AMERICAN PSYCHOLOGICAL ASSOCIATION

Developmental Psychology

ISSN: 0012-1649

2022, Vol. 58, No. 11, 2022–2035
https://doi.org/10.1037/dev0001411

False Belief Understanding and Moral Judgment in Young Children

Karlena D. Ochoa[1], Joseph F. Rodini[1], and Louis J. Moses[2]
[1] Department of Psychology, University of Oregon
[2] School of Psychology, Victoria University of Wellington

Although the influence of intent understanding on children's moral development has been long studied, little research has examined the influence of belief understanding on that development. In two studies we presented children with morally relevant belief vignettes to examine the extent to which they incorporate both intent and belief information in their moral judgments. In Study 1 (N = 64), 5-year-olds with higher false belief understanding (FBU) rated agents with false beliefs as more positively intentioned in good intent trials (even though the outcome was bad) than in bad intent trials (even though the outcome was good). In contrast, 4-year-olds with higher FBU were generally unable to integrate their belief understanding with their moral evaluations, performing no better on intention questions than children with lower FBU. Neither age group significantly differentiated reward and punishments as a function of intent when a false belief was involved. In Study 2 (N = 109 children, N = 42 adults), we found that by simplifying our study design and reducing the task demands, 4-year-olds with higher FBU were able to make appropriate intent judgments. Yet, as in Study 1, all children had difficulty assigning punishment/reward based on intent. For both moral intentions and moral consequences, 4- and 5-year-olds with higher FBU differed from those of adults in several respects, indicating that moral reasoning develops substantially beyond the preschool years.

Keywords: development, moral judgments, social cognition, theory of mind

Supplemental materials: https://doi.org/10.1037/dev0001411.supp

Sometime during the preschool years, before formal schooling, children begin to display understanding of their own and others' mental states such as desires, beliefs, intentions, and emotions. This Theory of Mind (ToM), allows them to understand, explain, and predict behavior (Premack & Woodruff, 1978) and is critical for other domains of social functioning (Astington, 2003; Baron-Cohen et al., 1985; Leslie, 1988; Wellman, 2020). One such domain is that of moral judgment (Baird & Astington, 2005; Cushman et al., 2013; Killen et al., 2011; Lagattuta & Weller, 2014; Lane et al., 2010; Smetana et al., 2012; Wainryb & Ford, 1998). The current article addresses the interplay between ToM and moral judgment in young children.

Understanding intentions is one aspect of ToM that is central in moral judgment. In our everyday ethical reasoning, as well as in our legal codes, the intentions that underlie actions are critical in assessing the moral status of actors and whether or not their actions are deserving of reward or punishment. Yet there is a long history of research going back at least to Piaget (1932/1965), indicating that young children often rely more on outcomes than intentions when making moral judgments (Cushman et al., 2013; Killen et al., 2011; Zelazo et al., 1996). Nonetheless, in simple scenarios in which the outcome is unknown or otherwise held constant, children as young as three can make appropriate use of information about intent in making moral judgments (Smetana, 2006), and even infants appear to show some sensitivity to moral intent (Hamlin, 2013). That said, weighing intentions against outcomes in making moral evaluations and in assigning punishment can be complex even for adults and is only mastered slowly over development (Cushman, 2008; Cushman et al., 2013).

Understanding others' beliefs, including false beliefs, is also important for making moral judgments. In contrast to intentions, however, there has been very little research assessing how and when children integrate belief information into their moral judgments. Although some earlier work indirectly examined belief and knowledge understanding in relation to moral development (Wimmer et al., 1984; Yuill & Perner, 1988), it is only recently that a systematic analysis of the relation has been undertaken. Killen et al. (2011) assessed 3- to 8-year-old children's understanding in a moral transgression task embedded within a false belief story. In this story, a well-intentioned boy accidentally causes a negative outcome because he acts on the basis of a false belief about a container's contents. Specifically, as the boy was helping a teacher clear tables, he threw out a paper bag which,

This article was published Online First July 11, 2022.
Karlena D. Ochoa https://orcid.org/0000-0002-2283-4810
Joseph F. Rodini https://orcid.org/0000-0003-0042-2978
Louis J. Moses https://orcid.org/0000-0001-5588-5713
The authors confirm that there are no known conflicts of interest. Study 1 was not preregistered. Study 2 was preregistered and study materials and data can be found on OSF at https://osf.io/3p5m9/.
Correspondence concerning this article should be addressed to Karlena D. Ochoa, who is now at School of Education, University of California, Irvine, 401 East Peltason Drive, Irvine, CA 92617, United States. Email: karlenaochoa@gmail.com or kochoast@uci.edu

unbeknownst to him, contained another child's cupcake. Children then responded to a range of questions including those assessing their false belief understanding (i.e., what did the boy believe was in the bag?), their moral appraisal of the actor's intention (i.e., whether the boy thought he was doing something all right or not all right), and, in a second study, whether punishment was warranted (i.e., whether the boy should get in trouble). Killen et al. (2011) found that children without false belief understanding were more likely to attribute negative intentions and to assign punishment to the accidental transgressor than children with false belief understanding. Moreover, it was not until children were 7 or 8 that they attributed positive intentions to the boy at high levels.

This research establishes a link between belief understanding and moral judgment in childhood but leaves many questions unanswered. First, Killen et al. (2011) administered only one morally relevant belief story with only one combination of agent's belief (false) and intention (good). It remains to be seen whether there are developmental changes across other combinations of belief (true vs. false) and intention (good vs. bad), and whether belief and intention interact in some way. Children may have had similar difficulty even in a true belief context and may have responded differently had the intentions of the actor been negative. Research by Cushman et al. (2013) partially addresses this issue. Their research contrasted accidental harm (benign intent, bad outcome) with attempted harm (negative intent, good/neutral outcome), finding developmental shifts across 4 to 8 years of age. Children were initially outcome-focused in their evaluations of moral wrongness, then incorporated intent into those evaluations, but only later began to weigh intent in judgments of punishability. Nonetheless, Cushman's research did not address the integration of belief understanding with moral evaluation, the main focus of the current studies.

Second, in the Killen et al. (2011) study, to perform well children were required to make assumptions about the agent's intention (i.e., that the boy cleaning tables would not have thrown out the cupcake had he known it was in the bag). Hence, the younger children may have performed poorly either because they lacked false belief understanding or because they did not hold a default assumption of benign intent. They might show greater ability to incorporate belief understanding with moral evaluations if agents' intentions are made more explicit. In Cushman et al. (2013), the protagonists' intentions were explicit but, again, that study did not address the integration of belief understanding with moral reasoning.

Third, when children initially attain false belief understanding, they may still have difficulty applying that knowledge to moral judgments. There may be a lag such that their ability to integrate their newly acquired belief understanding with their moral understanding is delayed. Interestingly, a similar lag has been found with respect to integrating belief information with emotion understanding. De Rosnay et al. (2004) told children stories in which, for example, a character mistakenly believed that a container held a preferred food when in actuality it contained a disliked food. Although 4- and 5-year-olds understood the character's mistaken belief, they nonetheless incorrectly predicted that the character would feel sad on seeing the container. It was not until age 6 that children made correct emotion predictions (see also, Wellman & Liu, 2004). This finding, that children cannot immediately use belief information to inform judgments about another mental state (emotion), demonstrates difficulty in integrating concepts within the ToM domain. It remains to be seen whether similar difficulty, and a similar developmental lag, would be found in integrating belief reasoning with understanding in a different domain, that of moral judgment. Preliminary evidence for such a lag comes from the Killen et al. (2011) study. They found that while 7-year-olds rated the accidental transgressor as having positive intentions, 5-year-olds (most of whom correctly answered the false belief question) gave a neutral rating of the agent's intention. Whether that finding is replicable for positive intentions and extendable to negative intentions is not known.

