Developmental Science

WILEY

Developmental Science

REGISTERED REPORT

Infants' Social Evaluation of Helpers and Hinderers: A Large-Scale, Multi-Lab, Coordinated Replication Study

Kelsey Lucca[1] | Francis Yuen[2] | Yiyi Wang[3] | Nicolás Alessandroni[4] | Olivia Allison[5] | Mario Alvarez[1] | Emma L. Axelsson[6] | Janina Baumer[7] | Heidi A. Baumgartner[8] | Julie Bertels[9] | Mitali Bhavsar[10] | Krista Byers-Heinlein[4] | Arthur Capelier-Mourguy[11] | Hitomi Chijiiwa[12] | Chantelle S.-S. Chin[2] | Natalie Christner[13] | Laura K. Cirelli[14] | John Corbit[15] | Moritz M. Daum[16,17] | Tiffany Doan[14] | Michaela Dresel[18] | Anna Exner[19] | Wenxi Fei[20] | Samuel H. Forbes[21] | Laura Franchin[22] | Michael C. Frank[23] | Alessandra Geraci[24] | Michelle Giraud[25] | Megan E. Gornik[26] | Charlotte Grosse Wiesmann[27] | Tobias Grossmann[5] | Isabelle M. Hadley[26] | Naomi Havron[28,29] | Annette M. E. Henderson[30] | Emmy Higgs Matzner[31] | Bailey A. Immel[32] | Grzegorz Jankiewicz[33] | Wiktoria Jędryczka[33] | Yasuhiro Kanakogi[12] | Jonathan F. Kominsky[34] | Casey Lew-Williams[35] | Zoe Liberman[32] | Liquan Liu[36,37,38] | Yilin Liu[39] | Miriam T. Loeffler[16,17] | Alia Martin[18] | Julien Mayor[40] | Xianwei Meng[41] | Michal Misiak[42,43] | David Moreau[30,44] | Mira L. Nencheva[23,35] | Linda S. Oña[45,46] | Yenny Otálora[47] | Markus Paulus[13] | Bill Pepe[48] | Charisse B. Pickron[31] | Lindsey J. Powell[48] | Marina Proft[49] | Alyssa A. Quinn[6] | Hannes Rakoczy[49] | Peter J. Reschke[50] | Ronit Roth-Hanania[51] | Katrin Rothmaler[27,52] | Karola Schlegelmilch[45,46] | Laura Schlingloff-Nemecz[34,53] | Mark A. Schmuckler[14] | Tobias Schuwerk[13] | Sabine Seehagen[19] | Hilal H. Şen[54] | Munna R. Shainy[23,55] | Valentina Silvestri[25] | Melanie Soderstrom[26] | Jessica Sommerville[56] | Hyun-joo Song[57] | Piotr Sorokowski[33] | Sandro E. Stutz[16,17] | Yanjie Su[58] | Hernando Taborda-Osorio[59] | Alvin W. M. Tan[23] | Denis Tatone[34] | Teresa Taylor-Partridge[60] | Chiu Kin Adrian Tsang[2,61] | Arkadiusz Urbanek[62] | Florina Uzefovsky[63] | Ingmar Visser[7] | Annie E. Wertz[32,46] | Madison Williams[4] | Kristina Wolsey[30] | Terry Tin-Yau Wong[61] | Amanda M. Woodward[64] | Yang Wu[14] | Zhen Zeng[38,65] | Lucie Zimmer[13] | J. Kiley Hamlin[2]

Correspondence: Kelsey Lucca (klucca@asu.edu)

Received: 12 November 2019 | **Revised:** 26 September 2024 | **Accepted:** 29 September 2024

Funding: Piotr Sorokowski was supported by Being Human Incubator funds; Yenny Otálora was supported by CI: 5348, funded by Universidad del Valle; Hernando Taborda-Osorio was supported by code 00010364 Universidad Javeriana; Nicolás Alessandroni was supported by Concordia Horizon Postdoctoral Fellowship; Hannes Rakoczy was supported by DFG RA 2155/7-2; Tobias Schuwerk was supported by DFG SCHU3060/2-1; Denis Tatone and Laura Schlingloff-Nemecz were supported by ERC Horizon 2020 742231 awarded to Gergely Csibra; Lindsey J. Powell was supported by Hellman Fund Fellowship; Kelsey Lucca was supported by funds from Arizona State University Department of Psychology; Mario Alvarez was supported by an APA SUPER fellowship; Florina Uzefovsky was supported by ISF 561/18; Liquan Liu was supported by MSCA-IF-798658; Foundation of Graduates in Early Childhood Studies under the Forest Hill section of the Trust; Yanjie Su was supported by National Natural Science Foundation of China, 32071075; Mark A. Schmuckler was supported by NSERC Discovery Grant; Melanie Soderstrom was supported by NSERC Discovery RGPIN-2023-04285 and RGPIN-05367-2019; Tobias Grossmann was supported by NSF 2017229; Annette M. E. Henderson was supported by PBRF grants, School of Psychology, University of Auckland; John Corbit was supported by NSERC 2023–05954; Moritz M. Daum was supported by SNF 10001G_20768; J. Kiley Hamlin was supported by SSHRC Partnership Development Grant 890-2020-0059 and a Social Sciences and Humanities Research Council of Canada (SSHRC) grant (12R20573); Hilal H. Şen was supported by University of Akureyri Internal Grant R2308; Jessica Sommerville was supported by a grant from NICHD (1R01HD076949-01); Terry Tin-Yau Wong was supported by University of Hong Kong, Seed Fund for Basic Research, 104006653; Hyun-joo Song was supported by grant NRF-2020S1A5A2A01042840; Annie E. Wertz was supported by funding from the Max Planck Society.

Keywords: experimental methods | infancy | moral development | reproducibility | social cognition | social development

Kelsey Lucca and Francis Yuen share the first authorship.

A full breakdown of CRediT contributions can be viewed here https://osf.io/qntyd/.

The authors from ManyBabies4 Consortium are listed in the byline.

For affiliations, refer to page 24.

ABSTRACT

Evaluating whether someone's behavior is praiseworthy or blameworthy is a fundamental human trait. A seminal study by Hamlin and colleagues in 2007 suggested that the ability to form social evaluations based on third-party interactions emerges within the first year of life: infants preferred a character who helped, over hindered, another who tried but failed to climb a hill. This sparked a new line of inquiry into the origins of social evaluations; however, replication attempts have yielded mixed results. We present a preregistered, multi-laboratory, standardized study aimed at replicating infants' preference for Helpers over Hinderers. We intended to (1) provide a precise estimate of the effect size of infants' preference for Helpers over Hinderers, and (2) determine the degree to which preferences are based on social information. Using the ManyBabies framework for big team-based science, we tested 1018 infants (567 included, 5.5–10.5 months) from 37 labs across five continents. Overall, 49.34% of infants preferred Helpers over Hinderers in the social condition, and 55.85% preferred characters who pushed up, versus down, an inanimate object in the nonsocial condition; neither proportion differed from chance or from each other. This study provides evidence against infants' prosocial preferences in the hill paradigm, suggesting the effect size is weaker, absent, and/or develops later than previously estimated. As the first of its kind, this study serves as a proof-of-concept for using active behavioral measures (e.g., manual choice) in large-scale, multi-lab projects studying infants.

1 | Introduction

As adults, we are quick to judge other individuals' actions as praiseworthy or blameworthy. These judgments have a pervasive impact on our social interactions—we gravitate toward and befriend individuals with a history of behaving prosocially and avoid those with a history of behaving antisocially. Notably, our judgments and selective social preferences are not limited to those with whom we directly interact. Humans readily judge individuals on the basis of the prosocial and antisocial actions they direct toward unrelated third parties, and even incur personal costs to punish those who behave antisocially (Fehr and Fischbacher 2003). From where does this propensity to morally evaluate others originate?

Historically, some scholars have contended that infants are born either amoral, with no moral sense, or immoral, motivated solely by selfish impulses (Freud et al. 1961; Kohlberg 1969; Piaget 1932). According to these perspectives, it is only through cognitive development and extensive socialization that humans come to develop an adult-like moral sense (reviewed in Brownell 2013; Carpendale and Lewis 2004). This proposal has been recently challenged by a series of studies on the social evaluative abilities of preverbal infants (reviewed in Margoni and Surian 2018). This work suggests that key precursors of full-fledged moral competencies, such as the ability to evaluate others on the basis of their prosocial or antisocial acts, may already be in place within the first year after birth.

The first study to suggest that preverbal infants engage in social evaluations presented 6- and 10-month-old infants with scenarios in which novel characters directed prosocial and antisocial acts toward a third party (Hamlin, Wynn, and Bloom 2007). Specifically, infants watched a puppet show featuring a "Climber" (a red wooden circle with googly eyes) who repeatedly tried but failed to climb to the top of a steep hill. Infants were shown two distinct events in alternation: a hindering event and a helping event. During hindering events, a Hinderer character (e.g., a blue wooden square with googly eyes) prevented the Climber from reaching its goal by pushing it down to the bottom of the hill. During helping events, a Helper character (e.g., a yellow wooden triangle with googly eyes; character identity was counterbalanced across infants) facilitated the Climber's goal by pushing it to the top of the hill. Helping and Hindering events were repeated until infants reached a preset habituation criterion. Finally, infants were presented with the Helper and the Hinderer and were prompted to choose between the two. Infants at both 6 and 10 months of age robustly reached for the Helper over the Hinderer (12 of 12 6-month-olds and 14 of 16 10-month-olds), suggestive of early social evaluation.

A key control condition tested whether preferences for the Helper were truly social in nature or whether they were based on low-level features of the display. In this control condition, the procedure was similar to the experimental condition, except that the animate Climber was replaced with an inanimate (eyeless) red circular ball that produced no self-propelled motion. On alternating events, this ball was pushed up or down the hill by the same square and triangle characters that played the roles of Helper and Hinderer in the original condition. Critically, in this condition, pushing up and down was present, but should not have been interpreted as helping and hindering, given that the ball did not exhibit cues of agency or goal-directedness. Here, infants did not demonstrate a preference for either character, suggesting that infants' evaluations of the Helper and Hinderer in the experimental condition were guided by the social consequences of their actions (Hamlin, Wynn, and Bloom 2007).

In recent years, research has extended these findings across a variety of paradigms and social scenarios. This work demonstrated that infants, as early as 5 months, prefer characters who helped, rather than hindered, an agent achieve different types of goals, including opening a box containing an appealing toy, retrieving a toy from a high shelf, and obtaining a preferred object (Hamlin and Wynn 2011; Hamlin et al. 2013; Woo et al. 2017). Infants' preference for prosocial individuals has also been shown to extend to morally relevant actions beyond helping and hindering; for example, 6–10-month-olds disprefer those who physically batter others but prefer those who prevent others from being battered (Kanakogi et al. 2017; see also Buon et al. 2014), and by 12–16 months, infants demonstrate a preference for agents who act fairly versus unfairly (e.g., distributed resources equally

Summary

- We examined infants' preferences for prosocial (helping) over antisocial (hindering) individuals through a large-scale, multi-lab, coordinated replication study.
- Using the ManyBabies framework for team-based science, we tested 1018 infants (567 usable; ages 5.5–10.5 months) from 37 labs across five continents.
- Infants did not prefer a Helper over a Hinderer, nor did they prefer a character who pushed up, versus pushed down, an inanimate object.
- This study provides evidence against infants' prosocial preferences in the hill paradigm, suggesting the effect is weaker, absent, and/or develops later, than previously estimated.

vs. unequally; Burns and Sommerville 2014; Geraci and Surian 2011; Lucca, Pospisil, and Sommerville 2018). As in the original Helper/Hinderer study, control conditions have often helped to rule out various low-level explanations for infants' preferences, suggesting that they were based on the social nature of the interactions.

Findings suggesting that infants possess precocious social evaluation capacities have led researchers to probe the replicability of these results, as well as the underlying explanation for infants' success in these tasks. Direct replication attempts have been met with varying levels of success. Though some studies have found that infants prefer prosocial individuals in manual choice and preferential-looking paradigms (Buon et al. 2014, with 10- and 29-month-olds; Chae and Song 2018, with 6- and 10-month-olds; Hamlin, Wynn, and Bloom 2010, with 3-month-olds; Loheide-Niesmann, Lijster, and Hall 2020, with 24-month-olds; Scola et al. 2015; with 12–24 and 24–36-month-olds; Shimizu et al. 2018, with 15–18-month-olds), others have not (i.e., Abramson et al. 2016, unpublished, with 9- and 18-month-olds; Cowell and Decety 2015, with 12- to 24-month-olds; Nighbor et al. 2017, with 5- to 16-month-olds; Salvadori et al. 2015, with 9-month-olds; Schlingloff, Csibra, and Tatone 2020, with 14- to 16-month-olds; Shimizu et al. 2018, with 6-, 9-, and 12-month-olds; Vaporova and Zmyj 2020, with 9- and 14-month-olds). These disparate findings have raised critical questions about the robustness of the effect, as well as the conditions under which the effect can be elicited.

Other research has focused on identifying the reason for infants' choices in social evaluation tasks. Most notably, Scarf and colleagues (2012a) raised the possibility that infants' preference for helping characters could be explained not by a preference for helpful over unhelpful agents, but by perceptual features that co-occurred with the helping action in Hamlin and colleagues' (Hamlin, Wynn, and Bloom 2007) experiment. Specifically, they noted that the climbing character bounced after being helped, but not after being hindered. Thus, infants may have selected the Helper solely due to its association with a positively valenced stimulus (i.e., the bouncing event). To examine this possibility, Scarf and colleagues (2012a) conducted a series of experiments with 10-month-old infants that closely matched the original study, but with additional conditions in which bouncing actions also occurred during hindering events. Results revealed that infants preferred characters associated with bouncing, regardless of the type of event (helping or hindering) in which the action occurred. These results suggested that infants' evaluations may be based on physical rather than social aspects of the displays. Subsequent research challenged this interpretation (cf. Hamlin 2015), noting that stimuli differences may have hindered infants' ability to represent the Climber's behavior as goal-directed. Nevertheless, the study by Scarf and colleagues (2012a) highlights that infants' evaluations in these tasks could be based on features of the events devoid of sociomoral significance, thereby confirming the necessity of proper control conditions in studies claiming to examine the roots of social evaluation in infancy.

A recent meta-analysis aimed to provide an estimate of the effect size of infants' preference (or lack thereof) for prosocial over antisocial characters, as well as to provide insights into potential moderators of the effect (Margoni and Surian 2018; see also Holvoet et al. 2016). The meta-analysis included data from 26 studies (reporting a total of 61 effect sizes) in which a prosocial agent was defined in various ways, including helping (vs. hindering), giving (vs. taking), or distributing goods fairly (vs. unfairly), and in which preference was defined either by selective reaching or selective helping. Overall, the estimated average proportion of infants who preferred the prosocial character was 0.68, 95% CI [0.64, 0.72].

The overall effect size did not vary as a function of age (4 to 32 months), the type of dependent variable (reaching vs. helping), target type (foam shapes, hand puppets, or human experimenters), or type of stimulus presentation (real events or video displays). In contrast, infants' preference for prosocial characters did depend on the type of action presented. Studies depicting giving versus taking yielded larger effect sizes than did studies depicting helping versus hindering (although this finding should be interpreted with caution due to the small number of datapoints for giving vs. taking). The authors also noted higher effect sizes in studies with small sample sizes ($N = 16$ or fewer), suggestive of a file-drawer problem, as well as in the laboratories that produced the original studies on the topic (e.g., Hamlin's), indicative of a lab effect. Finally, they explored for evidence of a publication bias. Although most of the examined effects were published ($N = 44$), several were unpublished ($N = 17$). A common "trim-and-fill" procedure (Duval and Tweedie 2000) provided evidence of a publication bias (but see Carter et al. 2019). When adjusted for publication bias, the meta-analytic estimate was slightly smaller, revealing an average proportion of choice for the Helper character of 0.64, 95% CI [0.60, 0.69].

The results of the meta-analysis raised several important directions for future work. Since nearly half of the studies included in the meta-analysis were conducted by a single lab using small sample sizes ($N = 16$ infants per study), replication studies that incorporate a higher diversity of labs as well as larger samples are clearly needed. Moreover, some of the studies incorporated in the meta-analysis were failed replications in which no preference for prosocial individuals was observed. These failures suggest either that the true effect size might be smaller than originally thought (or nonexistent), and/or that slight variations in method or population were responsible for the failures (Makel, Plucker, and Hegarty 2012). Moreover, variability between studies can also be partially explained by sampling error or chance alone

(see Margoni and Shepperd 2020; Stanley and Spence 2014). Finally, the meta-analysis did not examine infants' behavior in control conditions, leaving open questions about whether infants' preferences are truly based on the social aspects of prosocial and antisocial actions.

The current study set out to address these outstanding questions as well as provide new insights into infants' early social evaluations. Specifically, we first aimed to establish a precise estimate of the true effect size of infants' preference for helping over hindering agents. Second, we sought to determine whether infants' preferences are social in nature; namely, that they require that positive and negative acts be directed toward agentic third parties capable of goal-directed behavior. To achieve these objectives, we conducted a large-scale, multi-site replication study of infants' preferences for helping characters, with a preregistered methodological and analytical plan (https://osf.io/qntyd/). Our multi-site replication approach provides crucial insights beyond those of the existing systematic reviews (Holvoet et al. 2016; Margoni and Surian 2018). First, by using a consistent methodology across laboratories around the world, we are able to more precisely identify sources of variation in infants' social evaluations (e.g., age, geographic location) that go beyond variation in the stimuli and experimental procedure, since each lab in our study will follow the same protocol and use the same video stimuli. Second, this approach allows us to compare infants' preferences for characters across social *and* nonsocial contexts, enabling us to measure whether infants' preferences are driven by social features of the helping/hindering events versus nonsocial or perceptual aspects of the events. Finally, given that meta-analyses in psychology have been shown to report effect sizes approximately three times as large as preregistered multi-lab replication projects (e.g., due to publication bias or selective reporting), our approach will allow us to obtain a more accurate estimate of the true effect size of infants' preference for Helpers (Kvarven, Strømland, and Johannesson 2020).

We utilized the ManyBabies model of coordinated replication efforts (Byers-Heinlein et al. 2020; Frank et al. 2017; The ManyBabies Consortium 2020; see also ManyLabs: Klein et al. 2014), wherein a hypothesis of interest is chosen and explored by a large group of interested laboratories, all following a standardized protocol in order to create a data set larger than any one laboratory could produce on its own. This method allows for the exploration of both participant- and laboratory-level variables. The ManyBabies model strives to adhere to Open Science principles and practices. For instance, all stimuli, protocols, code, and data are shared on open-access repositories. As opposed to exactly reproducing previously published methodologies, group-level decision-making is used in order to converge on a method that provides the best possible test of a hypothesis of interest. Laboratories from across the globe, especially from countries that are traditionally underrepresented in developmental research, are invited to contribute at all research stages. Whereas these collective efforts focus on carefully deciding and standardizing a study's critical manipulations, individual labs still utilize their own general research practices (e.g., recruitment strategies, research assistant training).

To examine whether infants do indeed engage in social evaluation of prosocial and antisocial characters, we chose to replicate the hill study by Hamlin and colleagues (Hamlin, Wynn, and Bloom 2007). This study was selected because (1) it is the most widely cited demonstration of infant social evaluation in the literature; (2) it has been successfully replicated in subsequent research by the original lab (Hamlin, Wynn, and Bloom 2010) and at least one independent laboratory (Chae and Song 2018), but not by others (Cowell and Decety 2015; Scarf et al. 2012a; Schlingloff, Csibra, and Tatone 2020); (3) the effect has been reported in studies employing multiple response measures (including preferential looking, anticipatory looking, and selective reaching) and different presentation formats (video stimuli presented on screen rather than live displays: Hamlin 2015). We reasoned that video stimuli would be easier to utilize on a global scale.

