University of Reading

Re-examination of Oostenbroek et al. (2016): evidence for neonatal imitation of tongue protrusion

Article

Accepted Version

Meltzoff, A. N., Murray, L., Simpson, E., Heimann, M., Nagy, E., Nadel, J., Pedersen, E. J., Brooks, R., Messinger, D. S., De Pascalis, L., Subiaul, F., Paukner, A. and Ferrari, P. F. (2018) Re-examination of Oostenbroek et al. (2016): evidence for neonatal imitation of tongue protrusion. Developmental Science, 21 (4). e12609. ISSN 1363-755X doi: 10.1111/desc.12609 Available at https://centaur.reading.ac.uk/73061/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1111/desc.12609

Publisher: Blackwell Publishing

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

University of Reading

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Re-examination of Oostenbroek et al. (2016) – Evidence for Neonatal Imitation of Tongue Protrusion

Andrew N. Meltzoff
University of Washington
Lynne Murray
University of Reading & University of Cape Town
Elizabeth Simpson
University of Miami
Mikael Heimann
Linköping University
Emese Nagy
University of Dundee
Jacqueline Nadel
Hôpital de la Salpêtrière
Eric J. Pedersen
University of Miami
Rechele Brooks
University of Washington
Daniel S. Messinger
University of Miami
Leonardo De Pascalis
University of Liverpool
Francys Subiaul
George Washington University
Annika Paukner
National Institute of Child Health and Human Development
Pier F. Ferrari
Université Claude Bernard Lyon 1

Author note

Andrew N. Meltzoff, Institute for Learning & Brain Sciences, University of Washington; Lynne Murray, Department of Psychology, University of Reading & Department of Psychology, University of Cape Town; Elizabeth Simpson, Department of Psychology, University of Miami; Mikael Heimann, Department of Behavioral Sciences and Learning, Linköping University; Emese Nagy, School of Psychology, University of Dundee; Jacqueline Nadel, Centre Emotion, Hôpital de la Salpêtrière, Paris, France; Eric J. Pedersen, Department of Psychology, University of Miami; Rechele Brooks, Institute for Learning & Brain Sciences, University of Washington; Daniel S. Messinger, Department of Psychology, University of Miami; Leonardo De Pascalis, Department of Psychological Sciences, University of Liverpool; Francys Subiaul, Department of Speech, Language, & Hearing Sciences, George Washington University; Annika Paukner, Eunice Kennedy Shriver National Institute of Child Health and Human Development; Pier F. Ferrari, Institut des Sciences Cognitives – Marc Jeannerod, Université Claude Bernard Lyon 1

Address for correspondence: Andrew N. Meltzoff, Institute for Learning & Brain Sciences, University of Washington (Box 357988), Seattle, WA 98195, USA; email: Meltzoff@uw.edu

Submitted for Publication. ***Please do not circulate.***

Research Highlights

- Oostenbroek et al. used an insensitive procedure to test neonatal imitation, demonstrating 11 acts in succession to 1, 3, 6, and 9-week-olds.
- Some target acts were not within the motor capabilities of neonates making them impossible to imitate.
- We identify 11 flaws in the experimental design that can be predicted to bias the results toward null effects, based on extant literature.
- We re-analyze the authors' data and find significant imitation of tongue protrusion at all 4 ages tested, despite the weak design.

Abstract

The meaning, mechanism, and functions of imitation in early infancy have been actively discussed since Meltzoff and Moore's (1997) report of facial and manual imitation by human neonates. Oostenbroek et al. (2016) claim to challenge the existence of early imitation and to counter all interpretations so far offered. Such claims, if true, would have implications for theories of social-cognitive development. Here we identify 11 flaws in Oostenbroek et al.'s experimental design that biased the results toward null effects. We requested and obtained the authors' raw data. Contrary to the authors' conclusions, new analyses reveal significant tongue protrusion imitation at all 4 ages tested (1, 3, 6, and 9 weeks old). We explain how the authors missed this pattern and offer 5 recommendations for designing future experiments. Infant imitation raises fundamental issues about action representation, social learning, and brain-behavior relations. The debate about the origins and development of imitation reflects its importance to theories of developmental science.

Keywords: Infant imitation, visual processing, motor behavior, perception-action, social learning

Re-examination of Oostenbroek et al. (2016) – Evidence for Neonatal Imitation of Tongue Protrusion

Introduction

In a paper in *Current Biology*, Oostenbroek et al. (2016) claim to present data showing that infants from 1 to 9 weeks of age do not imitate facial gestures such as tongue protrusion. The existence, mechanisms, and meaning of early imitation have been actively discussed since Meltzoff and Moore's (1977) report of facial and manual imitation by neonates. What makes the Oostenbroek et al. paper unique is its claim to counter all interpretations so far offered. The authors recognize that the imitation of tongue protrusion is the most common gesture reported in the literature, but they claim to challenge this phenomenon. In so doing, they argue against not only intermodal mapping and perception-action mechanisms for early imitation but all "leaner" interpretations including arousal, associative learning, and automatic reflexes. If no early behavioral matching exists, then these leaner accounts of the mediating processes must also be rejected.

Here we rebut Oostenbroek et al.'s sweeping claims. First, we show that the Oostenbroek et al. study has 11 flaws in the design that lead to an underestimation of infants' imitative competence. Second, we re-analyze their raw data (we thank the authors for providing these data) and show that there is, contrary to the authors' report, strong evidence for the imitation of tongue protrusion. These results lead to different conclusions from those drawn by Oostenbroek and colleagues. Third, we make recommendations that will help researchers design effective eliciting conditions in future studies of infant imitation, and we draw broader lessons about replication studies in developmental science.

