2	
Field et al. (1982)	96	1	Widened lips Tight protruded lips Wide open mouth
Field et al. (1983)	74	1	Widened lips Tight protruded lips Furrowed brow Wide open mouth Widened eyes
Fontaine (1984)	83	30	Tongue protrusion Mouth opening Cheek swelling Eye closing Hand opening Index finger point
Heimann & Schaller (1985)	11	18	Tongue protrusion Mouth opening
Heimann, Nelson, & Schaller (1989)	23	3	Tongue protrusion Mouth opening
Time 1	24	21	
Time 2			
Heimann (1998)	33	2	Tongue protrusion Mouth opening
Heimann & Tjus (2019)	33	2	Tongue protrusion Mouth opening
Kennedy-Costantini (unpublished)	48	8	Tongue protrusion Mouth opening
Koepke, Hamm, & Legerstee (1983)	14	19	Tongue protrusion Mouth opening
Legerstee (1991)	12	46	Tongue protrusion Mouth opening
McKenzie & Over (1983)	14	23	Tongue protrusion Mouth opening Arm raising Hand to mouth
Meltzoff & Moore (1977)			Lip protrusion Mouth opening Tongue protrusion
Experiment 1	6	14	
Experiment 2	12	14	

			Sequential finger movement Mouth opening
Meltzoff & Moore (1983b)	40	1	Tongue protrusion Mouth opening
Meltzoff & Moore (1989)	40	2	Tongue protrusion Head movement
Meltzoff & Moore (1992)	32	43	Tongue protrusion Mouth opening
Meltzoff & Moore (1994)	40	42	Tongue protrusion Mouth opening
Nagy et al. (2007)	41	2	Hand movement Finger movement
Nagy, Pal, & Orvos (2014)			Index finger point
Left condition	37	2	
Right condition	43	2	
Both condition	41	2	
Oostenbroek et al. (2016)			Tongue protrusion Mouth opening Happy face Sad face Index finger point Grasp MMM sound EEE sound Tongue click
1 week old	74	7	
3 weeks old	80	21	
6 weeks old	83	42	
Reissland (1988)	12	1	Wide lips Pursed lips
Soh et al. (unpublished)	136	1	Tongue protrusion Mouth opening
Soussignan et al. (2011)	18	2	Tongue protrusion
Ullstadius (1998)	14	1	Tongue protrusion Mouth opening

[a] *N* refers to the number of neonates in the study's final sample size.

Table 4. Summary of the Impact of Methodological Variations.

Methodological variation	**Meta-regression result**			
	k	b_{meta}	*se*	*p*
Model presentation time per burst (seconds)	312	.01	.01	.359
Infant response time per burst (seconds)	308	< .01	.01	.428
Total model presentation time (seconds)	310	> -.01	.01	.676
Total infant response time (seconds)	317	> -.01	< .01	.313
Total number of actions modelled	336	-.07	.06	.279
Total active experiment time	301	> -.01	<.01	.345
Experimental setting (categorical)	336	**$QM(4) = 4.89^a$**		**.299**
Home	*237*	*.12*	*.81*	*.886*
Hospital	*65*	*.92*	*.76*	*.224*
University	*14*	*.69*	*.82*	*.399*
Not reported	*18*	*.83*	*.82*	*.321*
Identity of modeler (categorical)	336	**$QM(5) = 8.62^a$**		**.125**
Experimenter	*313*	*.64*	*.73*	*.380*
Mother	*8*	*.21*	*.88*	*.809*
Video	*9*	*2.07*	*.89*	*.020**
Multiple	*2*	*.89*	*1.02*	*.384*
Not reported	*2*	*.87*	*1.02*	*.392*
Pre-experimental exposure to the modeler's face? (categorical, 0 = unclear/not reported)	336	**$QM(2) = 0.23^a$**		**.893**
Not Exposed	10	*.19*	*.44*	*.661*
Exposed	188	*-.03*	*.30*	*.930*
Infant testing position (categorical)	336	**$QM(7) = 8.04^a$**		**.329**
Experimenter arms	*236*	*-.27*	*.16*	*.080*
Experimenter lap	*20*	*-.60*	*.56*	*.286*
Infant seat	*33*	*-.22*	*.35*	*.533*
Multiple	*10*	*-.96*	*.56*	*.089*
Mother arms	*4*	*-.66*	*.79*	*.401*
Mother lap	*4*	*-.83*	*.78*	*.290*
Pillow	*2*	*.65*	*.84*	*.439*
Infant seat? (binary, 0 = no infant seat)	322	.58	.31	.062
Measure of infant alertness (categorical)	336	**$QM(5) = 7.44^a$**		**.190**
Behavioral criterion	*14*	*.86*	*.72*	*.337*
Experimenter judgement	*54*	*.45*	*.89*	*.546*
Looking time	*17*	*1.35*	*.81*	*.094*
Mixture (Brazelton Scale)	*225*	*.12*	*.95*	*.903*
Not reported	*24*	*.98*	*.77*	*.203*
Statistical analysis method (6 categories)	336	**$QM(6) = 9.49^a$**		**.148**