Consistent with the ManyBabies goal of conducting the best possible test of a given hypothesis as opposed to exact replications, we made several modifications to the original Hamlin and colleagues (Hamlin, Wynn, and Bloom 2007) paradigm. First, as just noted, we utilized a filmed puppet show as opposed to a live puppet show as in the original study. Using prerecorded stimuli standardizes the stimuli presented to infants across laboratories, thereby ensuring that any differences in results cannot be attributed to variations in habituation events. Further, videotaping these events rather than producing them in real-time allowed us to, within condition, match the speed and timing of the pushing-up and down actions along with the overall exposure to the push-up/push-down characters down to the millisecond. This method of presentation also increases the number of labs eligible for participation because it substantially reduces barriers to participation, such as financial/time constraints involved in purchasing puppet stage materials, constructing a puppet stage, and training researchers to execute a live puppet show. The video stimuli used here were recorded in Hamlin's lab and closely matched videos used in a successful replication (Hamlin 2015) of the live puppet show paradigm (Hamlin, Wynn, and Bloom 2007). Because these two studies (Hamlin 2015; Hamlin, Wynn, and Bloom 2007) differed only in the stimulus presentation modality and found results with comparable effect sizes, we expected that the effect size would not be moderated by this decision. Importantly, meta-analytic results also found no moderating effect of the modality of stimulus presentation (i.e., animations, videotaped, and real events; Margoni and Surian 2018).

Second, our design implemented controls for perceptual differences between helping and hindering events that were not present in the initial Hamlin and colleagues (Hamlin, Wynn, and Bloom 2007) study. As previously discussed, Scarf and colleagues (2012b) argued that infants' preferences in the hill paradigm were due to the Climber character bouncing after being helped but not after being hindered (but see Hamlin 2015, for evidence against these criticisms). To avoid this issue, our study utilized videos in which the Climber remained motionless, instead of bouncing, upon reaching its final position.

Third, rather than including two separate age groups (6- and 10-month-olds), as in the original study, we included a single group of infants ranging from 5.5 to 10.5 months. This age range was selected for several reasons. First, a manual reaching choice task can be used across this age window, allowing us to fully standardize the task across all infants. Second, infants in this age range demonstrate sensitivity to the causal power of

agents (Liu, Brooks, and Spelke 2019), and to both successful (e.g., Woodward 1998) and failed goal-directed actions (e.g., Brandone and Wellman 2009; Hamlin, Hallinan, and Woodward 2008). Third, although Margoni and Surian (2018) did not find a significant influence of age on infants' preference for prosocial individuals, several successful and failed replications fall within this age range (Hamlin 2015; Hamlin et al. 2011; Salvadori et al. 2015; Scarf et al. 2012a). Thus, including this broad age range allowed us to assess whether there are developmental changes in infants' preferences for prosocial others. Finally, as recruiting participants across a broad age range is presumably easier than recruiting within a narrow age range, we selected a wide age window to maximize the number of laboratories able to participate.

As in the original Hamlin and colleagues (Hamlin, Wynn, and Bloom 2007) study, we included a nonsocial control condition to examine whether infants' preferences are driven by the social aspects of helping versus hindering actions as opposed to nonsocial perceptual features of the displays. In the control condition, infants viewed events similar to the helping and hindering events of the social condition, but with several notable differences. Most critically, the Climber was replaced by an inanimate, eyeless object that did not engage in self-propelled motion. Specifically, infants viewed an inert red ball being pushed up or down the hill by triangle and square agentic characters with eyes. Based on the estimate from Margoni and Surian's (2018) meta-analysis, we predicted that approximately 64% of infants would choose the Helper in the social condition where the animate Climber, a red ball with eyes, demonstrated an unfulfilled goal to climb the hill. We predicted that infants would not demonstrate a preference for the character who pushed an inanimate red ball (that had no eyes and demonstrated no goal-directed behavior) in the nonsocial control condition. Relatedly, we also predicted significantly greater preference for the Helper in the social compared to the pusher-upper in the nonsocial condition.

The social and nonsocial videos were designed to convey fundamentally different events—helping/hindering an animate character versus pushing an inanimate character up/down; therefore, it was necessary that they differed in several ways. First, we had to ensure that the ball was perceived as animate in the social videos, and as inanimate in the nonsocial videos. To do so, social videos included a hill-climbing action at the start, which demonstrated the ball's goal to go uphill. The nonsocial videos do not have this portion of the video, since the ball is inanimate and not capable of self-propelled motion. This difference led to the nonsocial videos being 4.4 s shorter than the social videos. Although the timing of the nonsocial videos could have been matched to the social videos by introducing additional still frames and/or adding in novel actions, we reasoned that these modifications might lead to inattentiveness and fussiness in the nonsocial displays, insofar as they do not add anything directly relevant to, or may even hamper the interpretability of the push-up or push-down character's goals. Despite the overall length across social and nonsocial videos, the length of the videos was nevertheless equated within condition (i.e., both social videos are 13.3 s and both nonsocial videos are 8.9 s), and the amount of time the Helper/Push-up and Hinderer/Push-down characters are on stage is exactly matched within and closely matched across conditions (social = 4.7 s, nonsocial = 5.9 s). Although we do not expect these differences in timing to impact our main results of interest, we will test for the possible influence of these timing differences by analyzing infants' attention (e.g., as measured by the number of habituation trials and the overall looking to the still frame events presented after each video), and explore whether (a) attention differs across conditions, and (b) differences in attention following each event relate to differences in infants' choices.

A second difference between the social and nonsocial conditions involves the location of the Climber at the start of each event. In the social condition, the Climber always has the goal to climb *up* a steep hill and needs assistance in doing so. Therefore, the Climber starts at the bottom of the hill during both helping and hindering events (before being pushed up or down). In the nonsocial condition, infants also see the ball being pushed up or down the hill. But, because the ball is inanimate and not-goal directed, the ball must begin at the bottom of the hill during pushing-up events, but at the top of the hill during pushing-down events. Although we could have created events where the character was trying to climb down a hill, we thought this goal might be perceived as relatively trivial in terms of costs (compared to climbing up a steep hill), rendering it more difficult to determine that the Climber needs assistance. We have no reason to hypothesize that these differences in starting location across conditions would influence infants' choices.

Because the critical event across conditions is the pushing action itself, it was important that the mechanics of the pushing actions were as similar as possible across conditions so the characters that interacted with the ball would appear equally agentive and goal-directed. Both conditions featured two pushing actions, each of which began with a forceful "smack." In the social condition, these smacks occur during the climbing action (i.e., on the steep part of the hill). In the nonsocial condition, on the other hand, the smacks occur on the flat portions of the hill. This difference across conditions is due to the physical properties of pushing objects: smacking an inanimate ball on the steep part of the hill should cause the ball to roll down on its own, rendering the second smack unnecessary. This difference in the location of the smacks required the characters in the nonsocial condition to stay in contact with the red circle for longer (~3 s) than the characters in the social condition (~1 s). Importantly, however, the contact time between characters *within* conditions is the same. Nevertheless, as with the differences in overall timing, we will conduct exploratory analyses to examine possible relations between infants' attention (e.g., as measured by looking-time to the freeze frames presented at the end of each video and the number of trials it took for the infants to habituate) and subsequent choices.

2 | Methods

2.1 | Participation Details

2.1.0.1 | Time-Frame. We circulated an open call for lab participation after our Registered Report received an in-principle acceptance, in April 2021. Data collection took place during a one and a half-year window, initiating in the summer of 2022 and ending in the fall of 2023[1]. Prior to receiving an in-principle

acceptance, we made an announcement on relevant listservs (i.e., *Cogdevsoc* and *Infancy*) to gauge general interest in the project, as having a sense of the number of labs that might be interested in participating helped us plan methodological decision-making (i.e., whether we would have sufficient power to propose two experimental conditions and whether labs would be using one vs. two experimenters). At the time we submitted the Stage 1 version of this Registered Report, 61 labs in 17 countries expressed interest in collecting data, pledging to test a total of 1414 infants. The obtained sample is described in the below *Participants* section. A data collection log that includes the range of testing for each participating lab can be found on OSF https://osf.io/qntyd/.

2.1.0.2 | Age Distribution. Participating labs were asked to recruit participants between 5.5 months (167 days) and 10.5 months (319 days), covering a 5-month age window.

2.1.0.3 | Power. Findings from the meta-analysis by Margoni and Surian (2018) revealed that the mean proportion of infants' preference for the Helper was 0.64, 95% CI [0.60, 0.69]. Using the Bayesian analysis tools described below, we computed the number of participants necessary to have a 0.8 probability (corresponding to 80% power) to find evidence against the null hypothesis of no preference between Helpers and Hinderers, indicated by a Bayes factor greater than 3 (Kruschke and Liddell 2018). In our simulations, this probability was achieved by including 140 participants in a study with a single condition (i.e., the social condition). To achieve the same probability in a study designed to detect a difference between an experimental group that shows an effect and a control group that performs at chance (i.e., the nonsocial control condition), our simulations revealed that 500 participants were needed.

2.1.0.4 | Lab Participation Criteria. Since the precision of our analysis is affected by the number of labs contributing data more than by the number of participants within each lab (Judd, Westfall, and Kenny 2017), and since including a large number of labs is important to the long-term objective of building a diverse community of researchers to engage in replication projects, we adopted a liberal inclusion criterion. We specified that labs should recruit typically-developing infants within the specified age range (see below for a full list of demographics collected). Sample sizes were asked to be calculated on the basis of the total number of infants who saw the entire video presentation, made a choice, and did not fit any of the exclusion criteria (described in detail below), meaning that most labs would likely need to recruit more than 16 infants. Labs were required to make sample size decisions prior to testing, and signed a contract confirming that they would not stop data collection based on results (e.g., whether or not an expected effect was observed). Thus, if labs were unable to achieve their initial goal of testing at least 16 infants, or if they tested more participants than they had initially registered for, we included their data.

2.1.0.5 | Ethics. Prior to collecting data, all labs were required to agree to a "Code of Conduct" where they agree to maintain a high level of integrity and ethics in their involvement in the project, including the explicit, detailed information about data collection integrity in the project instruction manual (e.g., not making decisions about stopping/extending data collection based on the data itself). Labs acquired their own Ethics Review Board protocols for data collection. All central data analyses used de-identified data. Individual video recordings of participants were coded and stored at each individual lab. In addition, if permitted by individual laboratories' Ethics protocols, participant videos were uploaded to a centralized video library accessible by ManyBabies investigators on the University of British Columbia's Chinook server, and by other co-investigators on Databrary (https://nyu.databrary.org/), a video data library used by behavioral scientists world-wide. Since not all laboratories had permission to share their videos, all primary data coding and exclusion criteria were determined by individual laboratories.

2.1.0.6 | Lab Research Practices. Labs were instructed to follow their individual protocols for training of research assistants and to maintain the same quality standards for this study as for other studies they conduct. Each lab completed a general questionnaire to report on their training practices, academic standing and experience of the experimenters involved in the present study (e.g., volunteer, undergraduate, graduate, postdoctoral, professor), and protocol for greeting families. Additionally, laboratories were required to create "lab tour videos" in which they walked through their lab setup and experimental procedure. These videos were uploaded to a central database.

2.1.0.7 | Lab Training and Reliability. All labs were provided with a "ManyBabies4 Big Manual" that provided a detailed overview of the experimental procedures (see https://osf.io/qntyd/). Prior to initiating data collection, labs were required to send three videos of the interactive part of the procedure (i.e., the "choice phase") to a centralized team of researchers for approval. To be approved, the pilot videos were required to follow the procedures outlined in an Experimenter's Manual located within the Big Manual linked above (e.g., timing of verbal prompts, presentation distance, and angle of choice characters). This review procedure was implemented to ensure labs strictly adhered to the experimental protocol. The decision to switch from piloting to data collection within each lab was made prior to data collection, and no pilot data was included in the final analyses.

To ensure standardization of coding for the primary variable of interest (i.e., infants' choice of puppet during the test phase), we created a centralized training process for choice coding using a centralized bank of videos with varying levels of coding difficulty. Prior to collecting data, all labs were required to complete this training with a reliability of 90%. The full protocol can be found at https://osf.io/qntyd/. During data collection, the experimenter recorded and coded the infants' choice behavior. Offline, a second researcher who was unaware of the conditions, coded infants' choices from video as a reliability check for the first experimenter's online coding. If the offline coder disagreed with the experimenter, the offline coder's choice was used for the analyses. Labs were required to report the percent agreement between the two coders in their final data reporting. The average agreement across labs was 98% (range = 80%–100%, 26/33 labs reported 100% agreement).

Labs were not required to participate in a standardized training process for coding looking-time data because some labs used automated techniques for coding (e.g., eye-tracking), and

looking-time was not the primary variable of interest in this experiment. Infants' looking time was coded either manually (82.5% of infants) or with an automated eye-tracker (17.5% of infants). Regardless, we asked all labs to conduct their own reliability coding for looking-time data within their lab on 25% of their sample and upload this data to the project's data repository (n = 35 labs reported, sample size = 1489). We calculated the intercoder reliability using the intraclass coder coefficient (ICC). The reliability was high, ICC = 0.96, $p < 0.001$, 95% CI [0.95, 0.96].

2.2 | Participants

A total of 567 typically developing, full-term infants participated. A total of 37 labs collected data, of which 34 collected data that met all inclusion criteria and were included in data analysis (mean laboratory sample size of babies included in data analysis = 16.68, SD = 10.16, range: 1–47). Labs participated from 18 regions, including: Australia (n = 25 included in analyses/44 tested), Austria (n = 16/34), Belgium (n = 17/23), Canada (n = 85/151), Colombia (n = 18/24), Germany (n = 71/134), Hong Kong (n = 29/63), Iceland (n = 15/30), Israel (n = 7/25), Italy (n = 19/48), Japan (n = 21/27), New Zealand (n = 44/73), Poland (n = 34/43), South Korea (n = 3/12), Switzerland (n = 19/28), United States (n = 80/131), mainland China (n = 47/93), and the Netherlands (n = 17/35). The average age of infants included in the study was 253 days (SD = 42.95, range: 167–319), and 46.91% were female. An additional 451 infants were tested but were not included in the final sample due to not having met the inclusion or eligibility criteria (see below). This exclusion rate (44.30%) is much higher than the original Hamlin et al. work (13%; 4/32), but more comparable to past replication attempts (e.g., 37.25% in Schlingloff, Csibra, and Tatone 2020). Further, the majority of exclusions resulted from experimenter errors and equipment errors, which highlights the complexity and difficulty of standardizing and implementing methodology with behavioral components at a large scale. Information on all participating and included labs, including information on all babies tested, is provided in Table 1.

2.3 | Procedure

The approved Stage 1 protocol can be found on the Open Science Framework (https://osf.io/qntyd/). Procedural instructions provided to each participating lab are viewable on OSF.io (https://osf.io/qntyd/). Infants sat in front of the screen, either in an infant seat or on their parent's lap. Infants were randomly assigned to one of two conditions: a social condition (N = 302) or a nonsocial control condition (N = 265; described below). The conditions differed only in the content of the video stimuli—the overall procedure was otherwise identical across conditions. While the video stimuli were displayed, parents were asked to close their eyes or wear occluding glasses. Parents also received standardized instructions on how to maintain proper positioning throughout the experiment so that infants could fully view the display and to avoid inadvertently biasing their infant: they were told to sit still (e.g., act like infants' "chairs," try to keep them in their lap, and hold them around the rib cage), not talk or gesture during the experiment, and not redirect infants' attention to the events if the

TABLE 1 | Details about participating labs.

University	Region	*N*
Peking University	Mainland China	93
The University of Hong Kong	Hong Kong	63
University of British Columbia	Canada	59
University of Auckland	New Zealand	50
University of Wroclaw	Poland	43
University of California, Santa Barbara	United States	42
University of Amsterdam	The Netherlands	35
Central European University (Vienna)	Austria	34
University of Toronto Scarborough	Canada	34
University of Manitoba	Canada	33
University of Akureyri	Iceland	30
Università degli Studi di Milano-Bicocca	Italy	29
Ludwig-Maximilians-Universität München: babylabLMU	Germany	28
University of Minnesota	United States	28
University of Zurich	Switzerland	28
Osaka University	Japan	27
Ludwig-Maximilians-Universität München: lmuMunich	Germany	25
Western Sydney University	Australia	25
Ruhr University Bochum	Germany	24
Universidad del Valle	Colombia	24
University of Göttingen	Germany	23
Université Libre de Bruxelles	Belgium	23
Victoria University of Wellington	New Zealand	23
Max Planck Institute for Human Cognitive and Brain Sciences	Germany	22
Concordia University	Canada	19
University of Newcastle	Australia	19
University of Trento	Italy	19
Princeton University	United States	18
University of Virginia	United States	16
The Academic College of Tel Aviv-Yaffo	Israel	14
University of California, San Diego	United States	14
Max Planck Institute for Human Development	Germany	12
Yonsei University	South Korea	12
University of Haifa	Israel	11
Arizona State University	United States	10
St. Francis Xavier University	Canada	6
University of the Incarnate Word	United States	3

Note: The *N* represents the number of total tested participants, including those who needed to be excluded from the analyses.

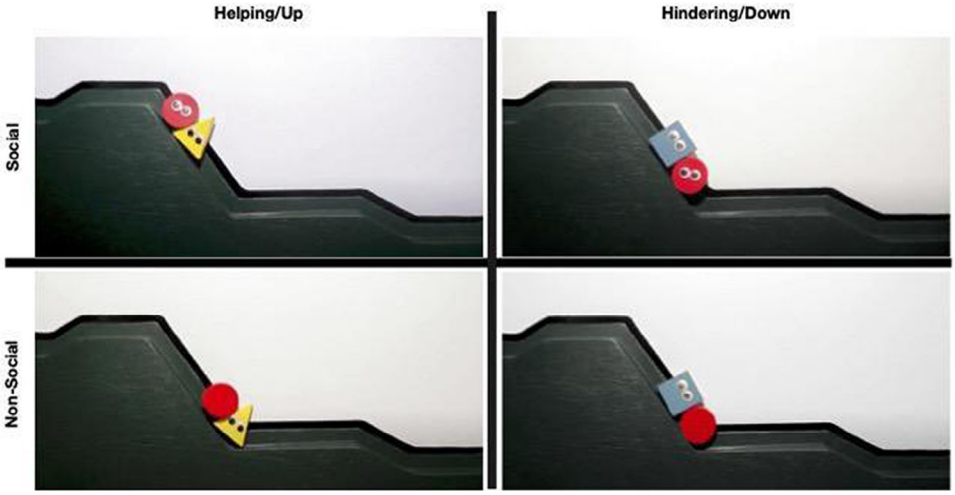

FIGURE 1 | Screenshots of study stimuli. *Note:* Screenshots of helping (top left) and hindering (top right) events in the social condition; pushing up actions (bottom left) and pushing down actions (bottom right) in the nonsocial condition.

infants grew uninterested. Infants were video recorded during the entire experiment.