Finally, a lag may also be present between children's appreciation of moral intentions and their accurate assignment of consequences to actors. Although Killen et al. (2011) questioned children regarding both moral intentions and moral consequences, their data are not broken down in such a way as to clearly determine whether the latter were more difficult to appreciate than the former. However, as noted earlier, Cushman et al. (2013) found that around age 5 children first incorporate intent information into judgments of moral wrongness and only later, between 6 and 8 years, into intent-based punishment judgments. They argue that this pattern is not driven by changes in relation to ToM but instead represents the reorganization of concepts within the moral domain itself, specifically in how intent relates to wrongness and punishment. We return to their findings in the general discussion but, for now, note that we do not yet know how belief understanding might interact with moral judgments of wrongness and punishability.

The current research was designed to address these issues. In two studies we manipulated agents' intentions and beliefs in morally relevant vignettes to examine preschoolers' ability to integrate belief understanding with moral judgment. In both studies we assessed children's understanding of beliefs, moral intentions, and moral consequences across the age range (4–5 years) in which children are acquiring false belief understanding. Doing so allowed us to determine whether developmental lags are present for integrating information across the domains of ToM and moral judgment, whether they are present for reasoning about both positive and negative moral intentions, and whether they are present for reasoning about both true and false beliefs.

Study 1

In Study 1, we manipulated agents' intentions (positive or negative) and beliefs (true or false) in vignettes conceptually comparable with those used in Killen et al. (2011). Each vignette featured an agent intending to deliver a benefit (positive intent) or a harm (negative intent) to another character. When the agent held a true belief, the harm or benefit was successfully delivered, whereas when the agent held a false belief the opposite outcome occurred (e.g., delivering benefit when harm was intended and vice versa). Children were then questioned about agents' beliefs and intentions and asked to assign reward or punishment.

For comparison we also assessed children's false belief understanding in a standard nonmoral context, and their moral understanding in a standard moral judgment task. To be clear, however, our main focus was on false belief understanding in the moral context not the standard context. From a theoretical standpoint, we were primarily interested in children who make correct false belief inferences in the very same context in which they are asked to

make moral evaluations. Only then could we say that children recognized the false belief but were unable to (or able to) to make use of it in moral reasoning. Put another way, we were not interested in whether a lag might be present between false belief understanding in general and moral evaluation but rather whether the lag would be present in moral contexts specifically.

For true belief trials we hypothesized that children would rate agents in the good intent condition as morally superior to those in the bad intent condition and would appropriately recommend reward or punishment. In contrast, we hypothesized that the intersection of intent and false belief would lead to differential response patterns across the sample concordant with the attainment of false belief understanding. We expected that children without false belief understanding would perform poorly, often responding in terms of outcome when assessing intent and assigning reward or punishment. For children with false belief understanding, we had two contrasting hypotheses. If children immediately integrate false belief understanding with moral evaluation they would use this information to correctly infer agents' intentions. In contrast, if there is an integration delay, the younger children might perform poorly, perhaps defaulting to outcome-based responding. Finally, based on prior research (Cushman et al., 2013), we hypothesized a further developmental lag between appropriately assigning intent and appropriately assigning consequences.

Method

Participants

Sixty-four children participated, 32 four-year-olds (13 girls; *M* age= 53.56 months, *SD* = 4.00) and 32 five-year-olds (15 girls; *M* age = 66.91, *SD* = 3.54). An additional 11 children were tested but excluded from analyses owing to experimenter error (two), refusal to cooperate/inattention (four), and for incorrectly answering all reality check questions for the four morally relevant belief vignettes (five). Although we did not conduct an a priori power analysis (this first study was conducted in 2014 at a time we were not yet conducting preregistrations), a post hoc sensitivity analysis revealed that we were well-powered to detect medium to large (f = .35) main effects and interaction effects. See the online Supplemental Materials (S1) for more details on our power analyses for both studies.

Children were typically developing and were recruited from a participant database at a large research university in the Western United States. The sample was representative of the population from which it was drawn. Forty-nine children were White, six were multiracial, five were Latinx, two were Native American/Native Alaskan, one child was African American, and one child was Asian/Pacific Islander. Parents were provided compensation in the form of $10, and children received a small toy. The research protocols for this and the following study were approved by the University of Oregon's Institutional Review Board under the project name "The Development of Moral Evaluations in Children and Adults" (Protocol Number: 05242018.032).

Procedure

Children were tested individually in a single 45-min videotaped session. This and the following study were part of a larger project examining theory of mind, executive functioning, and moral development. Here we present findings on ToM and moral development. Children received tasks in two counterbalanced testing blocks: a morally relevant belief vignette block, and a standard moral transgression and false belief block. In the first block, vignettes were presented in a counterbalanced order, as were the tasks in the second block.

Measures

Morally Relevant Belief Vignettes. Children in each age group were randomly assigned to one of two conditions: good intent (n = 32; 13 girls; *M* age = 60.34 months, *SD* = 8.24) or bad intent (n = 32; 15 girls; *M* age = 60.13 months, *SD* = 7.24). Within conditions, belief (true or false) was manipulated across four vignettes (see Table 1). The vignettes all featured two characters, one of whom (the "agent") discovered a pleasant object in an opaque container and an unpleasant object in a different opaque container. The agent then decided to share one of these containers with another character (the "recipient") after the objects switched containers. The sharing of objects in these stories served as instances of harming or helping behavior. Two of the stories involved true belief (TB) and two involved false belief (FB) on the part of the agent (i.e., for two stories the agent knew the contents of the container and for two stories the agent had a mistaken belief about the contents of the container). The order of story presentation was counterbalanced across children and fully decoupled with respect to belief and intent (e.g., the kitten and skunk story served as a bad intent, TB story for some children and a good intent, FB story for other children). A depiction of the experimental manipulations and a vignette script can be found on the Open Science Framework (OSF; https://osf.io/3p5m9/; Ochoa et al., 2020). As an example, in the kitten and skunk story, Bobby (the agent) discovered a bucket containing a kitten and a bin containing a skunk. After returning the animals to their containers, he explicitly stated his intent to make a second, offstage character, Jacob (the recipient), either happy or upset via sharing the container with the kitten or

Table 1
Proportion of Justification Types for Punishment/Reward by False Belief Group

False belief group	*N* of justifications	Content of justification		
		Mental state	Outcome	Undifferentiated
Children with lower FBU	100	.39	.15	.46
Children with higher FBU	122	.39	.22	.39
Adults with higher FBU	80	.88	0	.12

Note. FBU = false belief understanding.

skunk, respectively (in contrast to the Killen et al. (2011) study in which the intention needed to be assumed). The agent then left the scene, at which point the belief manipulation took place, with the agent either immediately returning to witness the animals switch location (TB trials) or returning only after the animals switched locations (FB trials). At this point, the recipient entered the scene, prompting the agent to offer either what he thought was a "good" container (the one containing the kitten) or what he thought was a "bad" container (the one containing the skunk) to the recipient. The other stories also featured positively and negatively valenced items: a butterfly and a spider, a cupcake and a moldy tomato, and a cookie and a rotten apple, respectively. In stories with valenced animals, the animals switched positions themselves, while in stories with valenced food items, a new animal appeared from offstage to switch the objects and left before the recipient arrived (a pet parrot and a field mouse, respectively).