ANOVA	*37*	*.72*	*.73*	*.330*
Difference score	*6*	*.04*	*.99*	*.968*
GLMM	*225*	*.12*	*.92*	*.900*
Q-test	*4*	*2.08*	*1.03*	*.044**
Signed-rank test	*56*	*.68*	*.72*	*.349*
t-test	*6*	*1.60*	*.86*	*.061*

---

[a] *QM* measures the overall category effect in lieu of b_{meta} and *se* $^{*}p < .05$

Table 5. Impact of Individual Studies on Size of Meta-Analysis Estimate

Study	**Number of effect size estimates**	**Change in meta-analysis estimate by including the study**	**SE**	***p***
Anisfeld (2001)	12	-.47	.77	.539
Barbosa (2016)	4	-.75	.76	.322
Coulon (2013)	8	.77	.71	.281
Field (1982)	5	.00	.77	.968
Field (1983)	3	-.02	.77	.982
Fontaine (1984)	6	-.69	.77	.373
Heimann (1985)	4	-.42	.80	.595
Heimann (1989)	8	-.62	.71	.385
Heimann (1998)	2	.17	.79	.824
Heimann (2019)	6	-.70	.76	.354
Kennedy-Costantini (unpublished)	4	-.64	.76	.400
Koepke (1983)	4	-.69	.78	.375
Legerstee (1991)	4	-.77	.78	.327
McKenzie (1983)	8	-.24	.78	.756
Meltzoff (1977)	8	1.26	.70	.072
Meltzoff (1983b)	2	.19	.78	.813
Meltzoff (1989)	2	.48	.78	.535
Meltzoff (1992)	2	.19	.79	.809
Meltzoff (1994)	4	1.69	.69	.014*
Nagy (2007)	3	-.02	.78	.978

Nagy (2014)	3	.64	.69	.357
Oostenbroek (2016)	225	-.62	.68	.364
Reissland (1988)	2	.94	.83	.257
Soh (unpublished)	2	-.73	.76	.333
Soussignan (2011)	1	2.53	.86	.003*
Ullstadius (1998)	4	-.59	.78	.449

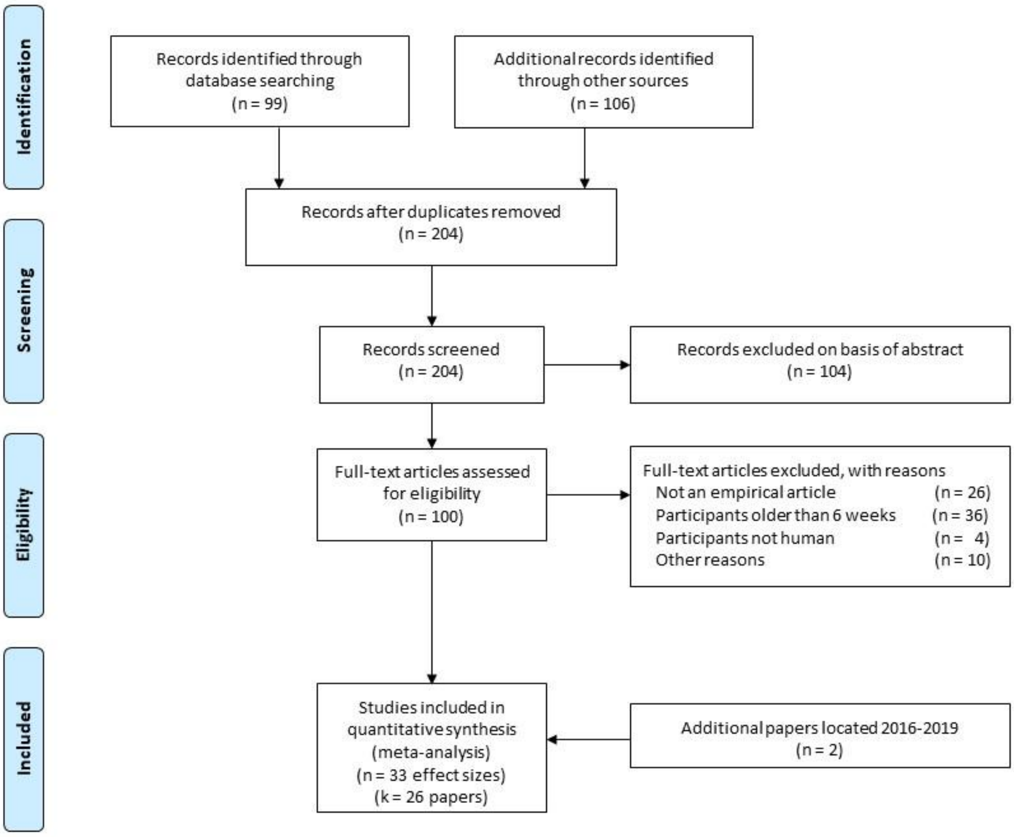
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 99)
Additional records identified through other sources (n = 106)
Records after duplicates removed (n = 204)
Records screened (n = 204)
Records excluded on basis of abstract (n = 104)
Full-text articles assessed for eligibility (n = 100)
Full-text articles excluded, with reasons
Not an empirical article (n = 26)
Participants older than 6 weeks (n = 36)
Participants not human (n = 4)
Other reasons (n = 10)
Studies included in quantitative synthesis (meta-analysis) (n = 33 effect sizes) (k = 26 papers)
Additional papers located 2016-2019 (n = 2)