2.3.0.1 | Apparatus. Laboratories were instructed to display the video stimuli using the setup with which they were most familiar (e.g., TV screen, projection screen, computer monitor). The mean screen size was 91.94 cm diagonal (*SD*: 57.48, range: 53.34–294.64) and screens stood at a mean distance of 94.67 cm (*SD*: 45.70, range: 23.60–196.00) from infants. Laboratories were instructed to position infants' faces at the center of the screen, so they were not required to look up or down to view the stimuli. To facilitate standardization across labs and experimenter blinding, laboratories were encouraged to use PyHab (Kominsky 2019) to display the video stimuli. PyHab is a free software program based in a PsychoPy environment (Kanbe 2019; Peirce et al. 2019) that randomly selects a counterbalanced order, presents stimuli, and records looking time, all while allowing the experimenter to remain blind to the onscreen stimuli display. Through this procedure, the experimenter is able to watch the infant's face, initiate a trial by pressing a key, and register when the infant is looking at the screen by pressing and releasing a key. Pyhab expects the experimenter and the infant to be viewing separate monitors, and rather than viewing what the infant is watching, the experimenter sees a display that simply indicates whether a trial is active or not. In this setup, it is possible for the entire experiment to be run by a single experimenter who is unaware of the puppets' identities (i.e., which puppet is the Helper vs. Hinderer). Labs that did not use PyHab were instructed to reproduce these testing conditions as closely as possible, and were strongly encouraged to use a second experimenter whenever possible in order to ensure the experimenter remained naive to the puppets' identities. Regardless of setup, we required that the experimenter administering the choice phase be naive to puppet identity. Thus, if labs did not use PyHab or a second experimenter, they were required to report how they ensured the experimenter remained naive to the puppets' identities.

2.3.0.2 | Personnel. Labs were given the option of running the experiment with one or two experimenters, as long as the experimenter administering the choice phase was naive to the puppets' helping/hindering identities. Labs were required to report to the central planning team any minimum academic achievements required to participate in data collection (e.g., a completed undergraduate degree), as well as a summary of the training procedures that the experimenters underwent prior to data collection.

2.3.0.3 | Calibration. Prior to displaying the habituation trials, an attention-getting stimulus (a spinning multi-colored circle) appeared in turn on each of the four corners of the screen and once in the center, allowing the online coder to calibrate infants' looking at the screen. The attention-getter moved to a new location after the experimenter pressed a key, indicating that the infant looked at it.

2.3.0.4 | Onset of Each Trial. Each trial began with an attention-getting stimulus at the center of the screen (e.g., a twirling/squeaking multi-colored circle). When the experimenter determined that the infant was fixating on the screen, the experiment proceeded to the next trial.

2.3.0.5 | Habituation Phase. The stimuli (Figure 1) used for habituation were modeled off those in Hamlin and colleagues (Hamlin, Wynn, and Bloom 2007), and are viewable at the following link in the folder "Video Files March 2021" (https://osf.io/qntyd/). The habituation phase of the experiment was the only aspect of the study that differed between the social and nonsocial conditions.

2.3.0.6 | Habituation Phase: Social Condition. Stimuli consisted of four 13.2-s, prerecorded, live-action puppet shows displayed via video depicting either a "helping event" or a "hindering event." Each event was filmed against a white background and featured a green hill rising from the bottom right to the top left of the scene. The hill had two inclines, the second steeper than the first, with a small flat plateau between them. Characters consisted of three colored wooden shapes (i.e., a red circle, a blue square, and a yellow triangle), each featuring googly eyes. The red circle was always the "Climber" character who tried but failed to climb the hill; the Climber's eyes were fixed, pointing upward

so that it continuously gazed toward her goal. The blue square and yellow triangle were the "Helper" and "Hinderer" characters; their eyes were not fixed, and the characters' identities were counterbalanced across participants. Soft instrumental music played in the background of the events to maintain infants' interest.

For each event, the Climber was observed trying but failing to reach the top of the hill. The Climber was first shown resting at the bottom of the hill, at the right edge of the display. The Climber then moved up the first mild incline to a plateau, where it "danced" briefly. Next, the Climber attempted but failed twice to reach the top of the second, steeper incline, moving slightly higher but sliding back down to the plateau after each attempt. During the Climber's third attempt, a second agent (the Helper or the Hinderer) entered the scene. In order to draw infants' attention to the screen just before the critical "hitting" aspects of the events occurred, a "ding" sound was played 300 ms prior to the second agent's appearance on stage. During helping events, the Helper entered the scene from the bottom of the hill, moved up the hill, and eventually contacted the Climber. The Helper hit the Climber twice from below, pushing it upward to the top of the hill. During hindering events, the Hinderer entered the scene from the top of the hill and moved down, eventually contacting the Climber. The Hinderer hit the Climber twice from above, pushing it downward to the bottom of the hill. Once the Climber reached its final destination (the top or bottom of the hill), the Helper or Hinderer exited the stage from where they entered, and the event paused. This last frame remained on the screen until either the infant looked away for 2 consecutive seconds or 30 s elapsed (Hamlin 2015). The helping and hindering events were matched in timing, speed, and sound down to the millisecond (see above OSF link for a comparison of videos). For example, the timing of the initial attempts of the Climber, the first frame where the Helper/Hinderer enters the stage, the first contact between the Helper/Hinderer and the Climber, and the exit speed of the Helper/Hinderer are identical across videos. All variations of the videos used the same audio track.

To ensure that infants viewed the helping/hindering events in their entirety, we identified a Critical Period from the first frame the Helper or Hinderer appeared on stage to the last frame where the Helper or Hinderer was moving (2500 ms). If infants looked away during the Critical Period for more than 1000 cumulative ms (i.e., 40% of the window), the trial was terminated and repeated up to three times[2]. Failure to attend the events within this Critical Period after three showings resulted in the exclusion of the infant due to inattentiveness. If such occurred within the first six trials, the experiment was nevertheless continued until the 6th trial and ended immediately after. This was done to avoid any frustration for caregivers that could arise from ending the experiment abruptly, and to ensure that each family had the opportunity to fully participate in the experiment.

Helping and hindering trials were shown in alternation until infants fulfilled the preset habituation criteria, adopted from Hamlin and colleagues (Hamlin, Wynn, and Bloom 2007). To fulfill these criteria, infants' summed looking times to any three consecutive trials had to decrease to less than half the summed looking time to the first three trials for which looking totaled 12 s or more. Once this criterion was met, or after a maximum of 14 trials, infants moved on to the choice phase. The presentation order of the videos (helping first vs. hindering first) was counterbalanced across participants.

2.3.0.7 | Habituation Phase: Nonsocial Control Condition. Infants viewed events highly similar to the helping and hindering events of the social condition but with one critical difference: the Climber was replaced by an inanimate, eyeless object that did not exhibit self-propelled motion. Because there were no initial climbing actions, videos in the nonsocial condition were 4.4 s shorter in duration than those in the social condition. In these videos, infants viewed an inert red ball get pushed up or down the hill by the same animated triangle and square characters from the social condition. Aside from these differences, the videos across conditions were closely matched on visual and sound cues. The screen time of the square/triangle characters was closely matched across conditions (i.e., screen time of Helper/Hinderer: 4.7 s per video; screen time of up-pusher/down-pusher: 5.9 s per video). The critical window was defined using the same boundaries as the social condition: it started from the first frame when the Helper or Hinderer appeared on stage to the last frame when the Helper or Hinderer stopped moving (3.5 s). As in the social condition, if infants looked away for more than 40% cumulative ms during this critical window (i.e., 1400 ms), the trial was terminated and repeated up to three times. The habituation criteria were identical across conditions. Within the nonsocial condition, pushing up versus down videos were closely matched on speed, timing, and sound cues. The presentation order of the videos (pushing up vs. down first) was counterbalanced across participants.

2.3.0.8 | Choice Phase. Complete instructions for administering the choice phase are viewable in Appendix A of the Supplemental Materials. The choice phase was identical across conditions.

Immediately following the end of the habituation phase, an experimenter (naive to the identity of the characters) presented infants with foam versions of the yellow triangle and blue square characters, attached with Velcro to a 45 cm × 60 cm board with a white background (Figure 2). The characters were standardized in size: blue square, 9.9 cm × 9.9 cm, and yellow triangle, 14.8 cm (base) × 13 cm (height). Labs were provided with hyperlinks to the material used and templates for creating the characters (see Appendix B for excerpts from the links). A subset of characters was centrally created by the project team and distributed to some participating labs via mail or at the International Congress for Infancy Studies conference in July 2018. Seventeen labs received characters made by the project team. Each character was located 7.5 cm from the base to the bottom of the board. Velcro pieces were placed at the center of each character. Measuring from the center of the Velcro, the pieces were 30 cm apart from each other, 15 cm from each side of the board, with each character placed ~9 cm from its outermost point to the side edges of the board. The location of characters (left/right) was counterbalanced across participants.

At the start of the choice phase, parents were instructed to grasp their infants securely around the waist and position them close to their knees to facilitate infants' reaching. Parents were then asked

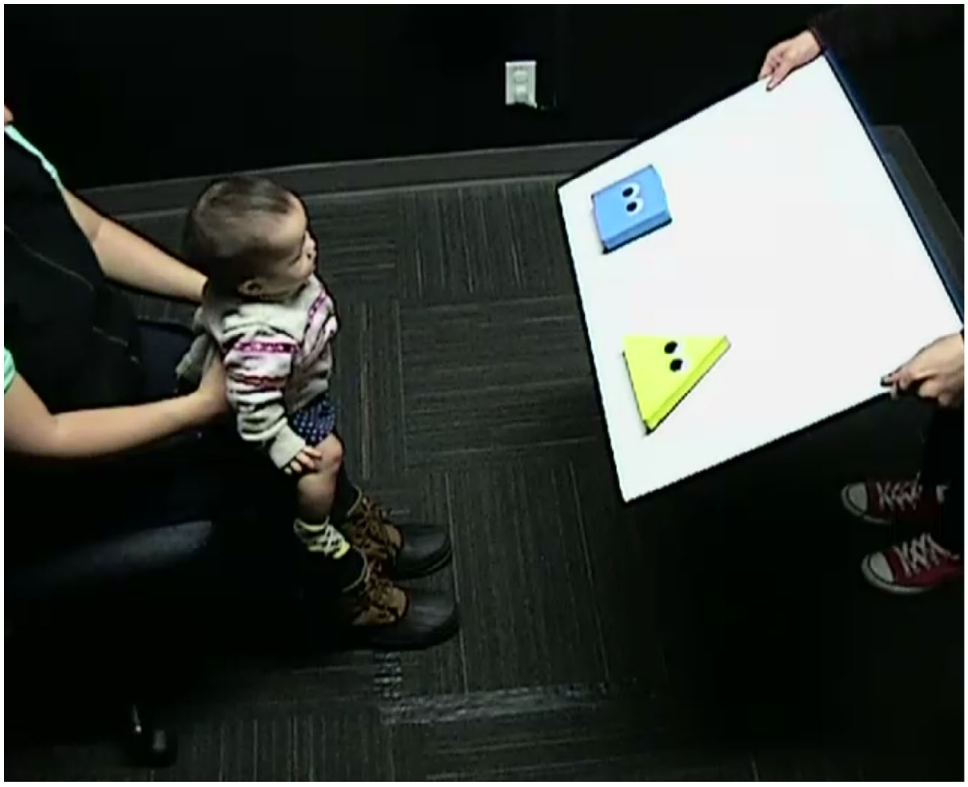

FIGURE 2 | Sample image of choice phase. *Note:* Choice phase foam 3-D stimuli presented on a white board. Board dimensions: 45 cm × 60 cm; Shape dimensions: 9.9 cm × 9.9 cm (blue square), 14.8 cm (base) × 13 cm (height; yellow triangle).

to close their eyes or wear occluding glasses to prevent them from biasing the child's attention toward a particular character.

Next, the experimenter leaned over in front of the infant and said, "Hi! Look!" while lowering the board directly at an approximately 30-degree angle. The board was placed just out of the infant's reach until the infant looked at both characters. When the infant had done so, the experimenter said, "Hi!" to direct the infant's attention away from the characters and back to the experimenter. Upon making eye contact with the infant, the experimenter said, "Who do you like?" while moving the board within the infant's reach. The experimenter kept the board extended toward the child for 60 s, while keeping track of time either in their head, using a stopwatch, or by referencing a wall clock. Any choices that were made after 60 s were excluded from the analyses (e.g., in the case the experimenter inadvertently extended the choice phase past 60 s).

Infants' choice of character was coded online by the experimenter conducting the choice phase. A *visually guided reach* (touching one character while looking at it) was indicative of infants' choices. Occasionally, infants touched both puppets at once; some of these instances were counted as valid choices according to our predefined criteria. In *usable both* touches, the infant clearly directed her gaze and reached toward one character but touched the other character as well. On the other hand, *unusable both* touches involved the infant touching both characters with unclear or inconsistent visual attention. Experimenters were instructed to continue the choice procedure until a usable touch was recorded by encouraging and reprompting the baby after 30 s had passed. If infants did not make a choice during the 60 s choice phase, their data were excluded from final analyses. Experimenters were required to administer the choice phase within 2 min of ending the habituation videos. If more than 2 min passed, the session was considered an "experimenter error" and excluded from data analysis.

2.3.0.9 | Order of Testing. Laboratories occasionally tested participants in a separate (unrelated) experiment during their visit. We encouraged, but did not require, labs to run the Helper/Hinderer experiment as their first experiment. All labs recorded whether another experiment was run with the same participants and whether it preceded or followed the Helper/Hinderer experiment.

2.3.0.10 | Demographics. Each lab collected a set of participant demographic information for each infant: gender, date of birth, estimated proportion of language exposure for language(s) heard daily, preterm/full-term status (i.e., more than 36 weeks gestation), hearing or visual impairments, developmental concerns (e.g., developmental disorders), infant handedness (right, left, not sure), parent handedness (right, left), and color blindness in the immediate family. Labs were given a standard participant questionnaire that they were encouraged to use (see Appendix C of the Supplemental Materials).

2.4 | Exclusion Criteria

All data collected for the study (i.e., every infant for whom a data file was generated, regardless of how many trials were completed) were uploaded to the central database for analysis. We instructed

labs to make note of any instances in which a procedural error or anomaly occurred during testing. Participants were excluded from the final analysis if they met any of the criteria below.

2.4.1 | Eligibility Criteria

1. *Full-Term*. Full term was defined as 36 weeks or more gestation. Caregivers were asked to report their child's due date, and a centralized research team calculated the child's gestational age. Ten (1.0%) of tested infants were excluded for not meeting this criterion.
2. *No Diagnosed Developmental Disorders*. If parents reported any known developmental disorders, their infants were excluded. One (0.1%) of the tested infants were excluded for not meeting this criterion.
3. *Within Age Range*. Infants were excluded if they were older or younger than our target age range. Fifty-four (5.3%) of tested infants were excluded for not meeting this criterion. Note that this category was added after the Stage 1 submission when we conducted an initial check for the data uploaded by labs.

2.4.2 | Experimental Exclusion Criteria

4. *Failure to Set a Habituation Criterion*. Infants set the habituation criterion if their looking time toward the paused frame at the end of the habituation video on any 3 consecutive trials during the first 6 trials summed to at least 12 s. If an infant did not set a habituation criterion in the first 6 trials, the study terminated, and their data was excluded. A total of 142 (14.0%) of tested infants were excluded for not setting a habituation criterion.
5. *Failure to View the Critical Period*. If infants looked away for more than 750 ms cumulatively during the Critical Period of a trial, the experimenter repeated the trial up to two times. If infants failed to attend a Critical Period after three consecutive trials, they were excluded. A total of 126 (12.4%) of tested infants were excluded for this criterion.
6. *Unclear Choice*. Infants must have produced visually guided reaches (as defined in the Supplementary Materials, Appendix A) toward one of the two characters in the choice phase. A total of 151 (14.8%) of the tested infants were excluded for not meeting this criterion.
7. *Parental/Outside Interference*. Infants were excluded if distracting events that were not part of the study protocol occurred during the Habituation or choice phase (e.g., a noise outside the testing room, the parent gesturing to the child). Thirty-nine (3.8%) of tested infants were excluded for parental interference.
8. *Experimenter Error*. Experimenter errors included any actions the experimenter inadvertently made that may have influenced infants' behavior (e.g., the experimenter failed to record an infant's looking time, failed to repeat the Critical Period after infants missed viewing it, or became aware of puppets' identity prior to the choice phase). A total of 133 (13.1%) of tested infants were excluded for this criterion.
9. *Equipment Error*. Any technical deviations that may have influenced infants' behavior were considered equipment errors (e.g., stimulus froze during presentation, software did not accurately set habituation criteria). A total of 232 (22.8%) of tested infants were excluded for equipment error.
10. *Fussiness*. Infants became too fussy to finish the experiment (e.g., by showing visible signs of distress, fussiness, or crying). Twenty-nine (2.9%) of tested infants were excluded for fussiness. Note that this category was added after the Stage 1 submission, but prior to the onset of data collection. The training manual defined fussiness as follows: "Infant becomes visibly distressed to the point where the experiment cannot continue. Note that this decision must be made by the experimenter and not the parent."
11. *Data Entry Error*. Labs entered the data incorrectly while reporting the experiment condition or infants' choices. Thirteen (1.3%) of tested infants were excluded for this criterion. Note that this category was added after the Stage 1 submission when we conducted a data entry check for the data uploaded by labs.

Exclusion based on the reasons above led to 451 total excluded and ineligible participants (44.30%). The categories above are *not* mutually exclusive (e.g., parental influence may have co-occurred with experimenter error), so the numbers in the individual categories listed above will total to more than 451. If we do not consider ineligible infants (i.e., infants who were tested but should not have been due to their age, preterm status, or diagnosed developmental disorder), the exclusion rate is 40.63%. Labs reported their exclusion rates and reasons, the central project team also manually checked the data for any missed exclusions and removed them from the final data analysis. Of the eligible infants tested by labs, the average (unweighted) exclusion rate across labs was 43.70% ($SD = 23.03\%$), with a range between 10% and 100%.

3 | Results

Note: Pilot data are reported in Supplementary Materials only (see Appendix D for details on pilot methods and Appendix E for pilot ananalyses). Pilot data only include infants in the social condition, as the decision to run a nonsocial control condition was made during the peer review process, after initial piloting was complete. We made a few minor modifications to the social condition videos used in the final data collection so they could be more closely matched to the nonsocial videos and more closely match the original Hamlin, Wynn, and Bloom (2007) *videos. Data simulations with the expected effect size were used to confirm the validity of our analysis structure to test the effects of interest. All analysis scripts, data, including simulations for Bayesian power analyses and choices for the statistical plan, can be found online (*https://github.com/manybabies/mb4-analysis*) and (*https://osf.io/qntyd/*). Bayesian power analysis was conducted in a similar fashion to frequentist power analysis, with the difference that power was estimated as the number of samples in which Bayes factors (BFs) were greater than the threshold value 3.*

3.1 | Analysis Structure

We used a Bayesian analysis framework to calculate *BF*s through Bayesian model comparisons. This framework provides

a principled method for testing evidence in favor of and against the null hypothesis (i.e., no difference in preference for Helpers over Hinderers), and allows us to fit hierarchical models to account for sparse data and variation across labs.

We fit generalized linear mixed-effects Bayesian models using Stan (Gelman, Lee, and Guo 2015) and the brms package (version 2.20.4; Bürkner 2017); 95% credible intervals (CI) were computed using the HPDInterval function from the coda package (version 0.19.4; Plummer et al. 2006). While standard frequentist 95% Confidence Intervals are defined such that if repeated samples are drawn from a population, the Confidence Interval will include the true population mean in 95% of the observed samples, Bayesian credible intervals represent the interval in which the true population mean is likely to be, given the data at hand and the model assumptions (Morey et al. 2016).