11 Design Flaws in Oostenbroek et al. (2016)

There are 11 weaknesses in the experimental design and execution that bias the Oostenbroek et al. (2016) study towards null results.

(1) Too many stimuli used in a within-subjects design. The procedure Oostenbroek et al. used was long (11 minutes), which leads to neonatal fatigue and disengagement, and it involved too many rapidly changing stimuli. Specifically, 11 different gestures were shown to each neonate in a within-subjects design. Previous papers with positive effects have used fewer gestures (typically 1-4 different gestures); no previous study in the literature has attempted to demonstrate 11 different gestures in a within-subjects design, requiring the same neonate to motorically switch from one gesture to the other in a rapid fashion (for reviews, see Meltzoff & Moore, 1997; Nagy, Pilling, Orvos, & Molnar, 2013; Simpson, Murray, Paukner, & Ferrari, 2014).

This 11-model procedure can give rise to response "carry over." To circumvent the problem of infants' responses to one demonstration contaminating their response to a subsequent one, Meltzoff and Moore (1994) recommended a shift from a within-subjects design to an independent groups design. Oostenbroek et al.'s procedure of showing neonates 11 different models within one test session has no precedent in tests of imitation at *any age* in infancy. There is no scientific justification to think that neonates could succeed using the 11-model within-subjects design.

(2) Infants cannot imitate behaviors that they are incapable of producing. Oostenbroek et al. test for imitation of several acts that are impossible for neonates to produce. For example, human neonates cannot produce the vowel "*ee*" (as in "peep"), yet imitation of that vocalization was tested. Research on phonological development indicates that the "*ee*" vowel is produced

only after the vocal tract matures later in the first year (Kent & Murray, 1982; Lieberman, Crelin, & Klatt, 1972). Oostenbroek et al. also tested for the imitation of a tongue-clicking sound, but again, there is no evidence in the phonological literature that neonates can produce such sounds. It is logically impossible for infants to imitate behaviors that they cannot generate. The decision to model behaviors that infants are incapable of producing biases the study towards null results.

(3) Stimulus and response periods were too brief. The duration of stimulus presentation is critical for eliciting early imitation. This factor is especially important for young infants who may not immediately fixate on the model and need time to process it. A review paper of 23 studies of early imitation found that a stimulus-presentation period of 60 s or more yielded positive evidence for imitation in all studies, whereas modeling the gesture for 40 s or less resulted in findings of imitation in only 31% of studies (Anisfeld, 1991; see also Simpson et al., 2014). The maximum duration of modeling used by Oostenbroek et al. was 30s, and some infants received only 15 s exposure to the stimulus. Thus all the infants in this study received a suboptimal stimulus-presentation duration. The relevant guidelines were published prior to the Oostenbroek et al. study. The eliciting conditions used by Oostenbroek et al. could be predicted, based on the literature, to bias the results toward null effects.

The length of the response period—the time allowed for the infants to imitate—is also an important factor in imitation (Meltzoff & Moore, 1983b, 1997; Simpson, et al., 2014). Neonates require time to organize their motor responses to visual stimuli (Heimann, 2002; Meltzoff & Moore, 1983a, 1997; Nagy, Pal, & Orvos, 2014; Soussignan, Courtial, Canet, Danon-Apter & Nadel, 2011). To accommodate this latency, Meltzoff and Moore (1977; Study 2) used a 2.5 min response period, and subsequent designs honed this to an even longer period, using electronically-timed 4-min response periods to allow for the slow motor organization in neonates

(Meltzoff & Moore, 1983a, 1989). Oostenbroek et al. used a shorter period, varying between 15 – 60 s, depending on the experimenter's decision *in situ*. This short duration may have cut short infants' responses and contributed to the weak effects.

(4) Flawed response criteria were used. Oostenbroek et al. report that they adopted the response criteria used in previous work, but, in fact, the criteria deviated from published work in several critical ways, and the new criteria are problematic. There are four problems with the response criteria used (see Oostenbroek et al. *Supplemental Information* Table S1 for criteria).

First, instances in which infants watched the model and then looked away for >2 seconds and imitated were not counted as imitation. The exclusion of motor behavior during a look-away was not done in any previous study reporting infant imitation. According to some reports, participants may look away when they are processing information or organizing a motor response (e.g., Previc, Declerck, & de Brabander, 2005; Simpson, Paukner, Suomi, & Ferrari, 2014). There is little justification for discounting imitative responses that occur 2 s after the infant looks away from the model.

Second, a lack of objectivity in response criteria could contribute to null effects. For example, the code used to determine whether infants imitated the tongue-click sound was: "A clear backward movement of the tongue to the roof of the mouth that produces an audible tongue click." But the authors had no way of seeing into the infant's mouth and could not have determined "a clear backward movement." The "mmm" sound was only scored if the infant "clearly and purposely produces a vocal gesture matching a 'mmm' sound." How purposefulness was assessed, especially in 1-week-old neonates, remains unclear.

Third, the scoring used for the infant mouth opening was problematic. Oostenbroek et al.'s requirement for scoring a full mouth opening was "the turning down of the sides of the mouth,"

which is questionable and does not match Meltzoff and Moore's (1983a, 1994) operational definition. Moreover, previous studies documented that the *duration* of infant mouth opening is an important response measure in 6-week-old imitation (Meltzoff & Moore, 1994). The wide-open mouth posture used in studies of mouth-opening imitation is a very distinctive act that involves a temporal component. Oostenbroek et al. did not score the durational aspects of the response. The distinction between frequency and duration measures and the utility of each is not unique to measuring the imitation of wide-open mouths; it has precedents in studying other infant phenomena, including infant looking (Aslin, 2012), tactile exploration (Ruff, 1984), vocalizations (Kent & Murray, 1982), and mother-infant interaction (Messinger, Ruvolo, Ekas, & Fogel, 2010).