Children were asked three comprehension questions over the course of the story: two of these concerned whether each of the items would make the recipient feel good or bad (e.g., "Do kittens make Jacob feel good or bad?"), and one concerned whether the agent was present or absent for the switch (e.g., "Was Bobby there to see the animals switch?").

After each vignette, children responded to seven questions in a fixed order: (a) intent evaluation 1 ("When [AGENT] handed [RECIPIENT] the container, was [AGENT] trying to be a good boy, bad boy, or just okay?" (b) agent belief ("What does [AGENT] think is in the container?"); (c) reality check ("What is really in the container?"); (d) intent evaluation 2 ("Do you think [AGENT] is being mean, nice, or just okay?"); (e) consequence evaluation ("Should [AGENT] get in trouble, like a timeout, a treat, like a trip to the zoo, or nothing? If [AGENT] gets nothing, he will not get a timeout or a trip."); (f) friend evaluation ("Do you want to be friends with [AGENT]?"); and (g) recipient emotion attribution ("How will [RECIPIENT] feel when he opens the container?"). The consequence assignment question featured depictions of the three consequences to which the child could point. Children were credited with FB understanding only if they correctly answered the agent belief and reality check question.

Perhaps not surprisingly, the two intent evaluation questions produced very similar results: Both questions assess whether children rely on intent in assessing what the agent was trying to do and in evaluating the moral status of the agent. For brevity, we only report on the intent evaluation 2 question (referred to as intent evaluation henceforward), but results for the intent evaluation 1 question can be found in the online Supplemental Materials (S2). The friend evaluation question appeared to be strongly affected by a yes response bias (many children wanted to be friends with all agents), and therefore is not analyzed further here.

Standard Moral Transgression Task. The moral transgression task consisted of a single, prototypic moral transgression story often used in the literature (see Smetana, 2006) that involved a character pushing someone off a swing. Children were presented with pictures of a swing set and two characters who matched their own gender. They were told a story in which one character pushes another off the swing, causing the second character to fall and hurt his or her knee. They then responded to a subset of questions from the morally relevant belief vignettes: intent evaluation 2, consequence evaluation, friend evaluation, and recipient emotion attribution.

Standard False Belief Task. The standard FB task (adapted from Leslie et al., 2005) had a broadly similar structure to the moral vignettes and featured a boy (Jamie), two doghouses, and two dogs—one with spots and one without. After discovering the dogs, the boy was described as wanting to give a bone to the dog with spots. The boy then left to get the bone and, while he was away, the dogs switched places. The boy then returned with the bone and approached the doghouse in which the spotted dog was originally located. Children were then asked which dog the boy thinks is in the doghouse and which dog is really in it.

Results

All analyses were conducted in R (R Core Team, 2014) and figures were produced using the *ggplot2* package (Wickham, 2009). For our central analyses, we used the *clmm* function in the *ordinal* package to conduct mixed-effects logistic regressions (Christensen, 2019).

Standard Moral and False Belief Tasks

In replication of previous research, children evaluated the agent in the prototypical moral transgression task as mean (count: 61/64) rather than nice or just okay, Cramer's $V = 97.50$, $p < .001$, and deserving to be punished (47/64) rather than receiving nothing or a reward, Cramer's $V = 132.5$, $p < .001$. Four and five-year-olds did not differ significantly for either the intent evaluation or consequence ratings. Further, almost all children recognized that the victim would feel sad (56/64).

On the standard FB task, there was a marginally significant age effect with 5-year-olds somewhat more likely to exhibit FB understanding than 4-year-olds, $\chi^2(1, N = 59) = 3.37$, $p = .07$, ϕ=.24. Four-year-olds (62% correct) did not differ from chance, $\chi^2(1, N = 29) = 1.69$, $p > .05$, whereas 5-year-olds did (83%), $\chi^2(1, N = 30) = 13.33$, $p < .001$.

Comparison of Morally Relevant Vignettes With Standard Tasks

To assess the comparability of our moral vignettes and the standard moral transgression task, we compared responses across tasks to the agent intention and consequence questions from the half of the sample ($n = 32$) that were given the standard moral transgression task and the analogous TB, bad intent vignettes. Because there was only one standard task, we compared it with only the first of the TB, bad intent vignettes that children received. A Wilcoxon signed-ranks test revealed that children rated the agent as meaner in the standard story than in the morally relevant belief story, $W = 638.50$, $p = .003$, and as more deserving of punishment in the standard story than in the morally relevant belief vignette, $W = 692.50$, $p = .002$.

Children's performance on the standard FB task (70% correct) did not differ significantly from their performance on the belief question on the first morally relevant FB vignette to which they responded (73% correct), McNemar's test, $p = .58$. Moreover, children's responses to the belief questions on these tasks were highly correlated, $\phi = .65$, $p < .001$. As noted earlier, for theoretical reasons our main point of comparison in the analyses below is performance on the moral false belief task rather than on the standard false belief task. That said, children's performance on the two

task types was highly correlated such that using standard false as the comparator produced the same general pattern of results as found with moral false belief. Additional information can be found in the online Supplemental Materials (S3).

Morally Relevant Belief Vignettes

Because our data were ordinal, we conducted a series of mixed-effects logistic regressions on our main dependent variables of interest. For each analysis we included intent condition, FB group (or age group) as fixed factors, and the interaction between the two, and included participant ID as a random factor. In most cases the negatively valenced response (i.e., "mean," "trouble") and the children with lower false belief understanding (FBU; vs children with higher FBU) were used as the reference categories. If the interaction term was not significant, we dropped it and included only the main effects in the model. If the interaction term was significant, we followed-up with simple effects analyses, and reported the *p* value and odds ratio for each comparison.

No main effects of gender, order (block one vs. block two), or story (e.g., kitten/skunk vs. others) were found for any morally relevant belief vignette test items (*p*s $>$ .15), and so these factors were collapsed in subsequent analyses. All children in these analyses answered the reality check question correctly for at least one TB story and one FB story. Five children were excluded for failing all reality check questions.

We used the recipient emotion prediction question as a comprehension check. Consistent with earlier work with standard moral transgression tasks (Smetana, 2006), children appropriately judged that recipients in good intent TB trials would feel happy (28/32) and that those in bad intent TB trials would feel sad (28/32). Similarly, children judged that recipients would feel sad (31/32) in FB good intent trials and happy (24/32) in FB bad intent trials. Importantly, these results for the last question asked following each vignette suggest that children were not seriously hampered by memory demands across the task.