We were also interested in assessing the evidence for specific hypotheses (e.g., a nonzero preference for Helpers over Hinderers). While Bayesian CIs allow us to assess the precision of our measurement of effects, they do not allow for the assessment of evidence in this way. For this purpose, *BF*s are the appropriate Bayesian tool. Crucially, while p values only allow us to reject the null hypothesis H_0, *BF*s obtained through model comparison allow us to either reject H_0 in favor of H_1, accept H_0 to the detriment of H_1, or conclude that the data do not provide sufficient evidence to support either. To obtain *BF*s, we computed two models for each specific research question, one representing the null hypothesis (H_0) and the other an alternative hypothesis (H_1). We then estimated the posterior distributions of each model and compared them to obtain a *BF*, using the bridgesampling package (version 1.1.2; Gronau, Singmann, and Wagenmakers 2017).

3.2 | Confirmatory Analyses

Our primary research questions were (1) whether there was an above-chance choice of the Helper in the social condition compared to the choice of the Push-Up Character in the nonsocial control condition, and (2) whether there were developmental differences in choice. Based on the original study (Hamlin, Wynn, and Bloom 2007), we predicted that the proportion of children choosing the Helper would be above chance.

In our Stage 1 Registered Report, we proposed two analysis plans with power analyses computed for each plan based on the final sample size (reported below). In the event that we did not have enough participants to reach 80% power to compare the social to the nonsocial condition ($n_{\min} = 500$), we planned to revert to testing the social condition only against chance ($n_{\min} = 140$). Below, we describe the analysis plan with a nonsocial control condition. If our sample size had only been sufficient for comparing the choices of infants in the social condition against chance performance, the same procedure would have been used, removing the main effect of the condition from the model, and instead testing the intercept. Our final sample size was 567, which is sufficient for achieving 80% power to compare the social to the nonsocial condition, so we used the below analysis and included the main effect of the condition and its interaction with age in the model.

We tested our prediction via a Bayesian generalized linear mixed effects model with a Bernoulli response model. In such models, the probability for the dependent variable (here, participants' choice) is transformed through the logit function such that an estimate of zero corresponds to a 50% chance for either choice, with values greater than zero representing a preference for the Helper. The specification of our model was:

$$\text{choice} \sim 1 + \text{condition} + \text{age} + \text{condition} : \text{age} + (1 + \text{condition} + \text{age} + \text{condition} : \text{age} \mid \text{lab})$$

Choice (Helper vs. Hinderer, or Push-Up vs. Push-Down character) and condition (social vs. nonsocial) were entered as binary variables. Age (in days) was scaled and centered to allow for better convergence of the statistical models and easier interpretation of the results. This type of variable scaling is standard (e.g., Marquardt 1980), and we adopt it because (a) keeping raw age in days would result in a meaningless intercept and main effect of condition (representing a hypothetical preference for either group at age = 0), and (b) estimating an intercept and a difference between conditions at age = 0 as well as a slope for age would lead to less precise estimates (as we found in simulations). To control for possible variation in preference across labs as well as in developmental trends across labs, the model also included random intercepts for each lab and random slopes for condition, age, and their interaction, by lab. For the effect of condition, we used treatment contrasts with the non-social condition as the reference level.

In Bayesian analyses, the most appropriate method for pooling data from a new experiment with previous knowledge is to set priors based on that previous knowledge. For our key effect of interest, namely the effect of condition (a higher or lower proportion of Helper choices in the social vs. nonsocial condition), we used an informative normal prior based on the effect size and Confidence Interval from the meta-analysis of Margoni and Surian (2018). We used the following formula to compute the standard deviation from the reported Confidence Interval ($\sigma = CI_{one-side\ width} \times \sqrt{N_{mean}}/z(95)$; here, $M = \text{logit}(0.64) = 0.58$, $SD = 0.1$). We believe that this choice was warranted given that the meta-analysis included both studies that showed and did not show the hypothesized effect, as well as published and unpublished data, making them unlikely to reflect the outcome of a biased literature selection toward the effect of interest. Further, the estimate of 0.64 represents the proportion of infants choosing prosocial characters after having corrected for possible publication bias. Sensitivity analyses using noninformative priors are presented alongside our results (hereafter referred to as the noninformative model, as compared to our main model, the informative model). For these analyses, we only specify a prior on the random effects to improve model convergence.

We further used weakly informative normal priors for the intercept ($M = 0$, $SD = 0.1$), relationship of age to choice and its interaction by condition ($M = 0$, $SD = 0.5$), and a restricted Student prior on random effects (with $df = 3$, $M = 0$, and $SD = 2$, as compared to the default in the brms package). Priors of this type provide very little information in cases like ours where we have large amounts of data; their primary purpose is to improve model convergence and subsequent bridge sampling for the computation of *BF*s.

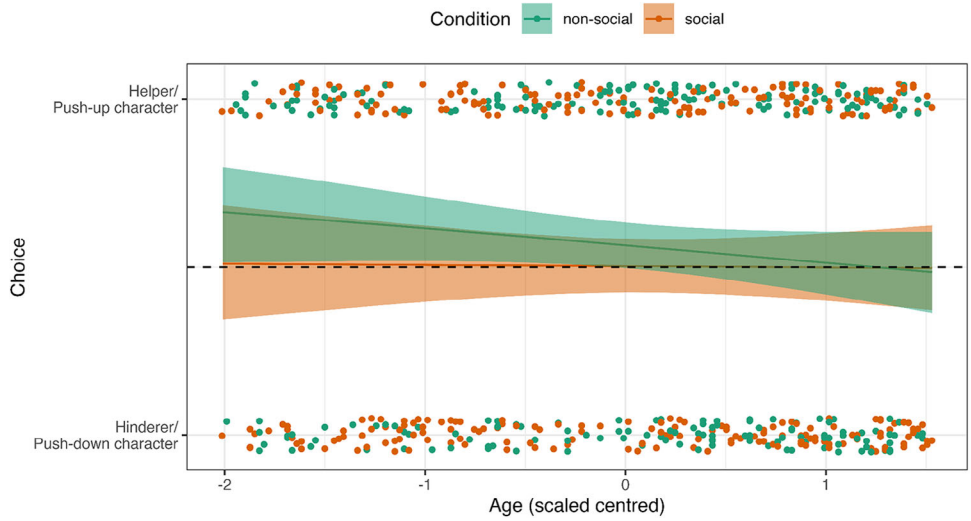

FIGURE 3 | Age × Choice plot. *Note:* Probability of choosing the Helper/Push-up character (over Hinderer/Push-down character) across ages. The smoothing line shows the predicted marginal effects from our noninformative Bayesian regression model along with their credible interval. The dashed line on the *y*-axis represents chance performance. Data are jittered slightly on the vertical axis to avoid overplotting.

Relative to the first research question, the main finding we aimed to replicate was infants' preference for the Helper at levels greater than chance, as reflected by a greater than zero estimate for the effect of the condition. To assess the evidence for this hypothesis, we computed the *BF* in favor of the full model described above (H_1) compared to a model that did not include a main effect of the condition or any higher-level terms including this effect (H_0). Given that random effects are only interpretable with respect to the corresponding fixed effects for nested model comparisons (Stroup 2012), random effects corresponding to the dropped fixed effects were also dropped[3]. More precisely, the model for H_0 was specified as:

$$\text{choice} \sim 1 + \text{age} + (1 + \text{age} \mid \text{lab})$$

Following Kass and Raftery (1995), we interpreted a $BF > 3$ in favor of either H_0 or H_1 as substantial evidence and a $BF > 10$ as strong evidence[4].

For all preregistered main models, models were checked for divergent transitions, $\hat{R}$, and effective sample sizes; all models converged properly, indicated by no divergent transitions, an effective sample size greater than 10% of the total sample size, and $\hat{R} < 1.1$. Since dispersion issues can arise with these models, they were also checked for dispersions and posterior predictions visualized against the data.

We found a $BF_{01} > 100$ in the informative model and a $BF_{01} = 10.23$ in the noninformative model, providing further evidence that infants' choices were not influenced by condition or the condition-by-age interaction. In other words, infants did not preferentially choose the Helper over the Hinderer more often in the social condition than in the nonsocial condition, which did not replicate previous findings (e.g., Hamlin 2015, and Bloom 2007). In the social condition, 49.34% of infants chose the Helper. In the non-social condition, 55.85% of infants chose the Helper (See Figures 3 and 4). Parameter estimates for this model with and without informative priors are reported in Table 2 and Table 3, along with estimated error and 95% credible intervals (CI) for those parameters. It is important to note that the 95% CI for the condition parameter did not overlap with zero in the informative model ($b = 0.42$, $SE = 0.09$, 95% CI [0.24, 0.61]). However, given a $BF_{01} > 100$ for the informative model, this parameter estimate alone cannot support the effect of condition on infants' choices (i.e., Type II conflict, Lovric 2019). It is likely that this effect was driven by our strong priors with a relatively small standard deviation of 0.1.

Relative to the second research question concerning developmental differences in infants' choice, we were interested in whether the addition of the condition-by-age interaction to the model contributed to model fit. We assessed evidence for this question via the same procedure as above, fitting a null model without the condition-by-age interaction:

$$\text{choice} \sim 1 + \text{condition} + \text{age} + (1 + \text{condition} + \text{age} \mid \text{lab})$$

The *BF*s for this comparison was $BF_{01} = 8.40$ in the informative model and $BF_{01} = 7.35$ in the noninformative model, suggesting that the condition-by-age interaction did not have an effect on infants' choices (informative: $b = 0.11$, $SE = 0.20$, 95% CI [−0.28, 0.49]; noninformative: $b = 0.20$, $SE = 0.20$, 95% CI [−0.20, 0.59]). In other words, infants' choices were not influenced by condition or age.

As a follow-up analysis, we refitted the same model, including only infants who successfully habituated to the events (n = 441, 77.78% of babies successfully habituated). We again fitted a Bayesian Bernoulli linear mixed effects model. Choice (Helper vs. Hinderer) and condition (social vs. nonsocial) were entered as binary variables. Age (in days) was scaled and centered. To control for possible variation in preference across labs as well as in developmental trends across labs, the model also included random intercepts for each lab and random slopes for condition and age by lab. Priors were as above. We used the same model comparisons and the same method to obtain *BF*s. We found a $BF_{01} > 100$ in the informative model and a $BF_{10} = 1.45$ in the noninformative model, suggesting that condition and the

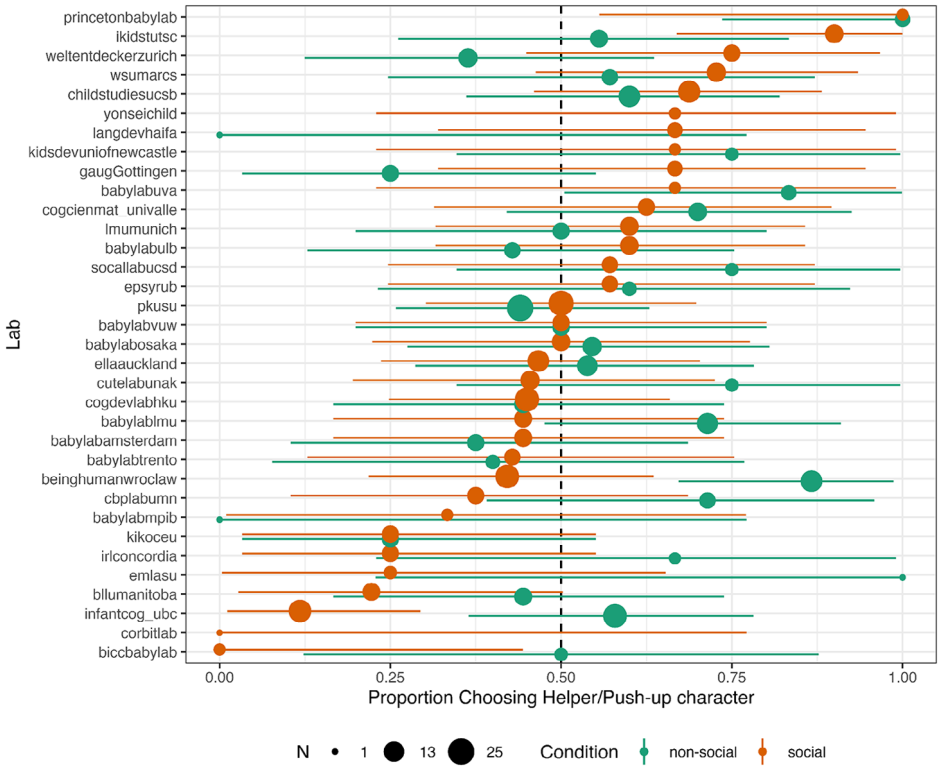

FIGURE 4 | Forest plot of character choice for all contributing labs. *Note:* "Forest plot" of estimates for proportion of participants selecting the Helper/Push-up character (over Hinderer/Push-down character) for each contributing lab. The dashed line at 0.50 on the *x*-axis represents chance performance. Error bars represent 95% Bayesian credible intervals.

TABLE 2 | Bayesian model parameter estimates—Informative priors.

Parameter	Estimate	Est. error	95% CI
Intercept	−0.11	0.1	[−0.3, 0.08]
Condition	0.42	0.1	[0.23, 0.61]
z_age_days	−0.12	0.16	[−0.44, 0.19]
condition:z_age_days	0.17	0.22	[−0.28, 0.60]

Note: Parameter estimates along with estimated error and 95% credible intervals (CI) for those parameters for the full Bayesian model with informative priors.

TABLE 3 | Parameter estimates for Bayesian models—Noninformative priors.

Parameter	Estimate	Est. Error	95% CI
Intercept	0.45	0.17	[0.12, 0.77]
Condition	−0.49	0.23	[−0.95, −0.02]
z_age_days	−0.23	0.18	[−0.58, 0.13]
condition:z_age_days	0.29	0.24	[−0.19, 0.74]

Note: Parameter estimates along with estimated error and 95% credible intervals (CI) for those parameters for the full Bayesian model with noninformative priors.

condition-by-age interaction did not have an effect on infants' choices. In other words, habituated infants did not preferentially choose the Helper over the Hinderer more often in the social condition than in the nonsocial condition, failing to replicate previous findings. In the social condition, 47.88% of habituated infants chose the Helper. In the non-social condition, 59.51% of habituated infants chose the Helper.

Additionally, we were interested in variation in infants' choice behavior across labs as, descriptively, effect sizes from individual labs greatly varied (see Figure 4). Following Klein and colleagues (Klein et al. 2014), we calculated the binary intraclass correlation coefficient (ICC) using the ICCbin package in R (version 1.1.1; Chakraborty and Hossain 2018). This measure captures the degree to which the proportion of Helper choices was correlated across labs by comparing participant-level deviations from the within-lab mean to the lab-level deviations from the between-lab mean. There was a low intra-class correlation of effects across labs (ANOVA Estimate ICC = 0.02, Smith's Large Sample Confidence Interval [0, 0.06], excluding one lab with only one data point), suggesting that the within-lab variance is much greater than the across-lab variance. Thus, it is unclear whether the observed variations in effect size reflect genuine cross-lab differences, or are a product of random variability inherent to complex behavioral paradigms.

Finally, we completed all confirmatory analyses using a frequentist approach to ensure their consistency with the Bayesian approach. Overall, the findings aligned with those reported above. Because the frequentist analyses were not included in our Stage 1 Registered Report, outputs from the frequentist analyses (including a summary table of the parameter estimates), additional information regarding how the frequentist analyses were carried out, and any discrepancies from the Bayesian analyses[5], can be

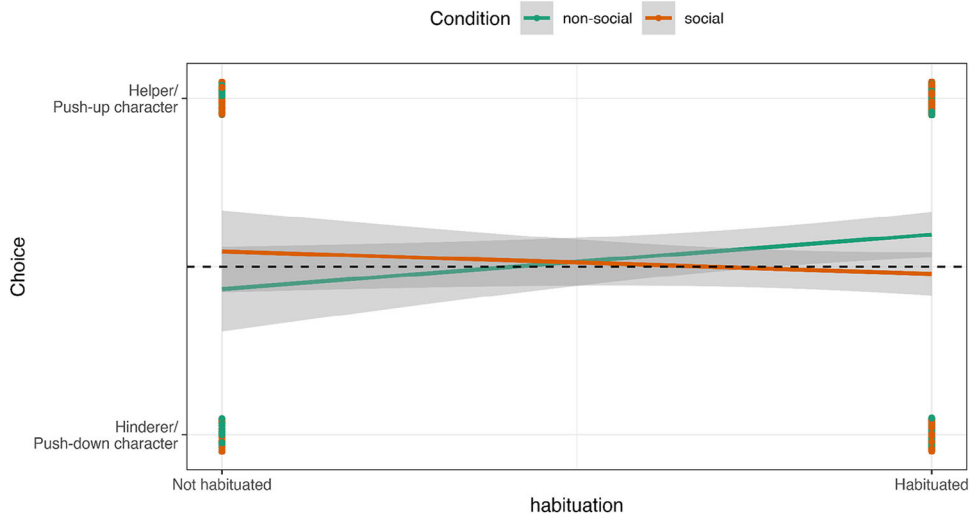

FIGURE 5 | Plot of habituation status × Probability of choosing Helper/Push-up character. *Note:* Probability of choosing the Helper/Push-up character (over Hinderer/Push-down character) based on infants' habituation status (i.e., whether they successfully habituated or not). The dashed line on the *y*-axis represents chance performance. The smoothing line shows the estimated effects with credible intervals. Data are jittered slightly on the vertical axis to avoid overplotting.

found in an annotated code script, available on GitHub (https://github.com/manybabies/mb4-analysis).

3.3 | Exploratory Analyses on Infants' Choices

An additional set of exploratory analyses was conducted. In our Stage 1 Registered Report we included the following exploratory analyses to test potential moderators of infants' preference for Helpers and Hinderers: (1) attention to the video events (i.e., as measured by the number of trials to habituation, mean looking time following the pushing events), (2) clear versus ambiguous choice actions (i.e., whether infants touched both characters during the choice phase), and (3) whether the experimenter had knowledge of the participant's condition (i.e., whether the experimenter administering the choice phase knew whether the infant participated in the social vs. nonsocial condition).

We also conducted a set of exploratory analyses that were not included in our Registered Report, using the following moderators: (1) *experiment-level factors*, including: the order of helping vs. hindering videos, Helper identity (yellow vs. blue shape), Helper side during the choice phase (right vs. left), infants' visual angle (calculated from infants' distance from the screen and screen size), whether the experiment was conducted as the first session of the infant's visit or not, whether the experimenter wore a mask or not, (2) *child-level characteristics*, including: habituation status (i.e., whether infants reached the habituation criterion prior to the choice phase), handedness, color blindness, participant sex, the percentage of primary language exposure, and (3) *lab-level factors*, including: the lab's overall exclusion rate (not including participants who were tested but were ineligible, e.g., due to age), the exclusion rate due to failure to make a choice for each lab, and the median dates of data collection of each lab (transformed into the number of days between the median date and the earliest median date). Most factors were selected based on theoretical relevance and interest. Others, especially experiment and lab-level factors, were selected to shed light on potential sources of cross-lab variation. The median testing date was analyzed to explore potential cohort effects as a result of the COVID-19 pandemic[6].