Fourth, the response criteria used to assess imitation were poorly justified in several cases. For example, the imitation of a manual gesture was only counted if the infant imitated at "midline" and not when the hand was extended out to the side. The imitation of the happy and sad faces was discounted if the infant vocalized.

(5) Distracting visual stimuli interfered with manual imitation. As displayed in Oostenbroek et al. (2016), the tests of finger movements had the experimenter's face as a visual distracter. The adult held her hand directly in front of her face when demonstrating the finger movements (Figure 1, E-F). Young infants' visual attention is selectively drawn to faces (e.g., Farroni, Johnson, Menon, Zulian, Faraguna, Csibra, 2005; Valenza, Simion, Cassia, & Umiltà, 1996). Inserting a face in infants' visual field could dampen infants' imitation of manual movements.

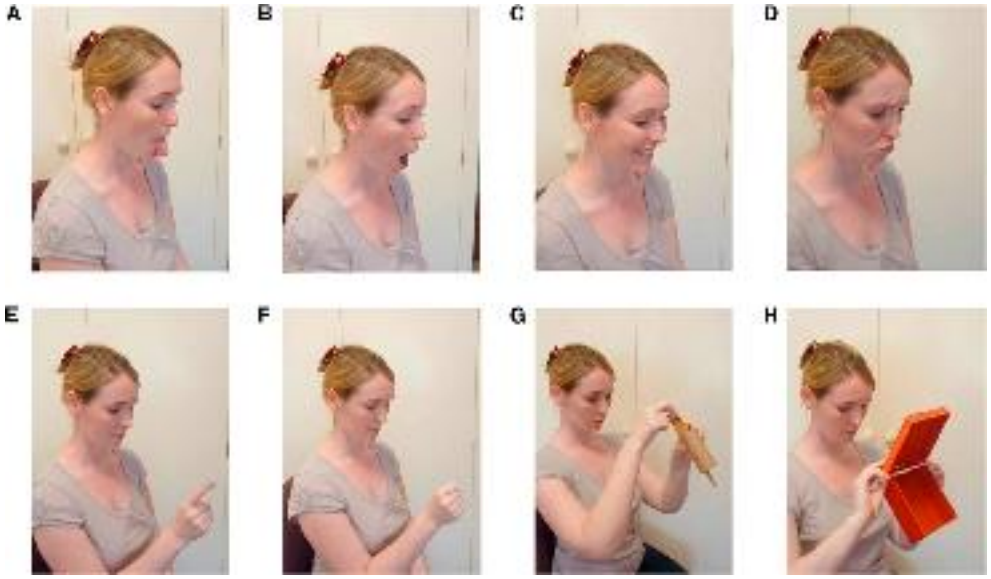

Figure 1 The face is a salient visual stimulus to young infants. In Oostenbroek et al.'s procedure, the adult's face was directly behind the finger movements (panels E and F), which may distract infants and dampen manual imitation. (From Oostenbroek et al., 2016, p. 1335.)

(6) Infants were tested in an unsatisfactory state of drowsiness. The main body of the Oostenbroek et al. paper reports that infants were tested when "in a suitable arousal state" (p. 1338). However, the *Supplemental Information: Missing Data and Subject Exclusion Criteria* reveals that infants were tested even if they were in a state of drowsiness, as defined in their study by Brazelton & Nugent's "state 3." According to Brazelton and Nugent's (1995) definition, state 3 entails, "Drowsy or semi-dozing; eyes may be open but dull and heavy-lidded, or closed, eyelids fluttering....Dazed look when the infant is not processing information and is not fully alert" (p. 15). Infants cannot imitate if they do not process the visual demonstrations. This confound of testing infants in an unsatisfactory state is likely to have biased the study toward null effects.

(7) Uncontrolled exposure to experimental stimuli is problematic. The Oostenbroek et al. study had procedural flexibility allowing infants to study the stimulus prior to the test, which is problematic. They state: "If the infant became sleepy or upset, testing was paused and calming

methods such as rocking, jiggling or walking the infant around the room were used to bring the infant back to a quiet alert state" (*Supplemental Information: Procedure*). Such walking around the room in the middle of the experiment opens up the possibility of experimenter bias, because the experimenter made these decisions *in situ*. Moreover, removing some infants, and not others, from the experimental setting changes their exposure to the adult tester (the stimulus). The literature highlights that exposure to the adult tester is a factor that must be controlled in studies of imitation. As noted in one publication: "Poor control over maternal leave-taking and the entrance of the experimenter was reported to dampen imitative responding in previous work with 6-week-olds….[thus] the infant was prevented from interacting with the experimenter (the experimental stimulus) before or between test sessions" (Meltzoff & Moore, 1994, p. 87). Appropriate control over the experimental stimulus (the experimenter's face) before and during the test was not achieved in Oostenbroek et al.'s design. [1]

(8) Post hoc subject selection occurred in the longitudinal sample. Oostenbroek et al.'s study design called for each infant to be tested starting at 1-week of age with repeated testing at 3, 6, and 9 weeks. Although some missing data are to be expected in longitudinal studies, the 11-min test at each age led to significant attrition. The authors included 64 infants (out of 106) in their longitudinal analyses, and there were questionable decisions about inclusion and exclusion for the 64 chosen for data analysis. Two of the 64 infants included in the longitudinal analysis (ID #28 and #60) were missing data for all of the models at a given age (one infant at 6 weeks and one at 9 weeks), and #28 had 45% of her data points missing across the four ages tested (infants were included if they had >50% of their data). Better justification is needed for selecting