Agent Belief. We conducted a 2 (belief: true vs. false) $\times$ 2 (intent: good vs. bad) $\times$ 2 (age: 4- vs. Five-year-olds) mixed-effects logistic regression, on agent belief responses. None of the interaction terms were significant and so were excluded from a reduced final model. In the final model, all fixed effects were significant. and in the true (86% correct) compared with false belief condition (72% correct), $B = 1.02$, $SE = .35$, $z = 2.92$, $p = .003$, $OR = 2.77$, 95% CI [1.40, 5.50]. Further, whereas 4-year-olds were only above chance on TB trials (One-sample Wilcoxon Test, $p <$. 001), 5-year-olds were above chance for both TB and FB trials (*p*s $<$ .001).

The sample was then divided into three groups as a function of belief understanding: lower (FBU) – children who failed one or both of the belief questions on the morally relevant FB vignettes ($N = 21$; seven girls; M age = 56.38 months, $SD = 6.30$); 4-year-old higher FBU—children of that age who passed both belief questions ($N = 17$; eight girls; M age = 53.94 months, $SD = 4.18$); and 5-year-old higher FBU– children of that age who passed both belief questions ($N = 26$; 13 girls; M age = 67.46 months, $SD = 3.56$). This grouping factor best allows for the assessment of whether FBU is immediately coupled with advances in moral evaluation or whether there is a developmental lag in integrating these forms of reasoning. For the remaining analyses, we present results separately for TB and FB trials, with TB being reported first as a baseline from which to evaluate moral judgment in FB conditions.

Intent Evaluation. Figure 1 depicts children's responses to the intent evaluation measure. For TB trials, we conducted a 3 (false belief group) $\times$ 2 (intent) mixed-effects logistic regression on intent evaluation responses. The group by intent interaction was not significant and so was excluded from the final model. The main effect of intent condition was significant in the final model: Children appropriately rated agents in the good intent condition as nicer than those in the bad intent condition, $B = 5.93$, $SE = 1.19$, $z = 4.97$, $p < .001$, $OR = 376.39$, 95% CI [36.32, 3901.20].

For FB trials, there was a significant main effect of FB group ($B = -4.40$, $SE = 1.44$, $z = -3.06$, $p = .002$) that was qualified by a significant interaction between intent condition and FB group ($B = 7.23$, $SE = 2.12$, $z = 3.41$, $p < .001$). Follow-up simple effects tests revealed that whereas f5ive-year-olds with higher FBU ($p <$ .001, $OR = 135$) appropriately evaluated agents based on intent condition, 4-year-olds with higher FBU ($p = .13$, $OR = 4.23$) and children with lower FBU did not ($p = .65$, $OR = .61$).

Agent Consequence. Figure 2 depicts children's responses to the agent consequence measure. For TB trials, we conducted a 3 (false belief group) $\times$ 2 (intent) mixed-effects logistic regression on agent consequence. There were significant main effects of FB group and intent condition that were qualified by a significant interaction. Across FB groups, children rated agents in the good intent condition as more deserving of a treat compared with the bad intent condition, $B = 3.06$, $SE = .72$, $z = 4.25$, $p < .001$, $OR =$ 21.28, 95% CI [5.20, 87.06]. Simple effects analyses revealed in the good intent condition the groups did not significantly differ ($p > .31$). In contrast, in the bad intent condition, 5-year-olds with higher FBU were more likely than children with lower FBU ($p <$ .001, $OR = 1.08$) and 4-year-olds with higher FBU ($p < .001$, $OR = 3.20$) to assign trouble. Importantly, however, simple effects tests also revealed that all groups distinguished consequence judgments based on intent condition (5-year-olds with higher FBU, $p =$ .002, $OR = 668.12$, 4-year-olds with higher FBU, $p = .02$, $OR =$ 18.87, and children with lower FBU, $p = .01$, $OR =$103.30).

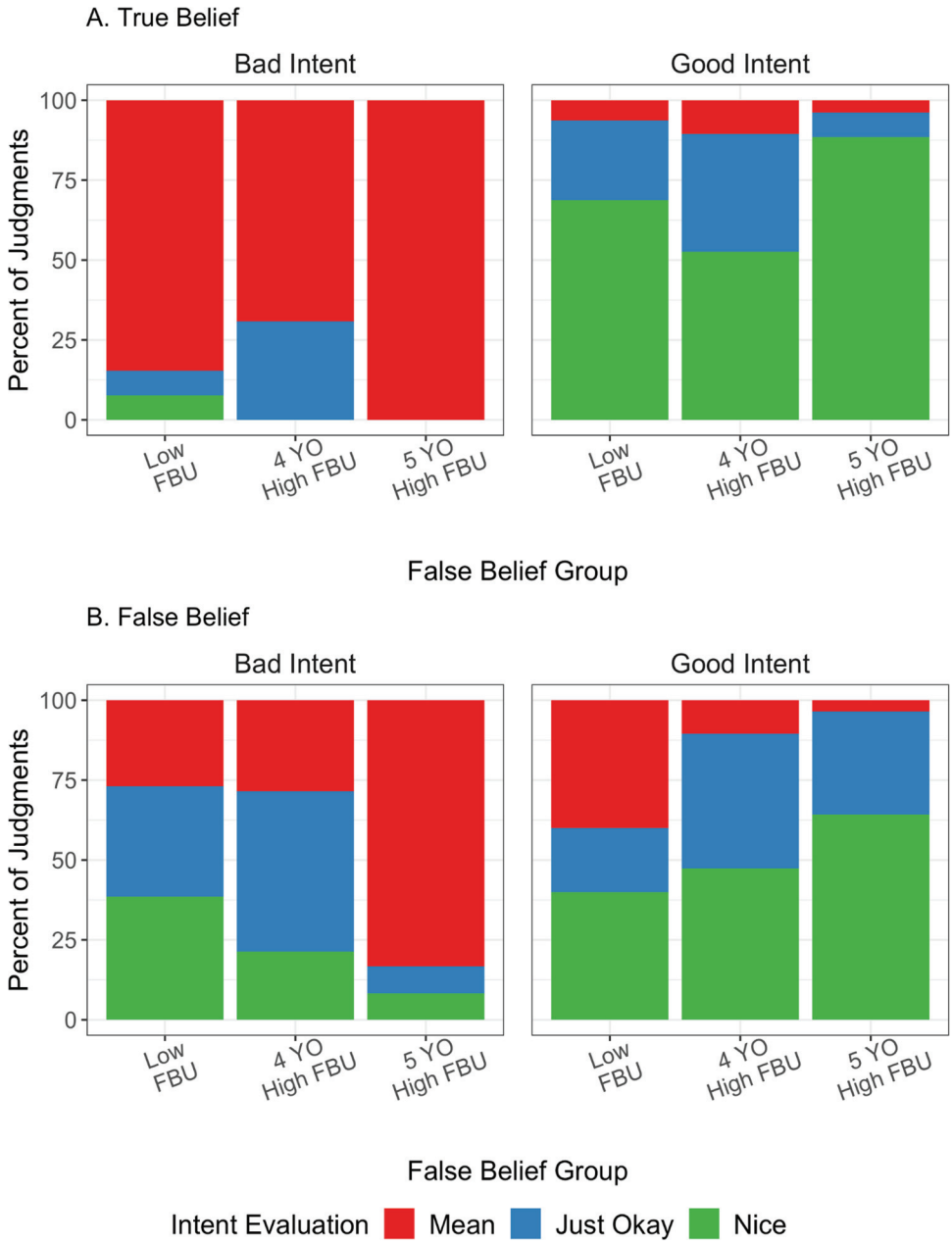
In contrast, in the FB condition there were no significant effects (see Figure 2B). Most importantly, children's consequence judgments did not differ significantly across intent conditions. Thus, although children were able to assign appropriate consequences in a TB context, doing so in a FB context was challenging for all groups.