For each moderator, we fitted a noninformative maximal model by including both the main effect of the moderator and its interaction with the condition:

$$\text{choice} \sim 1 + \text{condition} + \text{moderator} + \text{condition} : \text{moderator} + (1 + \text{condition} + \text{moderator} + \text{condition} : \text{moderator} \mid \text{lab})$$

We no longer used informative models here because previous studies provided little evidence on setting priors. We used treatment contrasts for all categorical variables.[7]

There was a main effect of habituation status, such that infants who habituated were more likely to choose the Helper/Push-Up character (53.29% chose the Helper/Push-Up character) than infants who did not habituate (49.21%; $b = 0.70$, $SE = 0.32$, 95% CI [0.09, 1.33]). However, this main effect was qualified by an interaction between condition and habituation status, $BF_{10} = 4.25$ ($b = -1.00$, $SE = 0.45$, 95% CI [−1.87, −0.13]). To unpack the interaction, we examined the effect of habituation status on infants' choices in the social and nonsocial conditions separately. We found no evidence in favor of the effect of habituation status in either condition, $BF_{01} = 2.56$ (social condition) and $BF_{10} = 1.77$ (nonsocial condition). However, when we examined the effect of condition on infants' choice of the Helper in nonhabituated and habituated infants separately, we found moderate evidence in favor of the effect of condition for habituated infants, $BF_{10} = 4.34$ ($b = -0.48$, $SE = 0.22$, 95% CI [−0.92, −0.06], See Figure 5)[8]. Habituated infants in the nonsocial condition were more likely to choose the Helper/Push-Up Character (59.51%) than habituated infants in the social condition (47.88%). For nonhabituated infants, we found no evidence in favor of the effect of condition, $BF_{10} = 1.14$. Nonhabituated infants' choices of the Helper/Push-Up Character in the nonsocial condition (43.33%) and the social condition (54.55%) were not different.

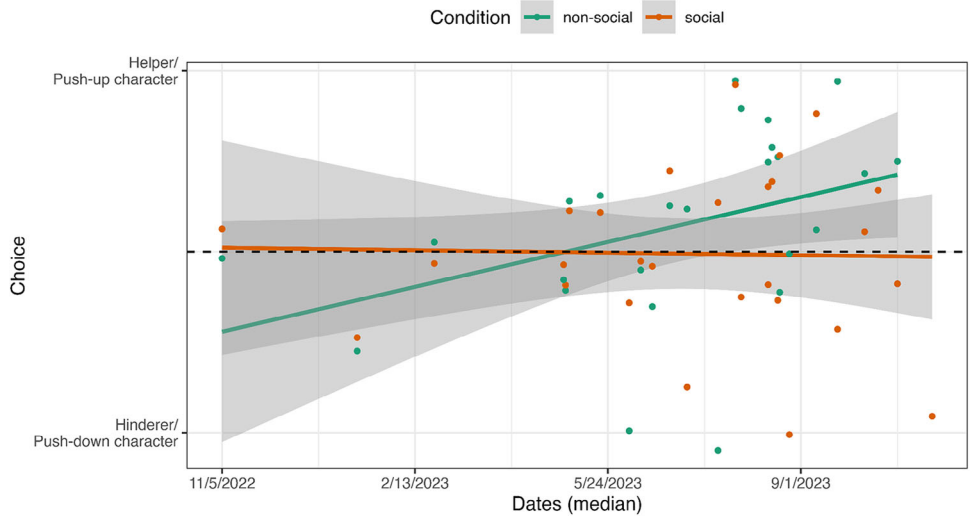

FIGURE 6 | Plot of median testing date × proportion of infants choosing Helper/Push-up character in each laboratory. *Note:* Proportion of infants choosing the Helper/Push-up character (over Hinderer/Push-down character) in each laboratory plotted by the laboratory's median testing date. The dashed line on the *y*-axis represents chance performance. The smoothing line shows the estimated effects with credible intervals. Data are jittered slightly on the vertical axis to avoid overplotting. Each point represents data from one lab.

There was also a main effect of laboratory median testing date, such that infants from labs with later median testing dates were more likely to choose the Helper/Push-Up character, $BF_{10} > 100$ ($b = 0.39$, $SE = 0.10$, 95% CI [0.01, 0.77]). However, the critical interaction between median testing date and condition did not have an effect on infants' choices, $BF_{01} = 3.70$ ($b = -0.29$, $SE = 0.27$, 95% CI [−0.82, 0.25], See Figure 6). All the other moderators and interaction effects had CIs that overlapped with zero (for details of the model parameter estimates, see Table 4).

3.4 | Exploratory Analyses on Infants' Looking Time

The above exploratory analyses focused on the effect of condition and possible moderators on infants' choices of the Helper. We also explored possible attentional differences across the social and nonsocial conditions. We conducted exploratory analyses on infants' gaze behavior during habituation, in particular how long they looked at the display following the pushing events (i.e., at the still image of the final frame presented after each video). To compare infants' looking time following the pushing events in the two conditions, we used the below model comparisons and the same method as above to obtain *BF*s. Noninformative priors were used in the models.

The model for H_1 was specified as:

$$looking\ time\ following\ pushing\ events \sim 1 + condition + age + condition : age + (1 + \text{condition} + \text{age} + \text{condition} : \text{age} \mid \text{lab})$$

The model for H_0 was specified as:

$$\text{looking time following pushing events} \sim 1 + \text{age} + (1 + \text{age} \mid \text{lab})$$

We found a $BF_{10} = 111.20$ in favor of H_1, providing strong evidence for the effects of condition and the interaction between age and condition on infants' looking time. The 95% CI for the condition parameter did not overlap with zero, $b = 0.97$, $SE = 0.34$, 95% CI [0.29, 1.62]. In other words, infants looked longer following the pushing events in the social condition ($M = 8.99$, $SD = 4.20$) than in the nonsocial condition ($M = 8.02$, $SD = 3.58$), suggesting infants were more attentive following the social videos relative to the nonsocial videos. Furthermore, we examined whether the addition of the condition-by-age interaction to the model contributed to model fit. The H_1 model was the same as above.

The model for H_0 was specified as:

$$\text{looking time following pushing events} \sim 1 + \text{age} + \text{condition} + (1 + \text{age} + \text{condition} \mid \text{lab})$$

We found a $BF_{10} = 3.33$ in favor of H_1, providing weak evidence for the effect of the interaction between age and condition on infants' looking time ($b = -0.69$, $SE = 0.33$, 95% CI [−1.35, −0.05]). Furthermore, we unpacked the interaction and analyzed the effect of condition for younger infants (i.e., 1 *SD* below the mean age) and older infants (i.e., 1 *SD* above the mean age). For younger infants, we found strong evidence in favor of the effect of the condition, $BF_{10} = 40.03$. Younger infants looked longer following the pushing events in the social condition ($M = 10.10$, $SD = 4.83$) than in the nonsocial condition ($M = 8.19$, $SD = 3.69$), $b = 1.80$, $SE = 0.83$, 95% CI [0.17, 3.43]. For older infants, we found no evidence in favor of the effect of the condition, $BF_{01} = 1.75$. We also examined the effect of age on infants' looking time in the social and nonsocial conditions separately. In the social condition, we found strong evidence in favor of the effect of age, $BF_{10} = 358.87$. Infants' looking time following the social events decreased with age, $b = -0.88$, $SE = 0.27$, 95% CI [−1.43, −0.34] (see Figure 7). In the non-social condition,

TABLE 4 | Parameter estimates for Bayesian models—Exploratory moderators.

Moderator	Parameter	Estimate	Est. error	95% CI
Number of trials to habituation	Intercept	−0.01	0.55	[−1.08, 1.08]
	Condition	−0.91	0.76	[−2.46, 0.52]
	hab_trial	0.05	0.06	[−0.07, 0.16]
	condition:hab_trial	0.05	0.08	[−0.11, 0.21]
Looking time following pushing events	Intercept	−0.12	0.34	[−0.79, 0.53]
	Condition	0.15	0.46	[−0.75, 1.06]
	Looking	0.05	0.04	[−0.04, 0.12]
	condition:looking	−0.05	0.05	[−0.15, 0.06]
Clear versus ambiguous choice	Intercept	0.23	0.15	[−0.07, 0.51]
	Condition	−0.23	0.21	[−0.63, 0.19]
	touch_both	0.17	0.41	[−0.62, 0.98]
	condition:touch_both	−0.25	0.69	[−1.60, 1.12]
Experimenter blindness	Intercept	0.27	0.56	[−0.80, 1.37]
	Condition	0.38	0.77	[−1.12, 1.91]
	Blindness	−0.02	0.58	[−1.18, 1.09]
	condition:blindness	−0.68	0.80	[−2.24, 0.89]
Order of helping versus hindering events	Intercept	0.30	0.18	[−0.07, 0.65]
	Condition	−0.10	0.26	[−0.61, 0.40]
	Order	−0.11	0.27	[−0.65, 0.41]
	condition:order	−0.32	0.36	[−1.02, 0.40]
Helper color	Intercept	0.34	0.18	[0.00, 0.69]
	Condition	−0.32	0.26	[−0.84, 0.17]
	helper_color	−0.22	0.27	[−0.75, 0.30]
	condition:helper_color	0.17	0.37	[−0.55, 0.90]
Helper side during choice	Intercept	0.26	0.19	[−0.11, 0.63]
	Condition	−0.24	0.29	[−0.82, 0.33]
	choice_side	0.01	0.31	[−0.60, 0.61]
	condition:choice_side	−0.09	0.39	[−0.81, 0.71]
Infants' visual angle to the display	Intercept	0.36	0.48	[−0.53, 1.33]
	Condition	−0.67	0.64	[−1.93, 0.57]
	v_angle	0.00	0.01	[−0.02, 0.01]
	condition:v_angle	0.01	0.01	[−0.02, 0.03]
Whether the experiment was the first session of infant's visit	Intercept	0.21	0.14	[−0.06, 0.48]
	Condition	−0.22	0.20	[−0.61, 0.17]
	first_sess	0.89	1.17	[−1.39, 3.31]
	condition:first_sess	−1.16	1.82	[−4.77, 2.41]
Whether or not the experimenter wore a mask	Intercept	0.31	0.16	[−0.01, 0.62]
	Condition	−0.26	0.23	[−0.71, 0.18]
	Mask	−0.23	0.38	[−1.00, 0.51]
	condition:mask	0.18	0.59	[−0.98, 1.32]
Infants' handedness	Intercept	−0.05	0.24	[−0.53, 0.42]
	Condition	−0.01	0.35	[−0.71, 0.66]

(Continues)

TABLE 4 | (Continued)

Moderator	Parameter	Estimate	Est. error	95% CI
	HandednessL	0.38	1.86	[−2.76, 4.29]
	HandednessR	0.36	0.43	[−0.47, 1.23]
	HandednessU	0.49	0.31	[−0.11, 1.09]
	condition:handednessL	−1.92	2.80	[−7.26, 2.41]
	condition:handednessR	−0.21	0.59	[−1.40, 0.9]
	condition:handednessU	−0.50	0.45	[−1.39, 0.37]
History of colorblindness in infant's family	Intercept	0.23	0.15	[−0.06, 0.52]
	Condition	−0.34	0.21	[−0.74, 0.09]
	Colorblind	0.12	0.74	[−1.23, 1.70]
	condition:colorblind	−0.06	1.09	[−2.23, 1.94]
Infant's sex	Intercept	0.24	0.20	[−0.16, 0.63]
	Condition	−0.19	0.27	[−0.72, 0.36]
	Sex	0.02	0.27	[−0.53, 0.52]
	condition:sex	−0.14	0.39	[−0.90, 0.62]
Percentage of primary language exposure	Intercept	−0.19	0.93	[−2.00, 1.65]
	Condition	−1.25	1.20	[−3.58, 1.13]
	lang_exp	0.00	0.01	[−0.01, 0.02]
	condition:lang_exp	0.01	0.01	[−0.01, 0.04]
Laboratory's overall exclusion rate	Intercept	0.88	0.41	[0.10, 1.70]
	Condition	−0.18	0.58	[−1.25, 1.01]
	ex_rate	−1.80	1.11	[−3.94, 0.44]
	condition:ex_rate	−0.19	1.56	[−3.31, 2.84]
The exclusion rate due to failure to make a choice for each lab	Intercept	0.59	0.24	[0.10, 1.06]
	Condition	−0.06	0.34	[−0.71, 0.61]
	no_choice	−2.70	1.58	[−5.90, 0.40]
	condition:no_choice	−1.53	2.21	[−5.92, 2.78]
Median data collection date of laboratories	Intercept	0.34	0.17	[−0.01, 0.67]
	Condition	−0.35	0.25	[−0.83, 0.15]
	median_date	0.39	0.20	[0.00, 0.78]
	condition:median_date	−0.29	0.27	[−0.84, 0.23]

Note: Parameter estimates along with estimated error and 95% credible intervals (CI) for those parameters for Bayesian models testing exploratory moderators.

we found no evidence in favor of the effect of age, $BF_{01} = 6.67$ (Figure 7).

Following the approach taken for confirmatory analyses, we completed all exploratory analyses using a frequentist approach. The results are highly consistent with those obtained through the Bayesian approach. Any discrepancies are documented in the code script available on GitHub (https://github.com/manybabies/mb4-analysis).

4 | Discussion

The extent to which infants make early social evaluations and prefer prosocial over antisocial agents continues to be a highly studied and frequently debated topic. However, methodological and procedural differences between studies render it difficult to directly compare findings. The present international, large-scale, and multi-site replication—using the ManyBabies framework for collaborative team-based science—allowed us to recruit 1018 infants from 37 labs from around the world, with a final sample of 567 infants from 34 labs included in the analyses. We used this model to address critical limitations of prior work and to establish a precise estimate of the effect size of infants' preference for helping over hindering agents in the Hill paradigm (Hamlin, Wynn, and Bloom 2007). Furthermore, we sought to test the extent to which infants' preferences for Helpers are social in nature by including a nonsocial control condition that was physically matched to the experimental condition. As the first of its kind, this project served as a proof-of-concept for the use of behavioral components (e.g., manual choice) in a large-scale, multi-lab replication and as a model for openly sharing video data

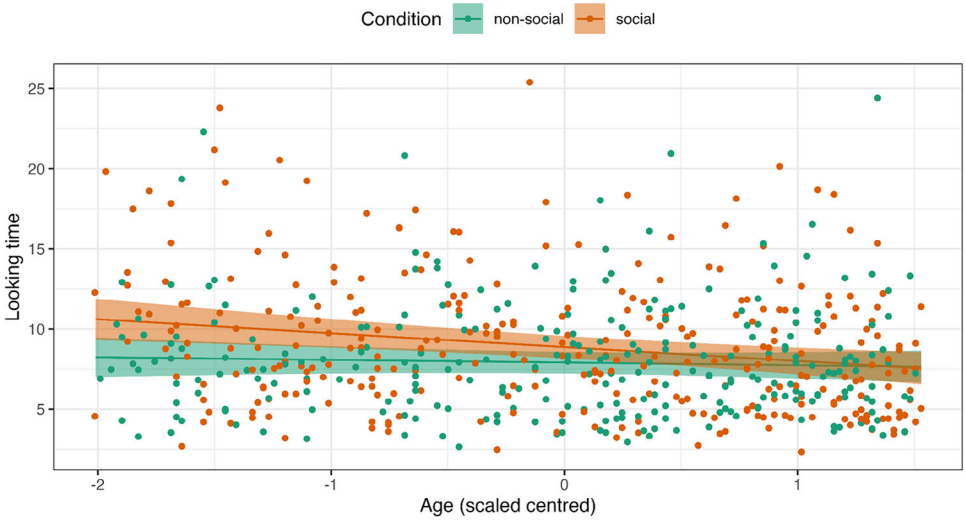

FIGURE 7 | Plot of age × Estimated freeze frame looking time. *Note:* Estimated looking time following the pushing events across ages. The smoothing line shows the predicted marginal effects from our noninformative Bayesian regression model along with their credible interval. Data are jittered slightly on the vertical axis to avoid overplotting.

across international borders. These videos can be reused in future "spin-off" projects to answer new research questions.

4.1 | Summary of Planned Analyses

The current study included infants across five continents, who participated in a standardized, preregistered paradigm that closely replicated the original hill study by Hamlin and colleagues (Hamlin, Wynn, and Bloom 2007) with a few key differences. Following our preregistered data analysis plan, we employed a Bayesian analysis framework and predicted that (1) infants in the social condition would show an above-chance tendency to choose the Helper over the Hinderer compared to infants in the nonsocial condition, and (2) infants' preference for the Helper over the Hinderer would not increase with age. In addition, we tested whether infants who habituated (vs. those who did not) would show a stronger preference for the Helper.

Inconsistent with our predictions, infants did not demonstrate a preference for the Helper over the Hinderer in the social condition, nor did they show a greater tendency to prefer Helpers in the social than in the nonsocial condition. Thus, the current work did not replicate the effect (preference for the Helper) observed in the original paradigm (Hamlin, Wynn, and Bloom 2007) and is inconsistent with a host of prior research demonstrating that infants show social preferences for Helpers and other prosocial agents over antisocial agents (e.g., Buon et al. 2014; Geraci, Simion, and Surian 2022; Geraci and Surian 2011, 2023; Hamlin, Wynn, and Bloom 2010; Kanakogi et al. 2017). Instead, these findings more closely align with past work that found no preference between Helpers and Hinderers (e.g., Schlingloff, Csibra, and Tatone 2020). These null results do not appear to be driven by age-related differences: Our analyses, consistent with Margoni and Surian (2018) and our prediction, provide no evidence that infants' social preferences for Helpers over Hinderers change across the second half of the first year of life. Similarly, our analyses showed that infants who habituated did not choose the Helper more often than infants who never habituated. Finally, the preregistered analysis revealed a low intra-class correlation, suggesting that while there was considerable variability in infants' preference for the Helper across labs, this variability was even higher across individuals. Together, this study provides evidence against infants' prosocial preferences in the hill paradigm, which suggests the effect is weaker, absent, and/or develops later, than previously estimated (e.g., Margoni and Surian 2018), and highlights that inconsistencies in past work may be partially due to overall high levels of variability in behavioral paradigms. We discuss these sources of variability in the sections below.

Our collaboration yielded a large sample size, which allowed a unique opportunity to explore potential moderating variables and interactions related to preferences for the Helper or Hinderer, usually not evaluated due to lack of power. The planned exploratory analyses tested whether infants' preference for the Helper was moderated by visual attention to video events, clarity of infants' choice actions, and experimenter awareness of the condition before choice presentation. We found no evidence supporting any moderating or interaction effects.

4.2 | Summary of Exploratory Analyses

We additionally conducted exploratory analyses to examine the relation between infants' attention to the stimuli and their subsequent social preferences. Infants' visual attention to the still frame following the videos varied by condition: Overall, infants in the social condition looked longer than those in the nonsocial condition. This conditional difference also interacted with age, such that younger infants showed a larger difference in visual attention between the social and nonsocial conditions than older infants. Together, these patterns suggest that infants, especially younger ones, found the social videos more engaging. Although this result is consistent with a host of other research demonstrating young infants' preferences for social over nonsocial stimuli

(e.g., preference for biological motion in Bardi, Regolin, and Simion 2011; face preference in Valenza et al. 1996), it remains unclear whether differences in infants' looking to videos across conditions were due to differences in the "plot" depicted in social versus nonsocial events (i.e., helping/hindering vs. push-up/push-down) or to various other lower-level differences (e.g., more eyes in the social videos, more motions, slight timing differences between social and nonsocial videos). Critically, infants' attention did not differ between helping/push-up trials and hindering/push-down trials, and the conditional differences did not predict infants' preference for the Helper or Hinderer. This finding is consistent with past work, which also found that infants' visual attention to the still frames following hill Helper/Hinderer events did not predict their preference, albeit with 15-month-olds and using a fixed number of familiarization trials rather than habituation (Schlingloff, Csibra, and Tatone 2020).