[1] To underscore this point by analogy: In studies of infant visual attention, one avoids uncontrolled exposure to the visual test patterns before or during the experiment. In Oostenbroek et al., uncontrolled exposure to the social test stimulus (the experimenter) introduced noise, potentially weakening results.

these particular 64 infants for the longitudinal analysis and moreover for including infants who were in the unsatisfactory state of drowsiness (see #6 above). The underlying problem is that the study was too long and demanding (11 rapidly shifting models) for neonates, which led to post hoc subject selection issues. Only 25 of 106 infants actually completed the pre-specified longitudinal design (11 models x 4 ages).

(9) Significant deviations from the intended procedure occurred. Oostenbroek et al.'s intended procedure involved a 60 s trial for each gesture. As Oostenbroek et al. state: "Infants (*n* = 106) were presented with 11 models for 60 s each…" (p. 1134). The 60 s trial consisted of four 15 s intervals in a burst-pause manner (15 s modeling, 15 s passive face, 15 s modeling, 15 s passive face). However, in actuality, the experimenter determined the trial length *in situ* depending on the infants' state: "There were a number of occasions when an infant remained in a suitable arousal state for only part of the 60-second trial before the trial had to be abandoned" (*Supplemental Information: Dependent Variable Selection*). Infants who did not complete the planned 60 s trial, were handled in a questionable fashion. Because some infants had incomplete trials (< 60 s) and trial fragments were counted, the results were plotted as a mean response per 15 s. Using this average can be misleading: If an infant has a response of 0, it could have derived from one to four 15 s periods, but this information is lost in averaging. Moreover, previously published studies indicate that infants often take time to organize a matching response (see #3 above), yet a trial fragment (15 s) was not treated differently from a complete trial (60 s).[2]

(10) Test order was not counterbalanced. Oostenbroek et al. did not counterbalance the order of the models. There were 11 models shown but the tongue-protrusion and mouth-opening

[2] The authors' shared data file did not tag whether the data derived from a 15, 30, 45, or 60 s period, and therefore we cannot provide further analyses of this point.

models were *always* immediately adjacent to one another. Thus, the five orders used in the study did not follow a random or principled selection from the possible orders. Moreover, in the sample of infants the authors selected for the longitudinal data set (n = 64), the order was skewed (45 cases had the tongue-protrusion demonstration immediately prior to the mouth-opening demonstration and 19 cases had the reverse order).

(11) Neonates were balanced on the adult's lap, resulting in poor postural support. Adequate postural control is fundamental to studies with neonates. Oostenbroek et al. used unsatisfactory postural support. The neonates were balanced on the lap for all 11 demonstrations. The stimuli involving object-movement required that the experimenter use *both* hands to manipulate the stimulus (see Figure 2), thus infants could roll from side to side (similarly, neonates were balanced on the lap and one hand was used to show the manual gestures). The threat of postural imbalance is disruptive to young infants (von Hofsten, 1982, 2004): "Several reflexes have been identified that serve that purpose….They typically interrupt action" (von Hofsten, 2007, p. 56). In Meltzoff and Moore's experiments, a procedure was instituted to eliminate postural imbalance. As stated in the published work, neonates were well supported in a padded infant seat, which assured a stable posture (e.g., Meltzoff & Moore, 1983a, 1994). Also, Nagy et al.'s (2013) and Soussignan et al.'s (2011) papers affirmed the importance of postural control in neonatal imitation. Oostenbroek et al. ignored this aspect of neonatal testing, which would bias the study towards null results.

Figure 2 In the Oostenbroek et al. procedure, infants were balanced on the experimenter's lap leading to poor postural control (Photos from: Kennedy-Costantini, Slaughter, Nielsen, 2016.)

Re-Analyses of the Raw Data Reveal Evidence for Neonatal Imitation of Tongue Protrusion

Oostenbroek et al. (2016) tested 106 infants at 1 week of age and attempted to re-test them at three subsequent ages (3, 6, and 9 weeks). Some infants did not complete sufficient testing for Oostenbroek et al. to conduct longitudinal analyses. This yielded a large number of infants in their cross-sectional data set (varying *n*s at different time points) and a smaller subset of infants in their longitudinal data set. The main body of the paper reports the longitudinal analyses; the *Supplemental Information* (Table S4) contains the cross-sectional analyses. We conducted new statistical analyses of both of their data sets based on the raw data files the authors provided.

The re-analyses yield results that contradict a central claim of Oostenbroek et al.'s published paper. The paper claims that even for tongue protrusion, which "has produced the most consistent evidence for neonatal imitation in the literature" (p. 1335), "there is no evidence infants were imitating the specific model" (p. 1335). Our analyses of the raw data reveal evidence for the imitation of tongue protrusion. Moreover, we can specify how the authors missed these positive results. This is elaborated below. We start with the re-analyses of the cross-sectional data set.

Re-analyses of the Cross-Sectional Data Yields Significant Effects

Oostenbroek et al.'s (2016) Table S4 (top panel) presents data for the tongue-protrusion measure in the cross-sectional data set. To test for imitation, the authors compared the number of tongue protrusions infants produced when shown the tongue-protrusion demonstration (TP) to the number of tongue protrusions infants produced when shown each of the 10 other demonstrations (the controls). The 10 other demonstrations were all dynamic stimuli designed to attract infants' attention. The list was: mouth opening, inanimate cylinder protruding from a tube (mimicking tongue protrusion), hinged-box opening/closing (mimicking mouth opening), happy face, sad face, finger protrusion, manual grasping motion, and faces articulating an *mmm* sound, an *ee* sound, and a *tongue-click* sound.