We conducted a follow-up analysis in the FB condition, recoding children's responses as either correct or incorrect. Specifically, in the bad intent condition, assigning trouble is a clear correct response but assigning nothing could also be considered reasonable because there is no negative outcome. Assigning a treat in this condition is clearly incorrect. Conversely, in the good intent condition, assigning a treat is a clear correct response but assigning nothing could also be considered reasonable because there is no positive outcome. Assigning trouble in this condition is clearly incorrect. Children thus received a score of 0 or 1 based on their incorrect/correct responses for each trial.

A 3 (false belief group) $\times$ 2 (intent) mixed-effects logistic regression on incorrect/correct consequence judgments revealed a main effect of intent, such that children made more correct consequence judgments in the good intent condition compared with the bad intent condition ($B = 3.39$, $SE = 1.29$, $z = 2.63$, $p = .009$, $OR =$ 29.51, 95% CI [2.37, 367.51]. The main effect of false belief

Figure 1

Percent of Intent Evaluation Judgments Separated by False Belief Group and Intent Condition for the True Belief Condition (Top) and False Belief Condition (Bottom) in Study 1



Note. FBU = false belief understanding. See the online article for the color version of this figure.

group and the interaction were not significant. Children with lower FBU (43% both correct), 4-year-olds with higher FBU (29% both correct), and 5-year-olds with higher FBU (58% both correct) all had difficulty assigning consequences in the FB context, with no group performing better than chance, *ps* > .06.

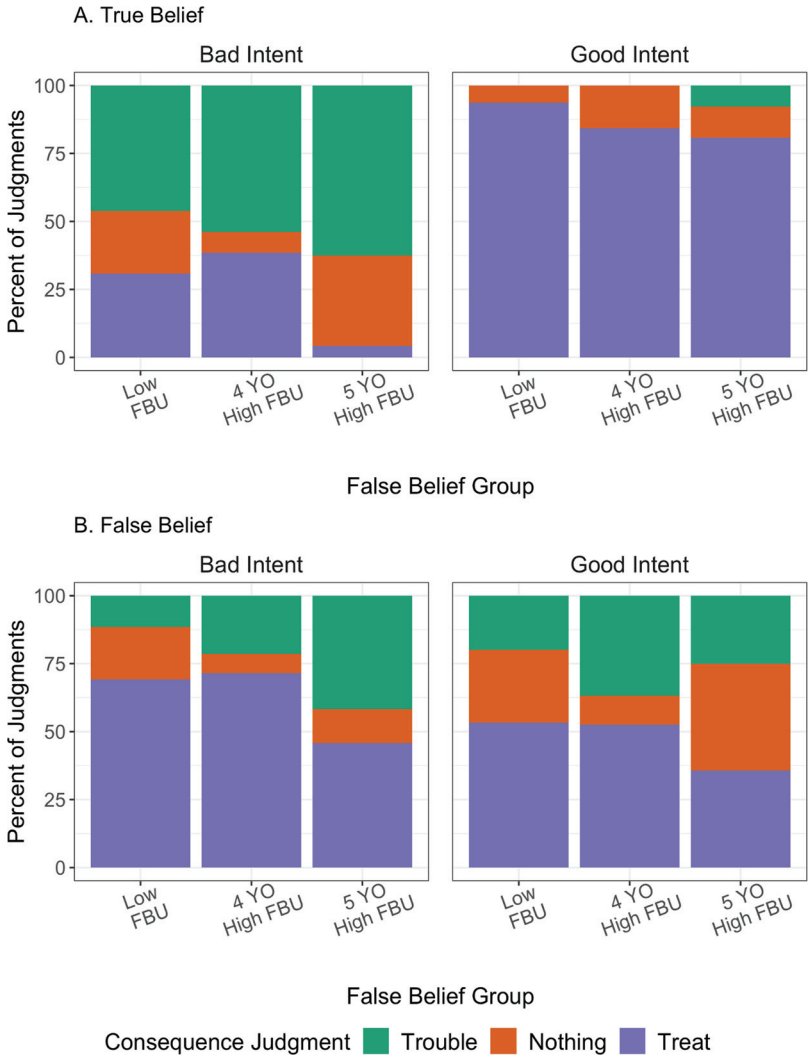
Discussion

By limiting the age groups in our sample to the window during which false belief understanding is acquired, we found intriguing interactions between ToM and moral judgments. The morally relevant belief vignettes served as dual assessments of false belief understanding and moral reasoning, and elicited responses from a sample of 4- and 5-year-olds demonstrating developmental changes in moral reasoning across the acquisition of false belief understanding.

We first evaluated whether our moral tasks were in relevant respects comparable in difficulty to a standard false belief and a standard moral transgression task. Performance on false belief questions in the moral vignettes was no more difficult than that on

Figure 2

Percent of Consequence Judgments Separated by False Belief Group and Intent Condition for the True Belief Condition (Top) and False Belief Condition (Bottom) in Study 1



Note. FBU = false belief understanding. See the online article for the color version of this figure.

the standard false belief task. In contrast, children performed somewhat less well on intention and consequence questions in the relevant moral vignette (true belief, bad intent) than on the same questions in the standard moral transgression task. Two possible factors may have generated this difference. First, our vignettes were informationally more complex than the standard task; second, the outcome was not as salient in our vignettes (an undesirable object hidden in a container) as in the standard task (a character falling down and hurting his or her knee).

We turn now to the central findings for the moral vignettes under different conditions. As expected, in true belief trials children generally performed well in responding to intent and consequence questions: They rated agents as meaner and more deserving of punishment in bad intent conditions compared with good intent conditions. Moreover, this pattern was found for children with lower FBU as well as those with higher FBU, which is of course not surprising as false belief understanding was not required for success on the true belief trials.

In contrast, on false belief trials performance on the intent questions interacted with false belief status. Confirming and extending previous research on good intent (Killen et al., 2011), we found that the acquisition of false belief understanding does indeed change children's patterns of moral evaluations in a false belief context, and does so for both good and bad intent.

Not surprisingly, children with lower FBU performed poorly on intent questions—assigning intent accurately depended on understanding that agents had false beliefs. Yet, children with lower FBU did not always default to assigning intent based on outcome; rather they seemed to be responding randomly (see Figure 1), a point to which we return in the General Discussion. More interestingly, a

developmental lag emerged between FBU and integrating that understanding with moral evaluations. Five-year-olds with higher FBU rated agents as more positively intentioned in good intent trials (even though the outcome was bad) than in bad intent trials (even though the outcome was good). In contrast, 4-year-olds with higher FBU were generally unable to integrate their belief understanding with their moral evaluations, performing no better on intent evaluation questions than false belief failers. Unlike children with lower FBU, this group had correctly predicted that the agent had a mistaken belief about the container they shared with the recipient but were unable to use this information in assessing the agent's moral intention. This finding is in accord with previous work (De Rosnay et al., 2004) in which children were not able to use false belief information to make false belief-influenced emotional predictions in a similarly aged sample. It is possible, however, that by reducing the information processing demands in the moral vignettes, 4-year-olds with higher FBU may be better able to use their belief knowledge to make moral judgments. We explore this possibility in Study 2.