We also explored whether infants' preference was moderated by various lab-level and experiment-level factors, as well as participant-level characteristics, and probed potential sources of cross-lab variation in infants' preference for the Helper. We found a main effect of median laboratory testing date, such that infants tested by labs with later median testing dates more often chose the Helper/Push-up character. However, the interaction of interest between median testing date and condition was not present. Other analyses revealed that none of the lab- or experiment-level variables of interest, such as exclusion rates or visual angle of the stimuli (i.e., the ratio between the screen size and infant distance from the screen), were associated with individual labs' results. Similarly, there were no moderating or interaction effects of counterbalanced experiment-level factors (e.g., order of video presentation). Finally, participant-level characteristics such as sex, handedness, and primary language exposure did not predict infants' preference for the Helper. This lack of sex difference, in particular, is consistent with a recent meta-analysis (Margoni et al. 2023). We found weak evidence for habituation status as a moderator of infants' choices, such that habituated infants tended to choose the Helper/Push-Up character more often in the nonsocial condition, compared to the social condition. However, this effect was not present when age was included in the model (in the confirmatory analyses), suggesting it is not robust and therefore difficult to interpret. Altogether, results from these exploratory analyses suggest that infants' preference for Helpers was not moderated by any variables collected in our study.

4.3 | Interpretations and Implications

There are several possible ways to interpret the null results observed in the current study. First, it may be that infants between 5 and 10 months of age do not prefer prosocial characters over antisocial characters after all. That is, past research on infants' social evaluations that obtained positive results (e.g., Buon et al. 2014, with 10- and 29-month-olds; Dunfield and Kuhlmeier 2010, with 21-month-olds; Geraci and Surian 2023 with 4-month-olds; Geraci, Simion, and Surian 2022, with 9-month-olds; Hamlin and Wynn 2011, with 3-, 5-, and 9-month-olds; Hamlin 2013, with 5- and 8-month-olds; Kanakogi et al. 2017, with 6-month-olds; Kanakogi et al. 2022, with 8-month-olds; Scola et al. 2015, with 12- to 14-month-olds) have all significantly overestimated the true effect size of infants' preference for prosocial others, perhaps due to factors such as publication bias, other researcher degrees of freedom, and small sample size (Margoni and Surian 2018). Therefore, with a large and different sample size and a tightly controlled methodology, the present work provided the most robust and accurate estimate of this (null) effect. Indeed, the present findings are consistent with a large body of work on infants' social evaluations which failed to detect a preference for helpful agents (e.g., Abramson et al. 2016, unpublished, with 9- and 18-month-olds; Cowell and Decety 2015, with 12- to 24-month-olds; Nighbor et al. 2017, with 16-month-olds; Salvadori et al. 2015, with 9-month-olds; Schlingloff, Csibra, and Tatone 2020, with 14- to 16-month-olds; Shimizu et al. 2018, with 6-, 9-, and 12-month-olds; Vaporova and Zmyj 2020, with 9- and 14-month-olds). Under this view, these findings serve as strong evidence against infants' social evaluative capacities at this early age. Although demonstrating a lack of sensitivity to moral actions during infancy would not in itself provide irrefutable evidence that morality is not innate (nor would positive evidence "prove" innateness), these findings highlight the importance of considering nonnativist possibilities for the origins of morality. For instance, future work should consider the role of children's early experiences in shaping socio-moral cognition and prosocial behaviors (Dahl 2013; 2015; Rogoff et al. 2018).

A second, narrower interpretation is that infants do not prefer Helpers in the Hill paradigm specifically, but may nonetheless prefer Helpers (and prosocial agents more generally) in other contexts. This view is plausible, and perhaps more likely than the first, for at least two reasons. First, although the Hill paradigm is the first scenario in which evidence was found for infants' social evaluation, it is far from the only scenario to show infants' preference for prosocial agents in a helping/hindering context, nor is helping/hindering the only context in which infants have demonstrated positive evaluation of prosocial agents. Recall that the Hill paradigm was chosen due to its high impact on and pervasiveness in the literature, rather than its capacity to elicit the strongest effects. Its simplicity, which allowed for the most precise test of the construct of interest, comes with lower ecological and face validity compared to other paradigms that boast richer cues to agency or more familiar actions (Kominsky et al. 2022), and these tradeoffs may have ultimately resulted in weaker average effects. In fact, in the time since the decision to replicate the Hill paradigm was made, meta-analyses have shown that helping/hindering scenarios in general have yielded the smallest average effect size relative to other socially-relevant scenarios, such as fair versus unfair resource distributions and giving versus taking (note that giving vs. taking is sometimes, but sometimes not, referred to as helping vs. hindering in the literature; Margoni and Surian 2018; A. Tan 2024). Therefore, even if we were to assume the evidence presented herein to be the "nail in the coffin" for the Hill paradigm, it need not directly speak to other helping/hindering scenarios (e.g., the Box opening/closing and Ball giving/taking show in Hamlin and Wynn 2011), let alone other significantly different but nonetheless socially-relevant scenarios such as fair versus unfair resource distributions (e.g., Burns and Sommerville 2014; Lucca, Pospisil, and Sommerville 2018) and physical battery versus protection (e.g., Kanakogi et al. 2017).

Second, since the conception of the present project in 2018, several works have been published by various research groups

using a diverse set of stimuli (e.g., Geraci and Franchin 2021; Loheide-Niesmann et al. 2021; Schlingloff, Csibra, and Tatone 2020; Singh 2020; Vaporova and Zmyj 2020; Woo and Spelke 2020; 2023; see Woo et al. 2022 for recent review; see A. Tan, 2024 for recent unpublished meta-analysis), most of which have found a preference for agents performing different kinds of prosocial actions. For these two reasons, it seems plausible that even if infants do not prefer Helpers over Hinderers in the Hill paradigm by 10.5 months, they might nonetheless prefer Helpers (or prosocial agents more generally) in other scenarios, perhaps when the intentions of prosocial/antisocial agents and/or the protagonist are overt (Geraci, Simion, and Surian 2022; Hamlin 2013; Strid and Meristo 2020; for pro-environmental behaviors, see also Geraci, Franchin, and Benavides-Varela 2023). Of course, it is unreasonable to require every single possible paradigm to succeed or fail in large-scale collaborations before one makes broad generalizations about social evaluation. However, it may be equally unreasonable, or at least premature, to discount all other paradigms based on results from one study alone. Indeed, these findings highlight the need for alternative methods of conducting large-scale replications that might be better suited for assessing the truth-value of a given theoretical hypothesis. For instance, one could randomly assign infants and/or laboratories to distinct instances of prosocial and antisocial behaviors from across the literature, which (given sufficient power) would have allowed paradigm-level similarities/differences to be explored while also assessing evidence for/against the broader question of whether infants prefer prosocial to antisocial others.

It is also possible that the overall weak/null effect observed herein is a product of cross-lab variations in methodology, an interpretation supported by the cross-lab variation in infants' preference for Helpers. Indeed, although we standardized critical components of the study (e.g., stimuli, methods of presentation) as much as possible, certain parts of the procedure inevitably deviated between labs due to practical considerations. As but one example, we allowed labs to "warm up" with infants according to their own practices; variation in warm-up practices may result in some infants being more/less comfortable in the testing environment, and subsequently lead to different patterns in selective reaching. Overall, given the large number of ways that labs vary in their routine practices, we were not able to quantify the full range of ways labs differed in their procedures, and so we were unable to test whether and how they may have impacted infants' behaviors.

More specific to the methodology is the potential variance in the manual choice procedure. Indeed, whereas other multi-lab replication attempts in the past used a "plug-and-play" methodology that required virtually no experimenter interaction with the participants and relied entirely on looking-based measures (e.g., ManyBabies1; The ManyBabies Consortium 2020), the current project was the first to use a behavioral component for the primary dependent measure. The difficulty in standardizing our procedure is evident in our relatively high exclusion rate (40.63%) if we do not consider ineligible infants, and this exclusion rate dropped to 14.86% if we do not consider experimenter or equipment errors. Although all participating experimenters underwent a standardized training process, and required approval from the central team to be eligible to both administer and code the choice phase, due to ethics constraints (e.g., laboratories unable to share participant videos), we were unable to centrally code all infant reaches. As such, variations both in how each individual laboratory administered the choice phase, as well as how infants' manual reaches were subsequently coded could have led to differences in infants' likelihood of choosing the Helper. That said, it should be noted that our attrition rate is comparable to past replication attempts that also involved the original author (e.g., 37.25% in Schlingloff, Csibra, and Tatone 2020). Thus, despite the null findings, we nonetheless consider the current project a successful proof-of-concept for adapting and scaling up paradigms with behavioral components.

On a different methodological level, our null findings could have been driven by differences between our stimuli and the original Helper/Hinderer "show" in Hamlin, Wynn, and Bloom (2007). First, we used videos versus live events as had been used in the original studies, both in order to maximize consistency across labs and to make it possible for labs that lacked puppet stage setups to run the study. Although we are aware of one published study that successfully utilized video recordings of the Hill show with a habituation design (Hamlin 2015), another lab failed to elicit a preference for Helpers using an animated cartoon version of the Hill show with a familiarization design (Schlingloff, Csibra, and Tatone 2020). Notably, the videos used in the current study were designed to match both each other and example videos of the original live show used in Hamlin, Wynn, and Bloom (2007) as closely as possible. However, because it is neither feasible nor practical for puppeteers performing the Hill events to perfectly reproduce the same show, this matching was achieved via heavy editing in postproduction (e.g., by speeding up and slowing down sections of the video).

It should also be noted that all of the videos used in the current study were created following the initial peer review process due to the need to include matching non-social videos, and therefore, due to COVID-19, none of these videos were piloted prior to testing (though these videos were closely matched to the ones that were piloted during Stage 1 submission, albeit slightly faster in order to minimize the timing difference between the social and nonsocial condition videos). Although there is no reason to suspect that the artificially sped up/slowed down sections impacted the perception of helping and hindering, it is notable that the live shows used in the original studies (12 s long on average), as well as the videos used in the social condition in the current study (13.2 s), were quite fast, whereas the cartoon videos used in subsequent hill displays shown on screens were on average significantly slower (e.g., E. Tan and Hamlin 2022; 16.8 s). These timing differences may be crucial, given that a recent video-based eye-tracking study found a positive association between anticipatory looking to the top of the hill during the Climber's failed attempts and subsequent visual preference for the Helper (E. Tan and Hamlin 2022). Specifically, infants who showed gaze shifts between the Climber and the hilltop (which has been argued to signal goal understanding) visually preferred the Helper at the test, whereas infants whose gaze fixated solely on the Climber (which has been argued as signaling a lack of goal understanding) did not. Thus, it is possible that increasing the overall pace of the show in turn decreased the amount of time infants had to process the Climber's goal, which may have influenced their subsequent social evaluations. This possibility should be explored in future work.

Second, as evidenced by a recent unpublished meta-analysis (A. Tan 2024), infant social evaluation studies in which stimuli are presented on video elicit smaller effect sizes than those with live puppet shows (note that a previous meta-analysis, with less data, did not observe this effect; Margoni and Surian 2018); this effect size reduction may be due to relatively lower attention and affective responses to video stimuli compared to real-life events (e.g., Barr 2010; Diener et al. 2008). Although, as noted above, there has been a successful replication using a similar video-based procedure (Hamlin 2015), those shows were slower than the ones in the current study. Thus, it is possible that the combination of faster-paced videos with potentially lower attentiveness/interest in video stimuli may have negatively impacted infants' understanding of the stimuli's overall message.

An additional notable methodological difference between our current work and the original 2007 Hill study was the (intentional) omission of a bouncing event that occurred when the Climber reached the top of the hill in the helping event of the social condition. This omission may have been problematic, as a previous study had failed to replicate a preference for Helpers over Hinderers when the bouncing event was also included in the hindering event (Scarf et al. 2012). Although this alternative explanation was subsequently ruled out in a successful replication that did not include bouncing during helping events and uncovered a preference for Helpers (Hamlin 2015), the absence of a significant preference for the Helper over the Hinderer in the current study with a broader sample could nonetheless suggest that bouncing was an important component of infants' social evaluations in the original Hill studies. Importantly, as we did not compare across conditions where the bouncing event was present or absent in the current study, we are not able to determine the impact of this factor relative to other differences from the original stimuli. We encourage future work to systematically investigate what lower-level factors, such as the bouncing action as well as the overall pacing discussed above, impact infants' understanding and interpretation of helping/hindering actions, which would in turn shed light on reasons behind failures to replicate the effects of seminal work.

Finally, the present findings may simply reflect fundamental differences between meta-analyses and multi-laboratory replications. Indeed, prior work has revealed that the effect sizes in meta-analyses tend to be three times larger than in multi-laboratory replication studies (Kvarven, Strømland, and Johannesson 2020), though the exact reasons underlying these discrepancies remain unclear (Lewis et al. 2022). It is noteworthy that even when effect size estimates are similar between meta-analyses and large-scale replication projects, the two sources may diverge in the effects of key moderators such as infant age and experimental method. This is the case of infant preference for infant-directed speech (IDS), where preference of IDS over adult-directed speech (ADS) reportedly increased with infant age according to a large-scale replication (The ManyBabies Consortium 2020), but remained stable according to meta-analytic estimates (Bergmann et al., 2018; Dunst, Gorman, and Hamby 2012, and community-augmented meta-analyses https://metalab.stanford.edu; Zettersten et al. 2024), despite having the same overall effect size (Zettersten et al. 2024). This may be because applying the very same procedure to infants across age, as happens in large-scale collaborations like ours but not in meta-analyses, can impact the size of an observed effect given age-related differences in attention, processing speed, and so forth. This suggests that meta-analyses and large-scale replications should likely be used in tandem to assess the robustness of effects of interest in the literature.

4.4 | Challenges and Limitations

Though this project was successful on many fronts, there are limitations that might have influenced the observed pattern of results. First, the initial planned window for participating labs to collect data spanned from summer 2021 to summer 2022. However, the COVID-19 pandemic posed significant challenges for participating laboratories and necessitated an extension, from summer 2022 to fall 2023. In addition, the pandemic introduced large cross-lab, -region, and -country variabilities in research protocols over which our researchers had no control (e.g., masking, time researchers spent warming up with infants prior to testing, infant familiarity with new environments and people). For instance, many labs had to shut down for substantial periods of time, whereas others remained open or had brief closures. These closures could have led to recruiting challenges when testing became possible (e.g., parents not yet comfortable with in-person sessions), and researchers who were out of practice for testing infants in person. In addition, we were unable to conduct planned in-person training at international conferences, which may have further contributed to across-lab procedural deviations.

Perhaps a more important factor is the potential impact on participating infants. Though many of the infants who participated in the current study were born after the height of pandemic-related restrictions (e.g., social distancing, mask wearing), the lingering effects of the pandemic may have led to a reduction in opportunities for social interaction and observation, which may have limited exposure to the types of social experience that are important for sociomoral development in infancy. Reduced social interactions with strangers may have also resulted in infants feeling more anxious about the testing environment, and this increased anxiety may have subsequently affected their performance. Although the exact impact of the COVID-19 pandemic on the study is difficult to quantify, and a growing body of research suggests that infants are surprisingly resilient toward direct adverse effects of the pandemic (see Lobue et al. 2023, for a review), no work to our knowledge has directly tested whether infants who grew up during the pandemic show heightened stranger anxiety, and how this potential heightened anxiety affects subsequent performance in social evaluative tasks. The present work crudely explored potential cohort effects by testing the effect of median testing dates; however, it should be noted that the date of testing is only meaningful if considered alongside related environmental factors (e.g., regional lockdown policies, laboratory testing policies). For instance, two laboratories with similar median testing dates may in fact have wildly different masking policies, or be located in regions with varying degrees of COVID restrictions. In addition, other factors that have been shown to impact infants' sociocognitive development (e.g. maternal stress; Nazzari et al. 2024) may be correlated with some or all of these environmental variables. As such, we note the theoretical and practical relevance of testing the effects of the pandemic, but are cautious in interpreting these exploratory results. We look forward to other future work, such as longitudinal work that includes cohorts tested before, during, and

after the pandemic, as well as spin-off work that capitalizes on the large dataset collected by this project and can more thoroughly shed light on these questions.

Finally, despite the participation of 37 labs from 18 different world regions, it is important to note that the participants included in this study mostly came from industrialized and educated societies. This concern applies broadly to previous studies investigating social evaluations in infancy (see the populations included in the Margoni and Surian 2018 meta-analysis), and within the field of developmental psychology broadly (Nielsen et al. 2017). Intriguingly, a recent study documented variation in infants' (12- to 20-months of age) expectations about third-party resource distribution between a Western (Sweden) and two non-Western populations (Samburu and Kikuyu infants in Kenya; Meristo and Zeidler 2022); these differing expectations for fairness in infants in the second year might reflect that infants are sensitive to the high levels of variation in fairness norms and behaviors that exist across diverse societies (Blake et al. 2015; Henrich, Heine, and Norenzayan 2010). In contrast, because helping behaviors have been shown to emerge early and consistently across diverse sociocultural contexts (Callaghan et al. 2011; Callaghan and Corbit 2018; Kärtner, Keller, and Chaudhary 2010), it may be that an early preference for helpful agents, if and when it appears, will emerge similarly consistently across diverse societies. It is crucial that future work continues to include participants from diverse sociocultural environments, especially from small-scale traditional societies, which are underrepresented in many developmental psychology samples, including the current study, in order to address these possibilities and otherwise gain a globally representative view of social evaluations in early childhood.

5 | Conclusion

Rigorous replication efforts are cornerstones to scientific progress; they contribute to bolstering the accuracy of findings, examining conditions under which findings are observed, and approximating findings' true effect sizes (Open Science Collaboration 2015; Zwaan et al. 2018). The current study, carried out through the ManyBabies collaboration, found no preference for Helpers over Hinderers in the Hill paradigm (Hamlin, Wynn, and Bloom 2007) in a large and diverse sample of 6- to 10-month-olds; these findings are inconsistent with the original study and with past meta-analyses. We additionally tested a number of moderators that could potentially explain the variability in results among the labs (e.g., attention to the stimuli, age, and sex), but did not find support for any moderating effects. Our study also found substantial variation in effect size across labs spanning diverse regions, though it remains unclear what exact factors contribute to these differences, and whether these differences merely reflect random variability associated with complex behavioral paradigms. Finally, this work took the first steps toward building best-practices for utilizing research paradigms with behavioral components in large-scale, multi-site collaborative research. We hope that the work presented herein lays the groundwork upon which future work on the development of social evaluations can build, and that novel works will elucidate what factors underlie replication failures and cross-lab differences.