Given their 11-model design, Oostenbroek et al. say that they faced a "dilemma" for their data analysis: "there is no widely accepted a priori reason to choose one control model over another" (p. 1335), and thus they were not sure "how to define a family of tests for the purpose of correcting *p*-values" (*Supplemental Information: Cross-sectional Analysis*). We find it puzzling, then, that the authors compared the TP demonstration to each other demonstration individually using 10 separate pairwise comparisons. If there is no a priori reason to choose one control over another, there are more informative tests. One can ask the question: Does the infant tongue protrusion response to the TP demonstration differ from the mean response to 10 other demonstrations that served as controls? Using their raw data, we tested this comparison at each age and found significant effects with paired *t*-tests (Figure 3).

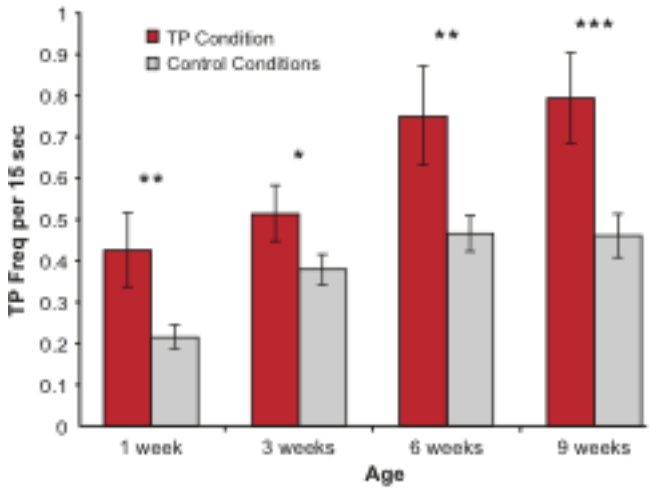

Figure 3 Infants produce significantly more tongue protrusions in response to the TP demonstration than to the mean of the 10 Controls at each age. * $p < .05$, ** $p < .01$, *** $p < .0005$. Error bars = *SE*.

As predicted by the hypothesis of infant imitation, there is significantly more infant tongue protrusion in response to the TP demonstration than to the Controls at each age. The results are: 1-week-olds, $t(74) = 2.75$, $p = .008$, $d = .32$; 3-week-olds, $t(80) = 2.16$, $p = .034$, $d = 0.24$; 6-week-olds $t(84) = 2.78$, $p = .007$, $d = .30$; 9-week-olds, $t(88) = 3.79$, $p = .0003$, $d = .40$. (These tests are also significant at each age using generalized linear mixed model [GLMM] analyses.)

The foregoing analysis is new, but we also draw readers' attention to Oostenbroek et al.'s Table S4 (top panel). The authors' approach was to conduct 40 individual pairwise comparisons (TP versus each of 10 controls at each of 4 ages). It is noteworthy that 39 of the 40 pairwise comparisons are in the direction predicted by the hypothesis of infant imitation. Infant responses to the TP demonstration were in the predicted direction (more infant tongue protrusions to the TP demonstration than to a control demonstration) for all 10 of the pairwise comparisons at 1 week, for 9/10 comparisons at 3 weeks, for 10/10 comparisons at 6 weeks, and for 10/10 comparisons at 9 weeks (Oostenbroek et al.'s Table S4).

Given the evidence for tongue protrusion imitation, one may wonder why the authors infer, "even our cross-sectional results do not provide any evidence for a true imitation effect" (p. 1335). There seem to be two streams of thought influencing the authors' inferences. First, infants do not show evidence of imitation for *all* 11 items demonstrated. However, some of the modeled behaviors are impossible for infants to produce (e.g., the vowel *ee*), and other models have problematic stimulus-presentation and response criteria (#4, 5, 7 above). A second reason the authors seem to discount the significant tongue protrusion results is that: "On no occasion, however, did the infants produce the gesture matching the model significantly more often than to *all* control models…" (p. 1335, emphasis added). This logic can be questioned. Although the authors are clearly conscious of the problem of inflating Type I error (i.e., false positives) associated with conducting many comparisons (40 pairwise comparisons), they seem to ignore the simultaneous problem of increasing Type II error (i.e., false negatives) by using a standard of evidence in which *all* of the individual comparisons must be significant. Consider the tongue protrusion response for the 9-week-olds (their Table S4, top panel). The table shows significant effects for 9 of the 10 pairwise comparisons (TP demonstration vs. each of 10 control conditions) ranging from $p < .001$ to .004, and the remaining comparison is in the predicted direction. The authors are holding out for 10/10 significant pairwise comparisons. However, by this logic there is no reason to stop at 10 comparisons; why not 100 control comparisons with any one failure refuting the hypothesis? [3]

[3] Assuming the authors' analyses had been powered at .80 (80% chance of detecting a true effect) and there was a true imitation effect, we would only expect to find approximately 8/10 significant results at each of the four time points. If one extends the authors' logic, it would suggest that a meta-analysis containing a single null or negative result undermines the hypothesis being tested, which is not a standard of evidence used in the field.