As noted earlier, children performed quite well on the agent consequence questions in true belief trials: They assigned greater punishment to agents with ill intent who generated negative outcomes for recipients than to those with good intent who generated positive outcomes for recipients. In contrast, on false belief trials, even 5-year-olds with higher FBU performed no better than chance when assigning a consequence to the agent across levels of intent. Thus, as hypothesized, and consistent with Cushman et al. (2013), a further developmental lag emerged between incorporating belief understanding in intent judgments and doing so in consequence judgments. What is not clear from the present study is how long that lag lasts because even the oldest children struggled with consequence questions. We begin to address this issue in Study 2.

Study 2

The purpose of Study 2 was fourfold. First, we aimed to investigate whether the developmental lags found in Study 1 would replicate when the processing demands of the task and the length of the testing session were reduced. We simplified the design of the moral vignettes in two ways. We used only animals (and not food items) so that children would only have to direct their attention to the animals of interest rather than a third animal entering and moving the food items from one box to another. In addition, we used only a single box such that children would not have to track the movements of two animals across two separate boxes.

To shorten the testing session, we also dispensed with the standard belief and moral tasks used in Study 1, and we reduced the number of test questions for each moral vignette by combining the two intention questions into a single question and dropping the friend question. In addition, because children performed so well on true belief trials in Study 1, we only administered false belief trials in Study 2.

Second, to gain deeper insights into children's thinking about assigning consequences, we asked open-ended questions probing the reasons behind their judgment of deserved reward or punishment. In this way, we hoped to uncover underlying moral principles that might guide children's reasoning (Cushman et al., 2006). Third, because even 5-year-olds in our first study did not appropriately assign reward/punishment on false belief trials (and neither did some of the 7-year-olds in Killen et al., 2011), we included an adult sample for comparison purposes to determine a developmental endpoint for consequence reasoning. Finally, because some of the cell sizes were on the low side for our false belief groups, we increased the sample size for Study 2 so that the central tests of our hypotheses would be better-powered.

Following from the findings of Study 1, we hypothesized that children with lower FBU would not be able to appropriately attribute agent intentions but would instead tend to answer intention questions based on outcome and would do so more often than those with higher FBU. We also hypothesized a developmental lag for use of false belief information in response to intention questions, such that adults and 5-year-olds with higher FBU would make more appropriate intent-based judgments than 4-year-olds with higher FBU. In addition, we hypothesized 4 four-year-olds with higher FBU would have more difficulty assigning punishment/reward than 5-year-olds with higher FBU and adults. Finally, we hypothesized that punishment/reward judgments would again be harder than intent judgments, particularly for 4- and 5-year-olds. This study's design and hypotheses were preregistered prior to analyzing the data on the OSF (https://osf.io/3p5m9/). Study 2 data can also be found on the OSF. Note, however, that our analysis plan was revised following an editorial recommendation to use a nonparametric approach with our ordinal data.

Method

Participants

One hundred nine children participated, 63 four-year-olds (34 girls; M_{age} = 53.80 months, SD = 3.56) and 46 five-year-olds (21 girls; M_{age} = 65.70 months, SD = 3.33). An additional five children were tested but excluded from analyses owing to experimenter error (three), inattention (one), and family interference (one). The sample was representative of the population from which it was drawn. Eighty-six children were White, five reported being two or more races/ethnicity, three were Asian, two were Hispanic, Latino, or Spanish, two were Middle Eastern, and one was Native American/Native Alaskan. Ten parents did not report their child's race or ethnicity. Seventy-five percent of families reported making at least $40,000 a year.

Eighty-nine children participated in the lab and were recruited from a participant database at a large research university in the Western United States. Parents were provided compensation in the form of $10, and children received a small toy. A further 25 children participated in a quiet space at a local children's museum—parents of these children did not receive compensation, but children received stickers.

Forty-two adults (69% female, M_{age} = 19.71 years, SD = 2.50 years) from an undergraduate participant database also participated. The majority (26) reported being White, eight were Asian or Asian American, two were Black, and six reported being of two or more race/ethnicities or other. Participants received compensation in the form of class credit.

Design

Participants were randomly assigned to one of two intent conditions (good or bad) in which they responded to two morally

relevant belief vignettes. An a priori analysis indicated that at least 30 participants with FBU would be required in each intent condition for the study to be adequately powered (80%). Further Information on our a priori power analysis can be found in the online Supplemental Materials (S1). Because we anticipated that many children would fail the false belief (FB) task, we oversampled to achieve our targeted numbers of higher FBU. Ultimately, we recruited 63 four-year-olds (30 higher FBU), 46 five-year-olds (31 higher FBU), and 42 adults (40 higher FBU).

Procedure

Children who participated in the lab were tested individually in a single 5- to 10-min videotaped session. Children who participated at the local children's museum were tested in a quiet space in a single session. Adults completed all tasks in the lab on an iPad.

Measures

Similar to Study 1, participants in each age group were randomly assigned to the two intent conditions until the required sample size of 30 participants, with higher false belief understanding, was achieved: for children, good intent (N = 46, 28 girls, M_{age} = 57.70 months) or bad intent (N = 59, 26 girls, M_{age} = 60 months) and, for adults, good intent (N = 22, 14 women, M_{age} = 19.86 years) or bad intent (N = 20, 15 women, M_{age} = 19.75 years). Each participant heard two morally relevant belief vignettes (one with a boy agent and one with a girl agent) in which the agent held a FB, with vignette order counterbalanced across conditions. As in Study 1, good intent stories involved an agent who wanted to make a friend happy by sharing a desirable animal (kitten/butterfly). Because of a FB the agent ends up sharing an undesirable animal (skunk/spider). In the bad intent stories, the agent wanted to make their friend upset, but, because of a FB, ended up sharing a desirable animal. Unlike Study 1, the vignettes involved only one box as follows. The agent first opened the box, stating what animal was in it and whether it would make their friend happy or sad. The next animal then appeared on the scene with the agent stating whether it would make the friend happy or sad. At this point the agent left, after which the animal in the box jumped out and also left. The other animal then jumped in the box. The agent then returned and, when their friend arrived, they gave them the box that now contained the unintended animal.

Following each vignette, participants responded to six questions in a fixed order: (a) intention evaluation ("When [AGENT] handed [RECIPIENT] the box, was [AGENT] trying to be nice, mean, or just okay?"; (b) agent belief; (c) reality check; (d) consequence evaluation; (e) open ended consequence explanation ("Why should [AGENT] get [assigned consequence])?; and (f) recipient emotion attribution. Response choices were offered in a fixed order. The same three comprehension questions from Study 1 were again included.