Author Contributions

Conceptualization: Annette M. E. Henderson, Denis Tatone, Florina Uzefovsky, Francis Yuen, J. Kiley Hamlin, Jessica Sommerville, Jonathan F. Kominsky, Kelsey Lucca, Laura Schlingloff-Nemecz, Michael C. Frank, Yiyi Wang, and Zoe Liberman. **Data curation**: Annette M. E. Henderson, Bill Pepe, Denis Tatone, Francis Yuen, Kelsey Lucca, Samuel H. Forbes, Wenxi Fei, and Yiyi Wang. **Formal analysis:** Alvin W. M. Tan, Amanda M. Woodward, Annette M. E. Henderson, Arthur Capelier-Mourguy, David Moreau, Denis Tatone, Francis Yuen, J. Kiley Hamlin, Julien Mayor, Kelsey Lucca, Michael C. Frank, Nicolás Alessandroni, Samuel H. Forbes, Wenxi Fei, Arthur Capelier-Mourguy, and Yiyi Wang. **Funding acquisition:** Annette M. E. Henderson, Arkadiusz Urbanek, Casey Lew-Williams, Francis Yuen, Hannes Rakoczy, Hernando Taborda-Osorio, Hilal H. Şen, Hyun-joo Song, J. Kiley Hamlin, Kelsey Lucca, Laura K. Cirelli, Lindsey J. Powell, Liquan Liu, Mark A. Schmuckler, Melanie Soderstrom, Moritz M. Daum, Piotr Sorokowski, Tobias Grossmann, Tobias Schuwerk, Yanjie Su, and Yenny Otálora. **Investigation**: Alia Martin, Alyssa A. Quinn, Anna Exner, Annette M. E. Henderson, Annie E. Wertz, Arkadiusz Urbanek, Arthur Capelier-Mourguy, Bailey A. Immel, Bill Pepe, Casey Lew-Williams, Chantelle S. -S. Chin, Charisse B. Pickron, Charlotte Grosse Wiesmann, Chiu Kin Adrian Tsang, Denis Tatone, Emma L. Axelsson, Emmy Higgs Matzner, Florina Uzefovsky, Francis Yuen, Grzegorz Jankiewicz, Hannes Rakoczy, Hernando Taborda-Osorio, Hilal H. Şen, Hitomi Chijiiwa, Hyun-joo Song, Ingmar Visser, Isabelle M. Hadley, J. Kiley Hamlin, Janina Baumer, John Corbit, Julie Bertels, Karola Schlegelmilch, Katrin Rothmaler, Kelsey Lucca, Krista Byers-Heinlein, Kristina Wolsey, Laura K. Cirelli, Laura Franchin, Laura Schlingloff-Nemecz, Linda S. Oña, Lindsey J. Powell, Liquan Liu, Lucie Zimmer, Madison Williams, Marina Proft, Mario Alvarez, Mark A. Schmuckler, Markus Paulus, Megan E. Gornik, Melanie Soderstrom, Michaela Dresel, Michal Misiak, Michelle Giraud, Mira L. Nencheva, Miriam T. Loeffler, Moritz M. Daum, Naomi Havron, Natalie Christner, Nicolás Alessandroni, Olivia Allison, Peter J. Reschke, Piotr Sorokowski, Ronit Roth-Hanania, Sabine Seehagen, Sandro E. Stutz, Teresa Taylor-Partridge, Terry Tin-Yau Wong, Tiffany Doan, Tobias Grossmann, Tobias Schuwerk, Valentina Silvestri, Wiktoria Jędryczka, Xianwei Meng, Yang Wu, Yanjie Su, Yasuhiro Kanakogi, Yenny Otálora, Yiyi Wang, Zhen Zeng, and Zoe Liberman. **Methodology:** Annette M. E. Henderson, Chantelle S.-S. Chin, Denis Tatone, Florina Uzefovsky, Francis Yuen, J. Kiley Hamlin, Jessica Sommerville, Jonathan F. Kominsky, Kelsey Lucca, Laura Schlingloff-Nemecz, Megan E. Gornik, Michael C. Frank, Sandro E. Stutz, and Zoe Liberman. **Project administration**: Francis Yuen, Heidi A. Baumgartner, J. Kiley Hamlin, Kelsey Lucca, Michael C. Frank, Michal Misiak, Munna R. Shainy, Wiktoria Jędryczka, Yilin Liu, and Yiyi Wang. **Resources:** Arkadiusz Urbanek, Emma L. Axelsson, Francis Yuen, Grzegorz Jankiewicz, Hannes Rakoczy, J. Kiley Hamlin, Katrin Rothmaler, Laura Franchin, Mario Alvarez, Moritz M. Daum, Sandro E. Stutz, and Yilin Liu. **Software**: Francis Yuen, J. Kiley Hamlin, Jonathan F. Kominsky, Kelsey Lucca, Michael C. Frank, Samuel H. Forbes, and Sandro E. Stutz. **Supervision:** Alia Martin, Annette M. E. Henderson, Annie E. Wertz, Arkadiusz Urbanek, Bill Pepe, Casey Lew-Williams, Chantelle S. -S. Chin, Charisse B. Pickron, Charlotte Grosse Wiesmann, Chiu Kin Adrian Tsang, Denis Tatone, Emma L. Axelsson, Emmy Higgs Matzner, Florina Uzefovsky, Francis Yuen, Hannes Rakoczy, Hernando Taborda-Osorio, Hyun-joo Song, Isabelle M. Hadley, J. Kiley Hamlin, Janina Baumer, Jessica Sommerville, Julie Bertels, Karola Schlegelmilch, Kelsey Lucca, Krista Byers-Heinlein, Laura K. Cirelli, Laura Franchin, Laura Schlingloff-Nemecz, Lindsey J. Powell, Liquan Liu, Lucie Zimmer, Mario Alvarez, Mark A. Schmuckler, Markus Paulus, Melanie Soderstrom, Michael C. Frank, Michal Misiak, Michelle Giraud, Miriam T. Loeffler, Moritz M. Daum, Natalie Christner, Piotr Sorokowski, Sandro E. Stutz, Teresa Taylor-Partridge, Terry Tin-Yau Wong, Tiffany Doan, Tobias Schuwerk, Valentina Silvestri, Yang Wu, Yanjie Su, Yenny Otálora, Yiyi Wang, Zhen Zeng, and Zoe Liberman. **Validation**: Alvin W. M. Tan, Amanda M. Woodward, Francis Yuen, Julien Mayor, Kelsey Lucca, Michael C. Frank, Samuel H. Forbes, and Yiyi Wang. **Visualization**: Alvin W. M. Tan, Francis Yuen, Julien Mayor, Kelsey Lucca, and Yiyi Wang. **Writing–original draft:** Alessandra

Geraci, Annette M. E. Henderson, Arthur Capelier-Mourguy, David Moreau, Denis Tatone, Florina Uzefovsky, Francis Yuen, J. Kiley Hamlin, John Corbit, Jonathan F. Kominsky, Kelsey Lucca, Laura Schlingloff-Nemecz, Mitali Bhavsar, Munna R. Shainy, Peter J. Reschke, Samuel H. Forbes, Teresa Taylor-Partridge, Yilin Liu, Yiyi Wang, and Zoe Liberman. **Writing–review and editing:** Alessandra Geraci, Alvin W. M. Tan, Amanda M. Woodward, Annette M. E. Henderson, Charisse B. Pickron, Charlotte Grosse Wiesmann, David Moreau, Denis Tatone, Emma L. Axelsson, Florina Uzefovsky, Francis Yuen, Heidi A. Baumgartner, Hilal H. Şen, J. Kiley Hamlin, Janina Baumer, John Corbit, Jonathan F. Kominsky, Julien Mayor, Kelsey Lucca, Laura Schlingloff-Nemecz, Liquan Liu, Michael C. Frank, Mitali Bhavsar, Moritz M. Daum, Munna R. Shainy, Nicolás Alessandroni, Peter J. Reschke, Samuel H. Forbes, Teresa Taylor-Partridge, Wenxi Fei, Yanjie Su, Yilin Liu, Yiyi Wang, and Zoe Liberman.

Affiliations

[1]Department of Psychology, Arizona State University, Tempe, Arizona, USA | [2]Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada | [3]Department of Psychology, University of Chicago, Chicago, Illinois, USA | [4]Department of Psychology, Concordia University, Montreal, Quebec, Canada | [5]Department of Psychology, University of Virginia, Charlottesville, Virginia, USA | [6]School of Psychological Sciences, University of Newcastle, Callaghan, New South Wales, Australia | [7]Department of Psychology, University of Amsterdam, Amsterdam, Noord-Holland, Netherlands | [8]Center for the Study of Language and Information, Stanford University, Stanford, California, USA | [9]Center for Research in Cognition and Neurosciences, Université Libre de Bruxelles, Brussels, Belgium | [10]ARISA Foundation, Baner, Maharashtra, India | [11]Department of Psychology, Lancaster University, Lancaster, Lancashire, UK | [12]Graduate School of Human Sciences, Osaka University, Suita, Osaka, Japan | [13]Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany | [14]Department of Psychology, University of Toronto Scarborough, Scarborough, Ontario, Canada | [15]Department of Psychology, St. Francis Xavier University, Antigonish, Nova Scotia, Canada | [16]Jacobs Center for Productive Youth Development, University of Zurich, Zurich, Switzerland | [17]Department of Psychology, University of Zurich, Zurich, Switzerland | [18]School of Psychology, Victoria University of Wellington, Wellington, Wellington, New Zealand | [19]Faculty of Psychology, Ruhr University Bochum, Bochum, Germany | [20]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Hong Kong | [21]Department of Psychology, Durham University, Durham, UK | [22]Department of Psychology and Cognitive Science, University of Trento, Trento, Italy | [23]Department of Psychology, Stanford University, Stanford, California, USA | [24]Department of Educational Sciences, University of Catania, Catania, Sicily, Italy | [25]Department of Psychology, University of Milano-Bicocca, Milano, Italy | [26]Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada | [27]Minerva Fast Track Group Milestones of Early Cognitive Development, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Sachsen, Germany | [28]School of Psychological Sciences, University of Haifa, Haifa, Israel | [29]Center for Child Development, University of Haifa, Haifa, Israel | [30]School of Psychology, The University of Auckland | Waipapa Taumata Rau, Auckland, New Zealand | [31]Institute of Child Development, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA | [32]Department of Psychological & Brain Sciences, University of California Santa Barbara, Santa Barbara, California, USA | [33]Institute of Psychology, University of Wrocław, Wrocław, Poland | [34]Department of Cognitive Science, Central European University, Vienna, Austria | [35]Department of Psychology, Princeton University, Princeton, New Jersey, USA | [36]Graduate School of Health, University of Technology Sydney, Sydney, New South Wales, Australia | [37]School of Psychology, Western Sydney University, Pentrith, New South Wales, Australia | [38]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Westmead, New South Wales, Australia | [39]School of Behavioral and Brain Sciences, University of Texas at Dallas, Richardson, Texas, USA | [40]Department of Psychology, University of Oslo, Oslo, Norway | [41]Graduate School of Informatics, Nagoya University, Nagoya, Aichi, Japan | [42]IDN Being Human, Institute of Psychology, University of Wroclaw, Wroclaw, Poland | [43]School of Anthropology & Museum Ethnography, University of Oxford, Oxford, UK | [44]Centre for Brain Research, University of Auckland, Auckland, New Zealand | [45]Institute of Biology, Department of Human Biology and Primate Cognition, University of Leipzig, Leipzig, Germany | [46]Max Planck Research Group Naturalistic Social Cognition, Max Planck Institute for Human Development, Berlin, Germany | [47]Facultad de Psicología, Universidad del Valle, Cali, Colombia | [48]Department of Psychology, University of California San Diego, La Jolla, California, USA | [49]Institute of Psychology, University of Göttingen, Gottingen, Niedersachsen, Germany | [50]School of Family Life, Brigham Young University, Provo, Utah, USA | [51]Department of Psychology, The Academic College of Tel-Aviv Yaffo, Tel-Aviv, Israel | [52]Humboldt Research Group for Child Development, Faculty of Education, Leipzig University, Leipzig, Sachsen, Germany | [53]TUM School of Social Sciences and Technology, Technische Universität München, Munich, Bavaria, Germany | [54]Faculty of Psychology, University of Akureyri, Akureyri, Iceland | [55]Axxonet Brain Research Laboratory, Bengaluru, Karnataka, India | [56]Department of Psychology, University of Toronto, Toronto, Ontario, Canada | [57]Department of Psychology, Yonsei University, Seoul, Republic of Korea | [58]School of Psychology and Cognitive Sciences, Peking University, Beijing, China | [59]Departamento de Psicología, Pontificia Universidad Javeriana, Bogota, Colombia | [60]Department of Psychology, University of the Incarnate Word, San Antonio, Texas, USA | [61]Department of Psychology, The University of Hong Kong, Hong Kong, Hong Kong | [62]Institute of Pedagogy, Faculty of Pedagogical and Historical Sciences, University of Wrocław, Wroclaw, Poland | [63]Department of Psychology, Ben-Gurion University in the Negev, Beer-Sheva, Israel | [64]Department of Psychology, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA | [65]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Sha Tin, New Territories, Hong Kong SAR

Acknowledgments

We would like to thank the many families who participated in this research. We would also like to thank all of the researchers who participated in various aspects of this study, including planning, pilot testing, participant recruitment, data collection, data management, and providing feedback on the manuscript, including: Siobhan Kennedy-Costantini, Florian Bednarski, Francesco Margoni, Yuju Shin, Alexander Withers, Georgina Mariyadas, Ninthujah Suthaharan, Angela Le, Beverly Lundeen, Cheyenne Engeser, Dulce Erdt, Duygu Hilal Bayir, Janek Stahlberg, Daniel Tatlow-Devally, Julia Wolf, Dorottya Mészégető, Chiara Vollmerhausen, Hafdís Kristný Haraldsdóttir, Stefanía Guðrún Eyjólfsdóttir, Léa Simon, Solenn Gouadon, Madeleine Gale, Mateo Belalcazar, Laura Mejía, Mayte Alvarado, Madalyn Brown, Meghan Pierce, Moxuan Liu, Xingyue Han, Sara Trnovska, Ariana Martinez, Silvia Saletti, Vanessa Y. Dawidowicz, and Laura I. Blank.

Conflicts of Interest

The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the funding agencies listed here. The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available on OSF at https://osf.io/qntyd/ and the code is available on GitHub at https://github.com/manybabies/mb4-analysis.

Endnotes

[1] Note that a brief extension was given to one lab.

[2] Pilot testing revealed that brief look-aways outside this Critical Period were common, especially during trials near the end of the habituation phase.

[3] In our Registered Report, we planned to run simulations to determine whether it is more appropriate to keep or drop random effects corresponding to the dropped fixed effects from our model comparisons. We planned to choose the model comparison method that yielded the best Type I and II error rates, so long as these models successfully converge. In analyzing the data, we came to the conclusion that because there are no fixed effects included in the null model, the random effect should be dropped because it indicates random variation of the fixed effect, which is not included in the null model (Stroup 2012). Therefore, in our analysis, we dropped the random effects of the predictors whose fixed effects were dropped for nested model comparisons.

[4] The notation BF_{10} refers to the Bayes factor for the alternative hypothesis. Conversely, the notation BF_{01} refers to the Bayes Factor for the null hypothesis.

[5] The frequentist analyses revealed a main effect of condition, where habituated infants were more likely to choose the helper in the nonsocial condition relative to the social condition. However, the addition of condition did not improve overall model fit, so this main effect was not interpreted further.

[6] Median testing date was added as a moderator after the helpful suggestion from one of the reviewers.

[7] The categorical variables and their reference levels are as follows: clear versus ambiguous choice actions (clear), whether the experimenter had knowledge of the participant's condition (naive), the order of helping versus hindering videos (helping first), Helper identity (blue shape), Helper side during the choice phase (left), whether the experiment was conducted as the first session of the infant's visit or not (first session), whether the experimenter wore a mask or not (no mask), habituation status (nonhabituated), handedness (use both equally), color blindness (no), participant sex (female).

[8] The model here is "choice ~ 1 + condition + (1 + condition | lab)" for habituated and nonhabituated infants run separately. Note that in our confirmatory analysis, described earlier, we included both age and the interaction between age and condition in the model, and did not detect an effect of condition.

References

Abramson, L., M. Dar, A. Te'eni, and A. Knafo-Noam. 2016. Preferences for Helpers and Hinders in 9- and 18-Month-Old Infants. Unpublished raw data.

Bardi, L., L. Regolin, and F. Simion. 2011. "Biological Motion Preference in Humans at Birth: Role of Dynamic and Configural Properties." *Developmental Science* 14, no. 2: 353–359. https://doi.org/10.1111/j.1467-7687.2010.00985.x.

Barr, R. 2010. "Transfer of Learning Between 2D and 3D Sources During Infancy: Informing Theory and Practice." *Developmental Review* 30, no. 2: 128–154. https://doi.org/10.1016/j.dr.2010.03.001.

Bergmann, C., S. Tsuji, P. E. Piccinini, et al. 2018. "Promoting Replicability in Developmental Research Through Meta-Analyses: Insights From Language Acquisition Research." *Child Development* 89: 1996–2009. https://doi.org/10.1111/cdev.13079.

Blake, P. R., K. McAuliffe, J. Corbit, et al. 2015. "The Ontogeny of Fairness in Seven Societies." *Nature* 528, no. 7581: 258–261. https://link.gale.com/apps/doc/A437223659/HRCA?u=anon~1cb3f10e&sid=googleScholar&xid=915c05ec.

Brandone, A. C., and H. M. Wellman. 2009. "You Can't Always Get What You Want: Infants Understand Failed Goal-Directed Actions." *Psychological Science* 20, no. 1: 85–91. https://doi.org/10.1111/j.1467-9280.2008.02246.x.

Brownell, C. A. 2013. "Early Development of Prosocial Behavior: Current Perspectives." *Infancy* 18, no. 1: 1–9. https://doi.org/10.1111/infa.12004.

Buon, M., P. Jacob, S. Margules, et al. 2014. "Friend or Foe? Early Social Evaluation of Human Interactions." *PLoS ONE* 9, no. 2: e88612. https://doi.org/10.1371/journal.pone.0088612.

Bürkner, P.-C. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80, no. 1: 1–28. https://doi.org/10.18637/jss.v080.i01.

Burns, M. P., and J. A. Sommerville. 2014. "'I Pick You': The Impact of Fairness and Race on Infants' Selection of Social Partners." *Frontiers in Psychology* 5: 93. https://doi.org/10.3389/fpsyg.2014.00093.

Byers-Heinlein, K., C. Bergmann, C. Davies, et al. 2020. "Building a Collaborative Psychological Science: Lessons Learned From ManyBabies 1." *Canadian Psychology/Psychologie Canadienne* 61, no. 4: 349–363. https://doi.org/10.1037/cap0000216.

Callaghan, T., and J. Corbit. 2018. "Early Prosocial Development Across Cultures." *Current Opinion in Psychology* 20: 102–106. https://doi.org/10.1016/j.copsyc.2017.07.039.

Callaghan, T., H. Moll, H. Rakoczy, et al. 2011. "Early Social Cognition in Three Cultural Contexts." *Monographs of the Society for Research in Child Development* 76, no. 2: vii–viii, 1–142. https://doi.org/10.1111/j.1540-5834.2011.00603.x.

Carpendale, J. I., and C. Lewis. 2004. "Constructing an Understanding of Mind: The Development of Children's Social Understanding Within Social Interaction." *Behavioral and Brain Sciences* 27, no. 1: 79–96. https://doi.org/10.1017/s0140525X04000032.

Carter, E. C., F. D. Schönbrodt, W. M. Gervais, and J. Hilgard. 2019. "Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods." *Advances in Methods and Practices in Psychological Science* 2, no. 2: 115–144. https://doi.org/10.1177/2515245919847196.

Chae, J. J., and H. Song. 2018. "Negativity Bias in Infants' Expectations About Agents' Dispositions." *British Journal of Developmental Psychology* 36: 620–633. https://doi.org/10.1111/bjdp.12246.

Chakraborty, H., and A. Hossain. 2018. "R Package to Estimate Intracluster Correlation Coefficient With Confidence Interval for Binary Data." *Computer Methods and Programs in Biomedicine* 155: 85–92. https://doi.org/10.1016/j.cmpb.2017.10.023.

Cowell, J. M., and J. Decety. 2015. "Precursors to Morality in Development as a Complex Interplay Between Neural, Socioenvironmental, and Behavioral Facets." *Proceedings of the National Academy of Sciences of the United States of America* 112, no. 41: 12657–12662. https://doi.org/10.1073/pnas.1508832112.

Dahl, A. 2015. "The Developing Social Context of Infant Helping in Two U.S. samples." *Child Development* 86, no. 4: 1080–1093. https://doi.org/10.1111/cdev.12361.