The authors could have compared the infant tongue-protrusion response to the TP demonstration versus the mean of the Controls to avoid their "dilemma" of 40 pairwise comparisons at each age. Our analyses show that TP is significantly different from the mean of the 10 Controls at each age tested. This buttresses previous reports of early tongue protrusion imitation (see reviews by Meltzoff & Moore, 1997; Nagy et al., 2013; Simpson et al., 2014), and also suggests that the tongue protrusion effect is not reducible to arousal. All 11 demonstrations used by Oostenbroek et al. were arousing dynamic stimuli with no a priori prediction of which would be more arousing than the other. The fact that infants produced significantly more tongue protrusions to the TP demonstration than to the mean of 10 controls—which included a variety of facial expressions, object movements, and auditory-visual events—contradicts the arousal account. (Oostenbroek et al. acknowledge as much in their *Supplemental Information*.)

Re-analysis of the Longitudinal Data Yields Significant Effects

A re-analysis of the longitudinal data set shows a systematic effect for tongue protrusion as well. Oostenbroek et al. used GLMM analyses to conduct pairwise contrasts of TP relative to each control condition, controlling for age. Again, since they provide no reason to prefer one control condition over another, the mean of all controls can be tested against TP, which is an informative test of the question of interest (i.e., did tongue-protrusion responses to the TP demonstration significantly differ, on average, from tongue-protrusion responses across all control conditions?). We acquired the syntax the authors used for their GLMM analyses of the longitudinal data set, and first reproduced exactly their results. Then, we modified their syntax to perform a post hoc linear contrast comparing TP (coded as 1) to all 10 control conditions (each coded as -0.1). The resulting coefficient tests the statistical significance of the difference between mean tongue-protrusion responses to the TP demonstration versus the overall mean

tongue-protrusion responses to the control demonstrations. This coefficient was statistically significant, *beta* = .45, *SE* = .09, $p < .0001$. [4]

Five Recommendations for Future Studies: Effective Eliciting Conditions for Studies of Early Imitation

Science depends on replications. In the spirit of paving the way for future investigations of neonatal imitation, we offer five design recommendations.

Recommendation #1: Number of models used in a within-subjects design. Showing neonates 11 models in a within-subjects design biases the study toward null results. Because contamination from earlier models to subsequent ones is a concern, an independent groups design can be useful, because only one model is demonstrated to each infant. This independent groups design has yielded especially strong results for early imitation (Meltzoff & Moore, 1994). Any attempt to use a within-subjects design should fully counterbalance the order of models and use a limited number of them.

Recommendation #2: Length of the test period. Infants do not imitate immediately, and research indicates that infants converge on the matching behavior over successive efforts (Meltzoff & Moore, 1997; Nagy et al., 2014). To accommodate such response sharpening, Meltzoff and Moore (1983a, 1989) used a 4-minute period. Although imitation may be documented at shorter latencies, our recommendation is to use 1.5 to 4 min so as to not cut short the response due to the slow motor organization in neonates.

[4] In the re-analysis we were careful to use the same infants ($n = 64$) infants that Oostenbroek et al. used in their longitudinal sample. Among these 64 infants, there were 25 infants who finished the designed study (11 models x 4 ages). We also analyzed the tongue protrusion response for this complete data set, using a two-way ANOVA with model and age as within-subjects factors. The results showed a highly significant effect of model, $F(10, 240) = 5.74$, $p < .0001$, $\eta_p^2 = .19$, a main effect for age, $p = .035$, $\eta_p^2 = .14$, and no significant model x age interaction. A planned contrast showed significantly more infant tongue protrusions to the TP model ($M = 0.75$, $SD = 0.52$) than to the 10 pooled controls ($M = 0.41$, $SD = 0.28$), $F(1, 24) = 15.62$, $p = .0006$, $\eta_p^2 = .39$.

Recommendation # 3: Control of the physical environment. Meltzoff and Moore's (1983b) methodological review of neonatal imitation listed four key issues: (a) the visual display should be presented against a homogenous (black, white, gray) backdrop to enhance attention to the face, (b) a spotlight should be used illuminate the adult's face (the stimulus), (c) distracting sounds should be eliminated from the test environment, and (d) parents should remain uninformed about the gestures under test to reduce practice. [5]

Recommendation #4: Control of the social environment. In tests of infant imitation, the stimulus is the adult experimenter. Infants should not receive uncontrolled access to the tester. This methodological point is key to eliciting neonatal imitation: "imitation is dampened if infants have competing expectations about the experimenter or his or her actions. Several steps were aimed at lessening such confusions" (Meltzoff & Moore, 1994, p. 87, which lists the procedures). Moreover, young infants develop expectancies about face-to-face interaction with adults, especially their mothers (Tronick, Als, Adamson, Wise, & Brazelton, 1978; Messinger et al., 2010; Murray, De Pascalis, Bozicevic, Hawkins, Sclafani, & Ferrari, 2016). These contingencies can interfere with a strictly imitative response (Meltzoff & Moore, 1992). We strongly recommend that studies of early imitation take measures to differentiate the mother and her familiar facial games from the experimenter. One approach previously recommended by Meltzoff and Moore (1992, 1994) was to use an experimenter with a different appearance (gender, hair color/style, glasses) than the primary caretaker. Similarly, in longitudinal studies some parents may be tempted to practice the gestures between visits. When Meltzoff and Moore

[5] Oostenbroek et al. (2016) instituted none of these previously published controls. Figures 1 and 2 show that a homogenous background is not common in home testing; homes also have disruptive sounds (siblings, street sounds, pets, household appliances). Oostenbroek et al. did not use a spotlight on the to-be-imitated stimulus, nor did they keep the parents blind to the gestures tested, possibly prompting practice by some parents for some of the gestures and not others. This allows unwanted noise in the study.

(1994) conducted a three-visit study, they kept the parents blind to the gestures, reducing noise in the data. (The effects of parental training and contingent responding are interesting to investigate in their own right, Murray et al., 2016).