Figure 1. Systematic search results and attrition of publications.

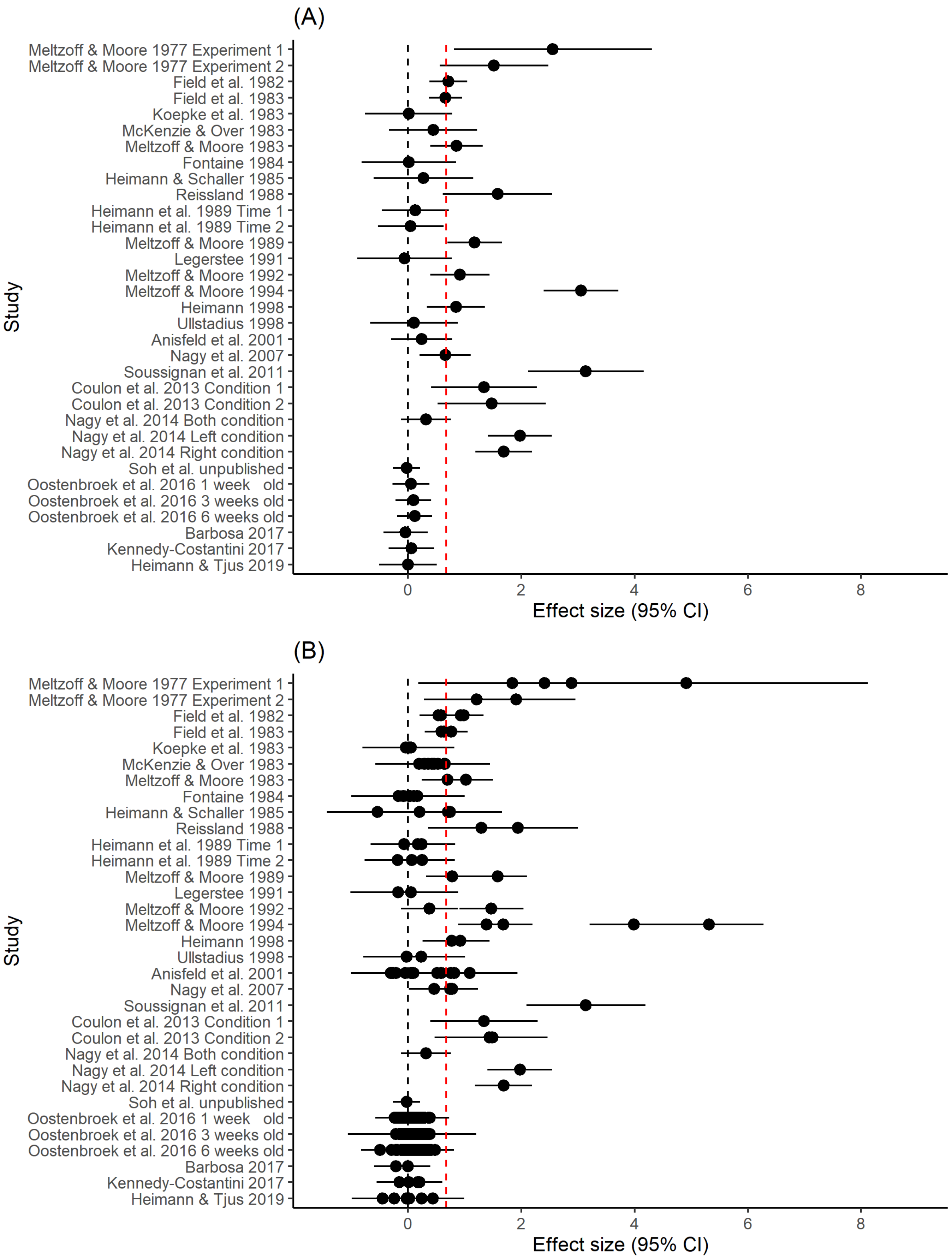

Figure 2. Forest plots of imitation studies showing (A) a reduced summary with one estimate per study, and (B) the full multi-level data with multiple estimates per study. The black dotted line indicates a zero effect; the red dotted line indicates the overall estimated effect.

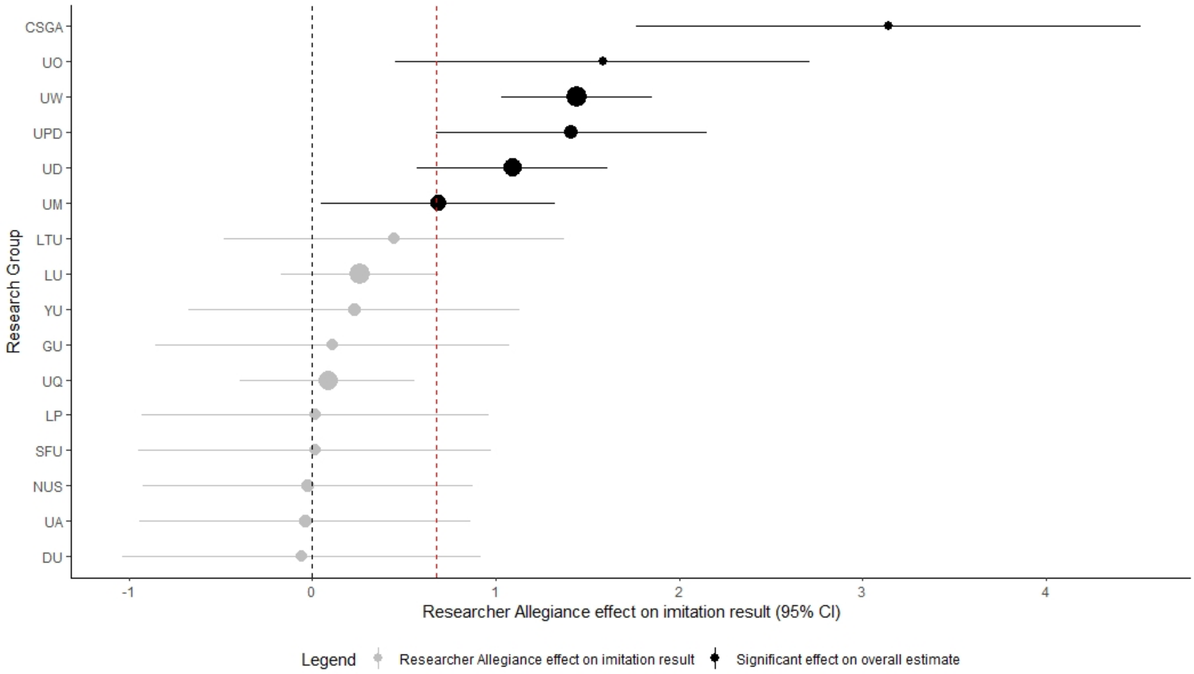

Figure 3. Effect of researcher allegiance on neonatal imitation effect estimates. Points show meta-regression effect sizes and 95% confidence intervals. Larger points indicate that this research group contributed more estimates to the meta-analysis and had greater weight on the overall result. Black points indicate the research groups reporting significant evidence of neonatal imitation. The black dotted line shows a zero effect; if the confidence interval crosses the black dotted line, that research group's papers do not collectively demonstrate a significant overall effect for neonatal imitation. The red dotted line indicates the overall (intercept) estimate for imitation found in the meta-analysis.