Participants' responses to the open-ended punishment/reward questions were coded for three purposes: (a) content of the justification, (b) correctness of the justification, and (c) whether the justification matched the assigned punishment/reward. The content of the justification was coded as referencing: Mental state, such as, intention, knowledge, or belief (e.g., "she didn't know the spider got in the box," "trying to make her friend happy"); Outcome (e.g., "there's a butterfly in the box"); or Undifferentiated/Uninterpretable (e.g., "I do not know," or "just because"). The justification was also coded for correctness (i.e., judging whether it matched facts from the story): Correct (e.g., "she was trying to be nice" in the good intent condition), Incorrect (e.g., "she was trying to be mean" in the good intent condition), or Undifferentiated/Uninterpretable. Last, the justification was coded for whether it matched the assigned punishment/reward (e.g., assign a reward to an agent because "she was trying to do something nice") or Incorrect (e.g., assign a reward to the agent even though they state that "She was being mean"). Assigning nothing (no treat or reward) was coded as a match because that would be considered acceptable in both scenarios. If the initial justification was coded as undifferentiated, it remained undifferentiated in this coding. Two independent coders attained high interrater reliability (Cohen's K = .82), with 90% agreement across 792 observations. All discrepancies were resolved in discussion between the two coders and the first author.

Adults answered the same questions as children. The only difference was that adults were asked the three comprehension questions at the conclusion of each vignette because we were not concerned about memory demands for them, whereas children were asked during each vignette.

Results

No main effects of gender, story order (girl vs boy story first), or experimenter were found. Therefore, these factors were collapsed in subsequent analyses. As in Study 1, we used the recipient emotion prediction question as a comprehension check. The majority of children and adults appropriately judged that the recipient would feel sad in the good intent condition (93% trials) and happy in the bad intent condition (84% trials).

Agent Belief

We conducted an intent (good vs. bad) × age (4- vs. 5-year-olds) mixed-effects logistic regression on agent belief understanding. Adults were not included in this analysis because they were essentially at ceiling in both conditions: good intent (100% of trials); bad intent condition (99.75%; only one trial was incorrect). In the first model we included age group, intent condition, and their interaction as fixed effects and participant ID as a random factor. The interaction term was not significant, and hence was dropped from the final model. As in Study 1, a main effect of age revealed 5-year-olds (79% correct) were more likely to attribute the appropriate belief than 4-year-olds (50% correct), B = 1.24, SE = .001, z = 1020.50, $p < .001$, OR = 3.46, 95% CI [3.45, 3.47]. The main effect of intent was also significant: Children were more likely to attribute the appropriate belief in the good intent (75% correct) than the bad intent condition (53% correct), B = .93, SE = .001, z = 761.10, $p < .001$, OR = 2.52, 95% CI [2.52, 2.53].

As in Study 1, the sample was then divided into groups as a function of belief understanding: participants who failed one or both belief questions (N = 50, 48 children, M_{age} = 57.80 months, SD = 6.76 months, and 2 adults), 4-year-olds who passed both (N = 30, M_{age} = 53.60 months, SD = 3.19), 5-year-olds who passed both (N = 31, M_{age} = 65.6 months, SD = 3.39), and adults who passed both (N = 40).

Agent Intention

Figure 3 depicts participants' responses to the agent intention question. Adults were at ceiling: They reported that the agent in the bad intent condition was trying to be mean and the agent in the good intent condition was trying to be nice. For children, a 3 (FB group) × 2 (intent) logistic regression revealed a main effect of group that was qualified by a significant interaction. Four-year-olds (B = 3.47, SE = 1.66, z = 2.09, p = .04, OR = 32.27 95% CI [1.25, 835.36]) and 5-year-olds with higher FBU (B = 6.16, SE = 1.83, z = 3.37, $p < .001$, OR = 471.59, 95% CI [13.12, 16953.43]) rated agents in the good intent condition as significantly better intentioned than agents in the bad intent condition. In contrast, children with lower FBU did not significantly differentiate intention ratings on the basis of intent condition, B = .64, SE = .92, z = .70, p = .49, OR = 1.91, 95% CI [.31, 11.66].

Agent Consequence

Figure 4 depicts participants' responses to the agent consequence question. A 4 (FB group) × 2 (intent) mixed-effects logistic regression on agent consequence revealed a main effect of FB group that was qualified by a significant interaction. Follow-up simple effects tests revealed that only adults appropriately rated agents in bad intent conditions as deserving more punishment than those in good intent conditions, $p < .001$, OR = 21.44. No other FB group distinguished punishment ratings for agents in the good versus bad intent conditions ($ps > .16$).

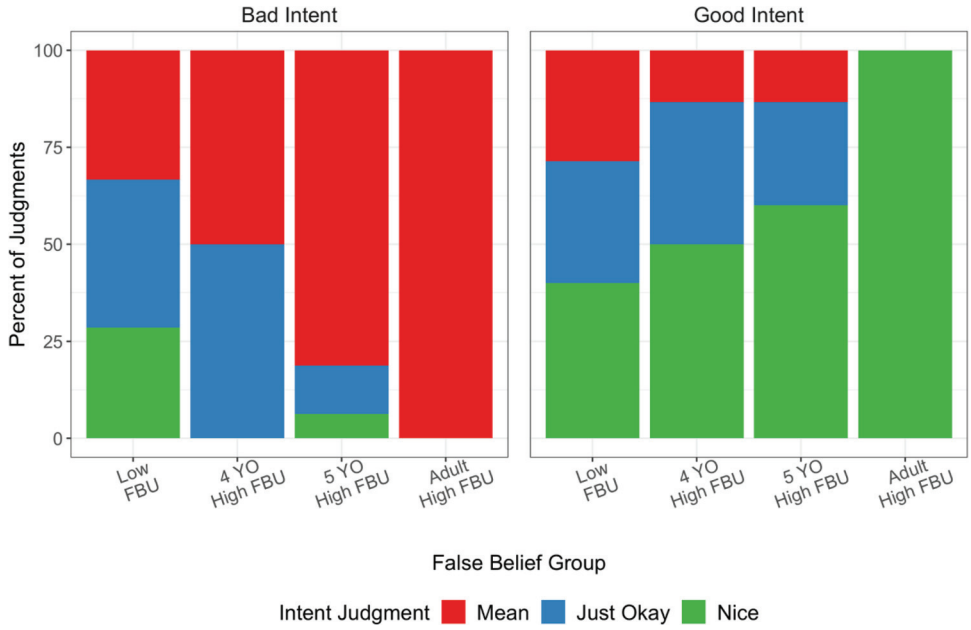
We conducted the same follow-up analysis as in Study 1 to examine whether FB groups differed in whether their consequence ratings were correct or incorrect. Adults were excluded because their ratings in both conditions were at ceiling on this measure. For children, we conducted a 3 (FB group) × 2 (intent) mixed-effects logistic regression on incorrect/correct consequence judgments. As in the preceding analysis, however, there were no significant effects on consequence judgments for children. Participants with lower FBU (52% both correct), 4-year-olds with higher FBU (53% both correct), and 5-year-olds with higher FBU (48% both correct) had similar difficulty assigning consequences in the FB context, with no group performing better than chance, Wilcoxon Ws = 637, 296, 270, $ps > .08$, respectively.