Recommendation #5: Pilot testing of new procedures. On the one hand, investigators should seek to profit from published designs with effective eliciting conditions. On the other hand, innovative procedures are also desirable. If researchers wish to introduce a radically new design, it is useful to run a pilot study. If infant matching behavior cannot be elicited at *any age* piloted, perhaps it is appropriate to consider whether it is the infants' competence or the experimental design that deserves attention.

What Inferences Can We Draw?

Our re-analyses of the Oostenbroek et al. (2016) paper provide support for the imitation of tongue protrusion in early infancy. The robustness of this tongue-protrusion effect is illustrated by its occurrence despite design flaws that biased the study towards null findings. The tongue-protrusion effect was found both in Oostenbroek et al.'s cross-sectional data set and in their longitudinal data set. [6]

Our new analyses call for a substantial revision in the conclusions of the paper. We draw three more general lessons from the re-examination of the Oostenbroek et al. (2016) paper.

(1) The null hypothesis. An old truism reminds us that there can be many sources of null effects. Oostenbroek et al. thought they had only null effects. They did not. However, even if this had been the case, it would have been useful for readers had the authors provided a list of design

[6] The 11 design flaws may have undermined imitation of a wider range of gestures. For example, flaws #2, 4, 5, 7, 11 would have biased the results towards null effect for particular models while leaving tongue protrusion relatively unperturbed. Other recent experiments without these problems have replicated neonatal imitation of mouth opening (Coulon, Hemimou, & Streri, 2013), manual gestures (Nagy et al., 2014), and other acts (Simpson et al., 2014).

differences between their study and previous experiments reporting significant effects. Such material would point towards potentially informative factors for future investigation. In this case, there are many significant deviations from published, effective eliciting conditions for neonatal imitation (see reviews by Meltzoff & Moore, 1983b, 1997; Simpson et al., 2014). Authors reporting null effects or failures to replicate have a special responsibility to call readers' attention to significant procedural changes from previous experiments that may have contributed to the null effects and to discuss the "limitations" of their study. [7]

(2) Towards a cumulative developmental science. Some literature reviews in infancy research simply "count up" the number of positive versus negative results in an area. However, a more useful meta-analytic approach is to sort studies according to their scientific design and adherence to "best practices" in an area. For example, since neonates cannot imitate what they cannot produce, it is not useful to tally a study as a "failure to replicate" if it uses an act that is impossible for neonates to produce. Similarly, since it is already known that short response periods are associated with weaker results in studies of neonatal imitation, the poor results based on 11 short-duration demonstrations might be put down to an insensitive design, rather than a failure to replicate. Ultimately, developmental scientists seek to create a cumulative science that both evaluates and profits from previous work. Novel designs can be a step forward; but they can be a step backward if they simply re-instate inadequate eliciting conditions that have already been identified, discussed, and corrected over the course of previous research programs.

[7] The 11-model protocol had never been used before. There was no reason to think it would be successful with neonates. Indeed we suggest that 12-month-olds would fail using this design, an age at which Piaget (1962) reported imitation of a range of facial gestures. Many of the 11 flaws in this study could be predicted to introduce noise into the data, based on previously published literature. The known weakness could have been listed as possible "limitations."

(3) Advancing developmental science. When young infants see an adult produce tongue protrusions it induces them to produce tongue protrusions themselves. A central question is what processes mediate this reaction? We have proposed accounts that address this question (see reviews by Meltzoff & Moore, 1997; Simpson et al., 2014). There are at least six open questions about early imitation that have implications for theories in developmental science. (a) What mechanisms underlie early imitation? (b) What functions does it serve? (c) Is early imitation a social response? (d) How does early imitation contribute to the growth of social cognition? (e) Are there individual differences in imitation and its development? (f) What are the neural correlates of infant imitation (Ferrari et al., 2012; Marshall & Meltzoff, 2014, 2015; Meltzoff & Moore, 1997)?

The phenomenon of infant imitation raises fundamental issues about action representation, self-other mapping, and social learning. An active debate about the origins and development of infant imitation may reflect its importance to theories of developmental science.

References

Anisfeld, M. (1991). Review: Neonatal imitation. *Developmental Review, 11*, 60-97. doi:10.1016/0273-2297(91)90003-7

Aslin, R. N. (2012). Infant eyes: A window on cognitive development. *Infancy, 17*, 126-140. doi:10.1111/j.1532-7078.2011.00097.x

Brazelton, T. B., & Nugent, J. K. (1995). *Neonatal behavioral assessment scale*. London: Cambridge University Press.

Coulon, M., Hemimou, C., & Streri, A. (2013). Effects of seeing and hearing vowels on neonatal facial imitation. *Infancy, 18*, 782-796. doi:10.1111/infa.12001

Farroni, T., Johnson, M. H., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns' preference for face-relevant stimuli: Effects of contrast polarity. *Proceedings of the National Academy of Sciences, 102*, 17245-17250. doi:10.1073/pnas.0502205102

Ferrari, P. F., Vanderwert, R. E., Paukner, A., Bower, S., Suomi, S. J., & Fox, N. A. (2012). Distinct EEG amplitude suppression to facial gestures as evidence for a mirror mechanism in newborn monkeys. *Journal of Cognitive Neuroscience, 24*, 1165-1172. doi:10.1162/jocn_a_00198

Heimann, M. (2002). Notes on individual differences and the assumed elusiveness of neonatal imitation. In A. N. Meltzoff & W. Prinz (Eds.), *The imitative mind: Development, evolution, and brain bases* (pp. 74-84). Cambridge: Cambridge University Press.