Dahl, A., R. K. Schuck, and J. J. Campos. 2013. "Do Young Toddlers Act on Their Social Preferences?" *Developmental Psychology* 49, no. 10: 1964–1970. https://doi.org/10.1037/a0031460.

Diener, M. L., S. L. Pierroutsakos, G. L. Troseth, and A. Roberts. 2008. "Video Versus Reality: Infants' Attention and Affective Responses to Video and Live Presentations." *Media Psychology* 11, no. 3: 418–441.

Dunfield, K. A., and V. A. Kuhlmeier. 2010. "Intention-Mediated Selective Helping in Infancy." *Psychological Science* 21, no. 4: 523–527. https://doi.org/10.1177/0956797610364119.

Dunst, C. J., E. Gorman, and D. W. Hamby. 2012. "Preference for Infant-Directed Speech in Preverbal Young Children." *Center for Early Literacy Learning* 5, no. 1: 1–13.

Duval, S., and R. Tweedie. 2000. "A Nonparametric "Trim and Fill" Method of Accounting for Publication Bias in Meta-Analysis." *Journal of*

the American Statistical Association 95, no. 449: 89–98. https://doi.org/10.1080/01621459.2000.10473905.

Fehr, E., and U. Fischbacher. 2003. "The Nature of Human Altruism." *Nature* 425, no. 6960: 785–791. https://doi.org/10.1038/nature02043.

Frank, M. C., E. Bergelson, C. Bergmann, et al. 2017. "A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-building." *Infancy* 22, no. 4: 421–435. https://doi.org/10.1111/infa.12182.

Freud, S., J. Strachey, A. Freud, A. Strachey, and A. Tyson. 1961. *The Ego and the Id: The Standard Edition of the Complete Psychological Works of Sigmund Freud, Volume XIX (1923-1925): The Ego and the ID and Other Works*. London: The Hogarth Press: The Institute of Psycho-Analysis.

Gelman, A., D. Lee, and J. Guo. 2015. "Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization." *Journal of Educational and Behavioral Statistics* 40, no. 5: 530–543.

Geraci, A., and L. Franchin. 2021. "Do Toddlers Prefer That Agents Help Similar or Dissimilar Needy Agents?" *Infant and Child Development* 30, no. 5: e2247. https://doi.org/10.1002/icd.2247.

Geraci, A., L. Franchin, and S. Benavides-Varela. 2023. "Evaluations of Pro-Environmental Behaviors by 7-Month-Old Infants." *Infant Behavior and Development* 72: 101865. https://doi.org/10.1016/j.infbeh.2023.101865.

Geraci, A., F. Simion, and L. Surian. 2022. "Infants' Intention-Based Evaluations of Distributive Actions." *Journal of Experimental Child Psychology* 220: 105429. https://doi.org/10.1016/j.jecp.2022.105429.

Geraci, A., and L. Surian. 2011. "The Developmental Roots of Fairness: Infants' Reactions to Equal and Unequal Distributions of Resources." *Developmental Science* 14, no. 5: 1012–1020. https://doi.org/10.1111/j.1467-7687.2011.01048.x.

Geraci, A., and L. Surian. 2023. "Intention-Based Evaluations of Distributive Actions by 4-Month-Olds." *Infant Behavior and Development* 70: 101797. https://doi.org/10.1016/j.infbeh.2022.101797.

Gronau, Q. F., H. Singmann, and E.-J. Wagenmakers. 2017. "Bridgesampling: An R Package for Estimating Normalizing Constants." *Journal of Statistical Software*. Published ahead of print, October 23, 2017. https://doi.org/10.31222/osf.io/v94h6.

Hamlin, J. K. 2013. "Failed Attempts to Help and Harm: Intention Versus Outcome in Preverbal Infants' Social Evaluations." *Cognition* 128, no. 3: 451–474. https://doi.org/10.1016/j.cognition.2013.04.004.

Hamlin, J. K. 2015. "The Case for Social Evaluation in Preverbal Infants: Gazing Toward One's Goal Drives Infants' Preferences for Helpers Over Hinderers in the Hill Paradigm." *Frontiers* 5: 1563. https://doi.org/10.3389/fpsyg.2014.01563.

Hamlin, J. K., E. V. Hallinan, and A. L. Woodward. 2008. "Do As I Do: 7-Month-Old Infants Selectively Reproduce Others' Goals." *Developmental Science* 11, no. 4: 487–494. https://doi.org/10.1111/j.1467-7687.2008.00694.x.

Hamlin, J. K., T. Ullman, J. Tenenbaum, N. Goodman, and C. Baker. 2013. "The Mentalistic Basis of Core Social Cognition: Experiments in Preverbal Infants and a Computational Model." *Developmental Science* 16, no. 2: 209–226. https://doi.org/10.1111/desc.12017.

Hamlin, J. K., and K. Wynn. 2011. "Young Infants Prefer Prosocial to Antisocial Others." *Cognitive Development* 26, no. 1: 30–39. https://doi.org/10.1016/j.cogdev.2010.09.001.

Hamlin, J. K., K. Wynn, and P. Bloom. 2007. "Social Evaluation by Preverbal Infants." *Nature* 450, no. 7169: 557–559. https://doi.org/10.1038/nature06288.

Hamlin, J. K., K. Wynn, and P. Bloom. 2010. "Three-Month-Olds Show a Negativity Bias in Their Social Evaluations." *Developmental Science* 13, no. 6: 923–929. https://doi.org/10.1111/j.1467-7687.2010.00951.x.

Hamlin, J. K., K. Wynn, P. Bloom, and N. Mahajan. 2011. "How Infants and Toddlers React to Antisocial Others." *Proceedings of the National Academy of Sciences of the United States of America* 108, no. 50: 19931–19936. https://doi.org/10.1073/pnas.1110306108.

Henrich, J., S. J. Heine, and A. Norenzayan. 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33, no. 2–3: 61–83. https://doi.org/10.1017/S0140525X0999152X.

Holvoet, C., C. Scola, T. Arciszewski, and D. Picard. 2016. "Infants' Preference for Prosocial Behaviors: A Literature Review." *Infant Behavior & Development* 45: 125–139. https://doi.org/10.1016/j.infbeh.2016.10.008.

Ioannidis, John P. A. 2018. "Why Replication Has More Scientific Value Than Original Discovery." *Behavioral and Brain Sciences* 41: e137. https://doi.org/10.1017/s0140525X18000729.

Judd, C. M., J. Westfall, and D. A. Kenny. 2017. "Experiments With More Than One Random Factor: Designs, Analytic Models, and Statistical Power." *Annual Review of Psychology* 68: 601–625. https://doi.org/10.1146/annurev-psych-122414-033702.

Kanakogi, Y., Y. Inoue, G. Matsuda, D. Butler, K. Hiraki, and M. Myowa-Yamakoshi. 2017. "Preverbal Infants Affirm Third-Party Interventions That Protect Victims From Aggressors." *Nature Human Behaviour* 1, no. 2: 1–7. https://doi.org/10.1038/s41562-016-0037.

Kanakogi, Y., M. Miyazaki, H. Takahashi, H. Yamamoto, T. Kobayashi, and K. Hiraki. 2022. "Third-Party Punishment by Preverbal Infants." *Nature Human Behaviour* 6, no. 9: 1234–1242. https://doi.org/10.1038/s41562-022-01354-2.

Kanbe, F. 2019. "Generating Strictly Controlled Stimuli for Figure Recognition Experiments." *Journal of Visualized Experiments*. https://app.jove.com/t/59149/generating-strictly-controlled-stimuli-for-figure-recognition.

Kärtner, J., H. Keller, and N. Chaudhary. 2010. "Cognitive and Social Influences on Early Prosocial Behavior in Two Sociocultural Contexts." *Developmental Psychology* 46, no. 4: 905–914. https://doi.org/10.1037/a0019718.

Kass, R. E., and A. E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90, no. 430: 773–795. https://doi.org/10.2307/2291091.

Klein, R. A., K. A. Ratliff, M. Vianello, et al. 2014. "Data From Investigating Variation in Replicability: A "Many Labs" Replication Project." *Journal of Open Psychology Data* 2, no. 1: e4. http://doi.org/10.5334/jopd.ad.

Kohlberg, L. 1969. *Stages in the Development of Moral Thought and Action*. New York, NY: Holt, Rinehart & Winston.

Kominsky, J. F. 2019. "PyHab: Open-Source Real Time Infant Gaze Coding and Stimulus Presentation Software." *Infant Behavior & Development* 54: 114–119. https://doi.org/10.1016/j.infbeh.2018.11.006.

Kominsky, J. F., K. Lucca, A. J. Thomas, M. C. Frank, and J. K. Hamlin. 2022. "Simplicity and Validity in Infant Research." *Cognitive Development* 63: 101213. https://doi.org/10.1016/j.cogdev.2022.101213.

Kruschke, J. K., and T. M. Liddell. 2018. "The Bayesian New Statistics: Hypothesis Testing, Estimation, Meta-Analysis, and Planning From a Bayesian Perspective." *Psychonomic Bulletin and Review* 25, no. 1: 178–206. https://doi.org/10.3758/s13423-016-1221-4.

Kvarven, A., E. Strømland, and M. Johannesson. 2020. "Comparing Meta-Analyses and Preregistered Multiple-Laboratory Replication Projects." *Nature Human Behaviour* 4, no. 4: 423–434. https://doi.org/10.1038/s41562-019-0787-z.

Lewis, M., M. B. Mathur, T. J. VanderWeele, and M. C. Frank. 2022. "The Puzzling Relationship Between Multi-Laboratory Replications and Meta-Analyses of the Published Literature." *Royal Society Open Science* 9, no. 2: 211499. https://doi.org/10.1098/rsos.211499.

Liu, S., N. B. Brooks, and E. S. Spelke. 2019. "Origins of the Concepts Cause, Cost, and Goal in Prereaching Infants." *Proceedings of the National Academy of Sciences* 116, no. 36: 17747–17752. https://doi.org/10.1073/pnas.1904410116.

Lobue, V., K. Pérez-Edgar, N. Kirkham, and J. Herbert. 2023. "The Impact of COVID-19 on Infant Development: A Special Issue of Infancy." *Infancy* 28, no. 1:. https://doi.org/10.1111/infa.12528.

Loheide-Niesmann, L., J. de Lijster, R. Hall, H. van Bakel, and M. Cima. 2020. "Toddlers' Preference for Prosocial Versus Antisocial Agents: No Associations With Empathy or Attachment Security." *Social Development* 30, no. 2: 410–427. https://doi.org/10.1111/sode.12487.

Loheide-Niesmann, L., J. de Lijster, R. Hall, H. van Bakel, and M. Cima. 2021. "Toddlers' Preference for Prosocial Versus Antisocial Agents: No Associations With Empathy or Attachment Security." *Social Development* 30, no. 2: 410–427.

Lovric, M. M. 2019. "Conflicts in Bayesian Statistics Between Inference Based on Credible Intervals and Bayes Factors." *Journal of Modern Applied Statistical Methods* 18.

Lucca, K., J. Pospisil, and J. A. Sommerville. 2018. "Fairness Informs Social Decision Making in Infancy." *PLoS ONE* 13, no. 2: e0192848. https://doi.org/10.1371/journal.pone.0192848.

Makel, M. C., J. A. Plucker, and B. Hegarty. 2012. "Replications in Psychology Research: How Often do They Really Occur?" *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 7, no. 6: 537–542. https://doi.org/10.1177/1745691612460688.

Margoni, F., K. Block, K. Hamlin, N. Zmyj, and T. Schmader. 2023. "Meta-Analytic Evidence Against sex Differences in Infants' and Toddlers' Preference for Prosocial Agents." *Developmental Psychology* 59, no. 2: 229–235. https://doi.org/10.1037/dev0001421.

The ManyBabies Consortium. 2020. "Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference." *Advances in Methods and Practices in Psychological Science* 3, no. 1: 24–52.

Margoni, F., and M. Shepperd. 2020. "Changing the Logic of Replication: a Case From Infant Studies." *Infant Behavior and Development* 61: 101483. https://doi.org/10.1016/j.infbeh.2020.101483.

Margoni, F., and L. Surian. 2018. "Infants' Evaluation of Prosocial and Antisocial Agents: A Meta-Analysis." *Developmental Psychology* 54, no. 8: 1445–1455. https://doi.org/10.1037/dev0000538.

Marquardt, D. W. 1980. "Comment: You Should Standardize the Predictor Variables in Your Regression Models." *Journal of the American Statistical Association* 75, no. 369: 87–91. https://doi.org/10.1080/01621459.1980.10477430.

Meristo, M., and H. Zeidler. 2022. "Cross-Cultural Differences in Early Expectations About Third Party Resource Distribution." *Scientific Reports* 12, no. 1: 11627. https://doi.org/10.1038/s41598-022-15766-7.

Morey, R. D., R. Hoekstra, J. N. Rouder, M. D. Lee, and E. J. Wagenmakers. 2016. "The Fallacy of Placing Confidence in Confidence Intervals." *Psychonomic Bulletin & Review* 23, no. 1: 103–123. https://doi.org/10.3758/s13423-015-0947-8.

Nazzari, S., M. P. Pili, Y. Günay, and L. Provenzi. 2024. "Pandemic Babies: A Systematic Review of the Association Between Maternal Pandemic-Related Stress During Pregnancy and Infant Development." *Neuroscience & Biobehavioral Reviews* 162: 105723. https://doi.org/10.1016/j.neubiorev.2024.105723.

Nielsen, M., D. Haun, J. Kärtner, and C. H. Legare. 2017. "The Persistent Sampling Bias in Developmental Psychology: A Call to Action." *Journal of Experimental Child Psychology* 162: 31–38. https://doi.org/10.1016/j.jecp.2017.04.017.

Nighbor, T., C. Kohn, M. Normand, and H. Schlinger. 2017. "Stability of Infants' preference for Prosocial Others: Implications for Research Based on Single-Choice Paradigms." *PLoS ONE* 12, no. 6: e0178818. https://doi.org/10.1371/journal.pone.0178818.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349, no. 6251: aac4716. https://doi.org/10.1126/science.aac4716.

Peirce, J., J. R. Gray, S. Simpson, et al. 2019. "PsychoPy2: Experiments in Behavior Made Easy." *Behavior Research Methods* 51, no. 1: 195–203. https://doi.org/10.3758/s13428-018-01193-y.

Piaget, J. 1932. *The Moral Judgment of the Child*. London, England: Kegan Paul, Trench, Trubner.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News* 6, no. 1: 7–11.

Rogoff, B., A. Dahl, and M. Callanan. 2018. "The Importance of Understanding Children's Lived Experience." *Developmental Review* 50: 5–15. https://doi.org/10.1016/j.dr.2018.05.006.

Salvadori, E., T. Blazsekova, A. Volein, et al. 2015. "Probing the Strength of Infants' Preference for Helpers Over Hinderers: Two Replication Attempts of Hamlin and Wynn (2011)." *PLoS ONE* 10, no. 11: e0140570. https://doi.org/10.1371/journal.pone.0140570.

Scarf, D., K. Imuta, M. Colombo, and H. Hayne. 2012a. "Social Evaluation or Simple Association? Simple Associations May Explain Moral Reasoning in Infants." *PLoS ONE* 7, no. 8: e42698. https://doi.org/10.1371/journal.pone.0042698.

Scarf, D., K. Imuta, M. Colombo, and H. Hayne. 2012b. "Golden Rule or Valence Matching? Methodological Problems in Hamlin et al." *Proceedings of the National Academy of Sciences* 109, no. 22: E1426. https://doi.org/10.1073/pnas.1204123109.

Schlingloff, L., G. Csibra, and D. Tatone. 2020. "Do 15-Month-Old Infants Prefer Helpers? A Replication of Hamlin et al. (2007)." *Royal Society Open Science* 7, no. 4: 191795. https://doi.org/10.1098/rsos.191795.

Scola, C., C. Holvoet, T. Arciszewski, and D. Picard. 2015. "Further Evidence for Infants' Preference for Prosocial Over Antisocial Behaviors." *Infancy* 20, no. 6: 684–692. https://doi.org/10.1111/infa.12095.

Shimizu, Y., S. Senzaki, and J. S. Uleman. 2018. "The Influence of Maternal Socialization on Infants' Social Evaluation in Two Cultures." *Infancy* 23, no. 5: 748–766. https://doi.org/10.1111/infa.12240.

Singh, L. 2020. "Bilingual Infants are More Sensitive to Morally Relevant Social Behavior Than Monolingual Infants." *Journal of Cognition and Development* 21, no. 5: 631–650.

Stanley, D., and J. Spence. 2014. "Expectations for Replications: Are Yours Realistic?" *Perspectives on Psychological Science* 9, no. 3: 305–318. https://doi.org/10.1177/1745691614528518.

Strid, K., and M. Meristo. 2020. "Infants Consider the Distributor's Intentions in Resource Allocation." *Frontiers in Psychology* 11: 596213. https://doi.org/10.3389/fpsyg.2020.596213.

Stroup, W. W. 2012. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Boca Raton, Florida: CRC Press.

Tan, A. W. M. 2024. Prosocial Evaluation: A New Meta-Analysis. RPubs. osf.io/preprints/psyarxiv/qnh92. https://doi.org/10.31234/osf.io/qnh92.

Tan, E., and J. K. Hamlin. 2022. "Mechanisms of Social Evaluation in Infancy: A Preregistered Exploration of Infants' Eye-Movement and Pupillary Responses to Prosocial and Antisocial Events." *Infancy* 27, no. 2: 255–276. https://doi.org/10.1111/infa.12447.

Valenza, E., F. Simion, V. M. Cassia, and C. Umiltà. 1996. "Face Preference at Birth." *Journal of Experimental Psychology: Human Perception and Performance* 22, no. 4: 892–903. https://doi.org/10.1037/0096-1523.22.4.892.

Vaporova, E., and N. Zmyj. 2020. "Social Evaluation and Imitation of Prosocial and Antisocial Agents in Infants, Children, and Adults." *PLoS ONE* 15, no. 9: e0235595. https://doi.org/10.1371/journal.pone.0235595.

Woo, B. M., and E. S. Spelke. 2020. "How to Help Best: Infants' changing Understanding of Multistep Actions Informs Their Evaluations of Helping." In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. COGSCI, 384–390.

Woo, B. M., and E. S. Spelke. 2023. "Toddlers' Social Evaluations of Agents Who Act on False Beliefs." *Developmental Science* 26, no. 2: e13314. https://doi.org/10.1111/desc.13314.

Woo, B. M., C. M. Steckler, D. T. Le, and J. K. Hamlin. 2017. "Social Evaluation of Intentional, Truly Accidental, and Negligently Accidental Helpers and Harmers by 10-month-old Infants." *Cognition* 168: 154–163. https://doi.org/10.1016/j.cognition.2017.06.029.

Woo, B. M., E. Tan, and J. K. Hamlin. 2022. "Human Morality Is Based on an Early-emerging Moral Core." *Annual Review of Developmental Psychology* 4: 41–61. https://doi.org/10.1146/annurev-devpsych-121020-023312.

Woodward, A. L. 1998. "Infants Selectively Encode the Goal Object of an Actor's Reach." *Cognition* 69, no. 1: 1–34.

Zettersten, M., C. Cox, C. Bergmann, et al. 2024. "Evidence for Infant-Directed Speech Preference Is Consistent Across Large-Scale, Multi-Site Replication and Meta-Analysis." *Open Mind: Discoveries in Cognitive Science* 8: 439–461. Advance publication. https://doi.org/10.1162/opmi_a_00134.

Zwaan, R. A., A. Etz, R. E. Lucas, and M. B. Donnellan. 2018. "Making Replication Mainstream." *Behavioral and Brain Sciences* 41: E120. https://doi.org/10.1017/S0140525X17001972.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.