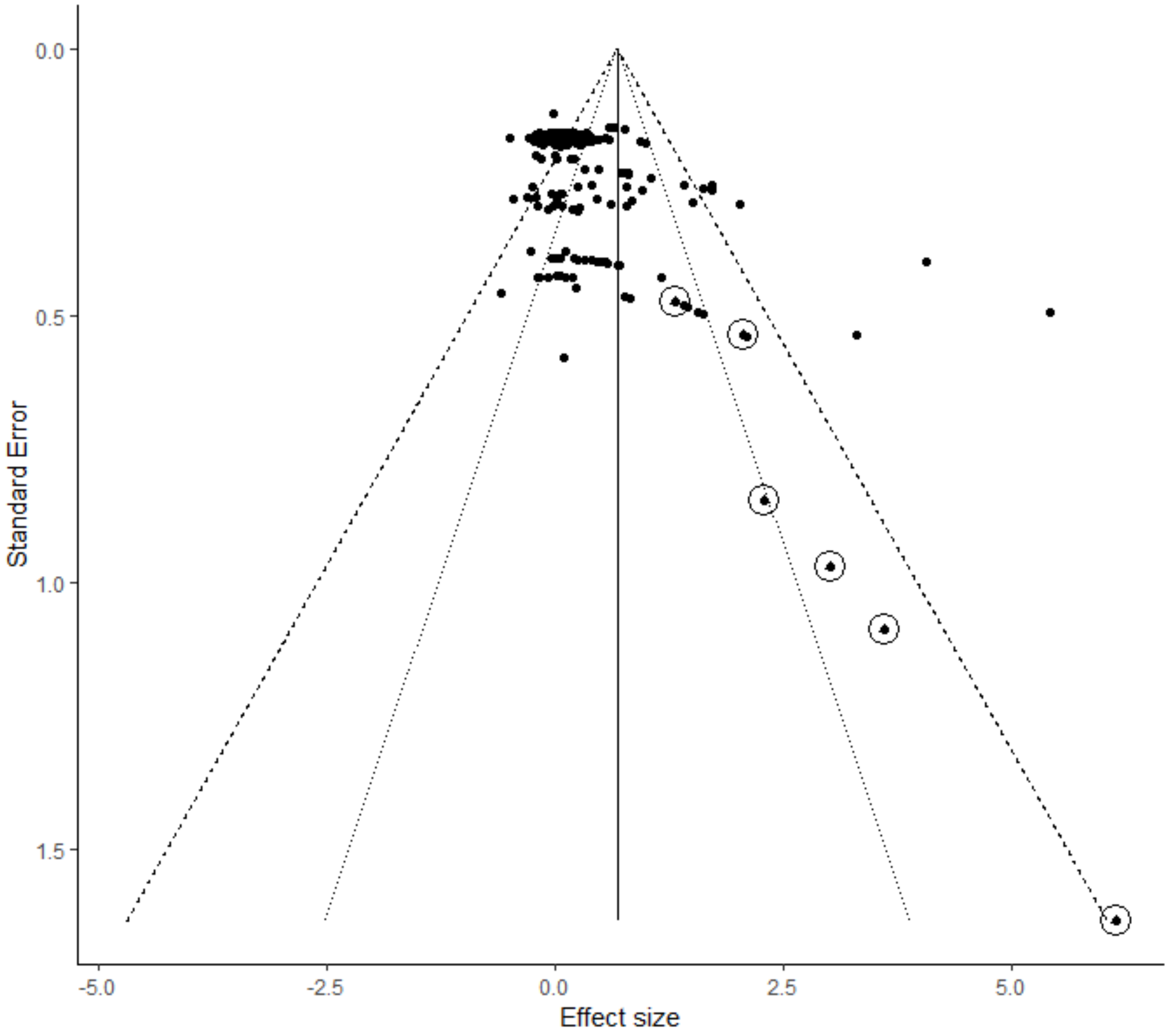

Figure 4. Funnel plot for all reported effect sizes of neonatal imitation of all gestures, showing significant asymmetry. Dotted lines show the 95% and 99% confidence intervals for the funnel plot. The vertical line shows the estimated overall effect size from the meta-analysis. The six circled effects are from Meltzoff and Moore's (1977) foundational paper on neonatal imitation.