Kennedy-Costantini, S., Slaughter, V., & Nielsen, M. (2016, May). *Why are you copying me? Functional explanations for neonatal imitation*. Poster presented at the XX International Congress on Infant Studies, New Orleans.

Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *Journal of the Acoustical Society of America, 72,* 353-365. doi:10.1121/1.388089

Lieberman, P., Crelin, E. S., & Klatt, D. H. (1972). Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee. *American Anthropologist, 74,* 287-307.

Marshall, P. J., & Meltzoff, A. N. (2014). Neural mirroring mechanisms and imitation in human infants. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*: 20130620. doi:10.1098/rstb.2013.0620

Marshall, P. J., & Meltzoff, A. N. (2015). Body maps in the infant brain. *Trends in Cognitive Sciences, 19,* 499-505. doi:10.1016/j.tics.2015.06.012

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198,* 75-78. doi:10.1126/science.198.4312.75

Meltzoff, A. N., & Moore, M. K. (1983a). Newborn infants imitate adult facial gestures. *Child Development, 54,* 702-709. doi:10.2307/1130058

Meltzoff, A. N., & Moore, M. K. (1983b). The origins of imitation in infancy: Paradigm, phenomena, and theories. In L. P. Lipsitt & C. K. Rovee-Collier (Eds.), *Advances in infancy research* (Vol. 2, pp. 265-301). Norwood, NJ: Ablex.

Meltzoff, A. N., & Moore, M. K. (1989). Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Developmental Psychology, 25,* 954-962. doi:10.1037/0012-1649.25.6.954

Meltzoff, A. N., & Moore, M. K. (1992). Early imitation within a functional framework: The importance of person identity, movement, and development. *Infant Behavior & Development, 15,* 479-505. doi:10.1016/0163-6383(92)80015-M

Meltzoff, A. N., & Moore, M. K. (1994). Imitation, memory, and the representation of persons. *Infant Behavior and Development, 17,* 83-99. doi:10.1016/0163-6383(94)90024-8

Meltzoff, A. N., & Moore, M. K. (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting, 6,* 179-192. doi:10.1002/(SICI)1099-0917(199709/12)6:3/4<179::AID-EDP157>3.0.CO;2-R

Messinger, D. M., Ruvolo, P., Ekas, N. V., & Fogel, A. (2010). Applying machine learning to infant interaction: The development is in the details. *Neural Networks, 23,* 1004-1016. doi:10.1016/j.neunet.2010.08.008

Murray, L., De Pascalis, L., Bozicevic, L., Hawkins, L., Sclafani, V., & Ferrari, P. F. (2016). The functional architecture of mother-infant communication, and the development of infant social expressiveness in the first two months. *Scientific Reports* 6: 39019. doi:10.1038/srep39019

Nagy, E., Pal, A., & Orvos, H. (2014). Learning to imitate individual finger movements by the human neonate. *Developmental Science, 17,* 841-857. doi:10.1111/desc.12163

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,* 1628-1638. doi:10.1037/a0031127

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,* 1334-1338. doi:10.1016/j.cub.2016.03.047

Piaget, J. (1962). *Play, dreams and imitation in childhood* (C. Attegno & F. M. Hodgson, Trans.). New York, NY: Norton.

Previc, F. H., Declerck, C., & de Brabander, B. (2005). Why your "head is in the clouds" during thinking: the relationship between cognition and upper space. *Acta Psychologica, 118,* 7-24. doi:10.1016/j.actpsy.2004.10.012

Ruff, H. A. (1984). Infants' manipulative exploration of objects: Effects of age and object characteristics. *Developmental Psychology, 20,* 9-20.

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. doi:10.1098/rstb.2013.0289

Simpson, E. A., Paukner, A., Suomi, S. J., & Ferrari, P. F. (2014). Visual attention during neonatal imitation in newborn macaque monkeys. *Developmental Psychobiology, 56,* 864-870. doi:10.1002/dev.21146

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,* 385-394. doi:10.1111/j.1467-7687.2010.00984.x

Tronick, E., Als, H., Adamson, L., Wise, S., & Brazelton, T. B. (1978). The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child Psychiatry, 17,* 1-13. doi:10.1016/S0002-7138(09)62273-1

Valenza, E., Simion, F., Cassia, V. M., & Umiltà, C. (1996). Face preference at birth. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 892-903. doi:10.1037/0096-1523.22.4.892

von Hofsten, C. (1982). Eye-hand coordination in the newborn. *Developmental Psychology, 18,* 450-461. doi:10.1037/0012-1649.18.3.450

von Hofsten, C. (2004). An action perspective on motor development. *Trends in Cognitive Sciences, 8,* 266-272. doi:10.1016/j.tics.2004.04.002

von Hofsten, C. (2007). Action in development. *Developmental Science, 10*, 54-60. doi:10.1111/j.1467-7687.2007.00564.x

Figure Legends

Figure 1 The face is a salient visual stimulus to young infants. In Oostenbroek et al.'s procedure, the adult's face was directly behind the finger movements (panels E and F), which may distract infants and dampen manual imitation. (From Oostenbroek et al., 2016, p. 1335.)

Figure 2 In the Oostenbroek et al. procedure, infants were balanced on the experimenter's lap leading to poor postural control (Photos from: Kennedy-Costantini, Slaughter, Nielsen, 2016.)

Figure 3 Infants produce significantly more tongue protrusions in response to the TP demonstration than to the mean of the 10 Controls at each age. * $p < .05$, ** $p < .01$, *** $p < .0005$. Error bars = *SE*.