Justification for Punishment or Reward

Finally, we examined the proportion of justifications for punishment/reward broken down by FB group for the content of the justification, correctness of the justification, and whether the justification matched the assigned punishment/reward. Because 4- and 5-year-olds with higher FBU revealed essentially the same pattern of justifications across the three codes, we collapsed them for this analysis. For children, a large portion of justifications were coded as undifferentiated/uninterpretable (43% vs. 10% for adults). Of the justifications that were interpretable, the majority for all FB groups matched their justifications with the punishment/reward they assigned: 92% for participants with lower FBU, 97% for children with higher FBU, and 100% for adults.

With respect to the content of the justifications, mental states were referred to more often than outcomes, and that was true even for children with lower FBU (see Table 1). In a further analysis, however, we examined whether the mental state reference was consistent with the story. Even though children lower FBU referenced mental states more than outcomes, they often did so incorrectly. Only 42% of participants with lower FBU who referenced a

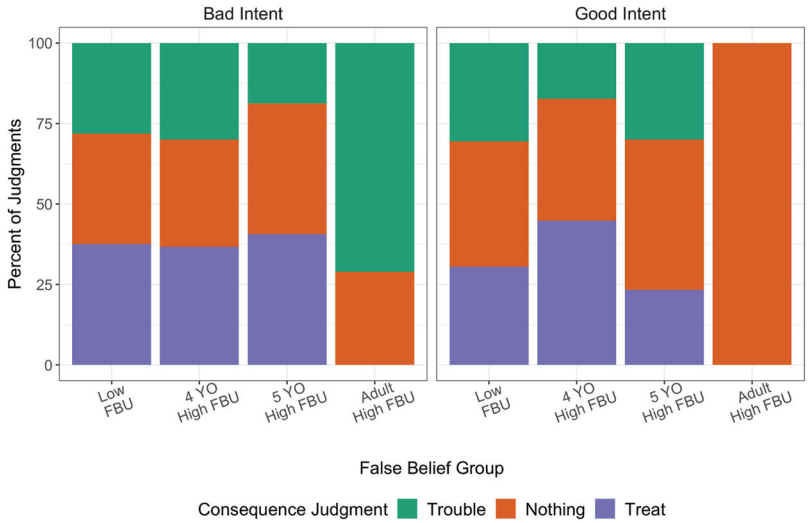
Figure 3
Percent of Agent Intent Judgments Separated by False Belief Group and Intent Condition in Study 2



Note. FBU = false belief understanding. See the online article for the color version of this figure.

Figure 4

Percent of Consequence Judgments Separated by False Belief Group and Intent Condition in Study 2



Note. FBU = false belief understanding. See the online article for the color version of this figure.

mental state did so correctly, versus 85% for children with higher FBU, and 97% for adults with higher FBU. The other 58% of participants with lower FBU referenced mental states that were inconsistent with the vignettes (e.g., stating that "he wanted to give his friend something nice" in the bad intent condition or "because she did something mean" in the good intent condition).

Discussion

In this study we examined whether the developmental differences found in the first study would persist when information processing demands were reduced. In contrast to Study 1, both 4- and 5-year-olds with higher FBU appropriately rated agents in the good intent condition as significantly better intentioned than agents in the bad intent condition. In the first study only 5-year-olds with higher FBU had made this distinction. It appears that reducing the task processing demands may have helped younger children focus on the relevant information needed to make accurate moral judgments. It is also possible that the new wording of the intent question (an amalgam of the two types of intention questions from Study 1) may have affected children's performance.

We had also hypothesized that 4-year-olds with higher FBU would have more difficulty assigning appropriate consequences than 5-year-olds with higher FBU. However, we again found that children of all ages, regardless of FBU, had difficulty doing so. Despite the reduced processing demands, children did not significantly differentiate consequence ratings for agents in the good versus bad intent conditions.

In addition, we hypothesized that consequence judgments would be harder than intent judgments for children. As in Study 1, this hypothesis was confirmed: 4- and 5-year-olds with higher FBU were often able to properly assign agent intent but generally did not make appropriate use of this information when assigning a punishment/reward. With respect to consequence justifications, we found that all groups tended to reference mental states more often than outcomes. This initial finding was surprising because previous research suggests that younger children, and especially those without false belief understanding, often focus more on outcomes when assigning consequences (Cushman et al., 2013; Zelazo et al., 1996). However, when examining whether participants' references to mental states were accurate, children without FBU often incorrectly referenced agents' intentions by aligning those intentions with their mischaracterizations of agents' beliefs. For example, in the good intent condition, they appeared to think that the agent knew there was an unpleasant object in the container, and therefore they thought he or she must have a bad intention. As a result, they recommended punishment for the agent. In that sense, these children were perhaps still outcome-focused—they assigned intentions to the agent in their justifications that matched the outcome of the vignette.

Last, we found that for children with higher FBU, moral judgments differed from those of adults in several respects. Both children and adults significantly distinguished good intentions from bad intentions. However, whereas adults were at ceiling in doing so, children were much more variable. Moreover, whereas children did not significantly distinguish good versus bad intent in their consequence judgments, adults had little difficulty in doing so.

General Discussion

Although the influence of intent understanding on children's moral development has been long studied (e.g., Piaget, 1932/1965), very little research has examined the influence of belief understanding on that development. In two studies we presented children with morally relevant belief vignettes to examine the extent to which they incorporate both intent and belief information in their moral judgments. By sampling within the narrow developmental window during which false belief understanding is acquired,

we found important developmental changes in children's ability to integrate mental state understanding with moral reasoning.

In Study 1, when agents held *true* beliefs, 4- and 5-year-olds appropriately shifted their moral judgments according to whether the agent had a good or bad intention, and children did so irrespective of whether they demonstrated understanding of false beliefs. Children appropriately rated agents sharing pleasant containers as positively intentioned and deserving of reward, while rating agents sharing unpleasant containers as negatively intentioned and deserving of punishment. These findings are important because they show that children of this age have a basic understanding of the role of intent in determining both the moral status of agents and whether or not those agents deserve punishment or reward.

Yet, in the otherwise similar *false* belief conditions, the pattern was quite different. In those conditions, we anticipated that the intersection of intent and false belief would lead to differential response patterns across the sample concordant with the attainment of false belief understanding. We predicted that children without FBU would perform poorly, responding in terms of outcome when assessing intent and consequences. Yet they did not show a distinct pattern of relying only on outcome information, instead children seemed to be responding randomly. It's likely that some children found it difficult to incorporate the contradictory (intent and outcome did not align) information and therefore may have reverted to guessing. For children with FBU, we had two contrasting hypotheses. If they immediately integrate such understanding with moral reasoning, they should use this information to correctly infer agents' intentions. In contrast, if there is an integration delay, the younger of these children might perform poorly, again defaulting to outcome-based or other incorrect responses. Our findings were consistent with the latter hypothesis: 5-year-olds with higher FBU, but not 4-year-olds with higher FBU, appropriately shifted their evaluations as a function of intent in cases where agents held false beliefs. Further, as hypothesized, we found support for a developmental lag between appropriately assigning intent and appropriately assigning consequences. Even 5-year-olds with higher FBU did not distinguish ratings of punishment or reward as a function of intent in false belief contexts (whereas they had done so successfully in true belief contexts).

In Study 2, we attempted to replicate the developmental lag in conditions in which the task processing demands were reduced. We did so by using only animals (and not food items) so that children would only have to direct their attention to the animals of interest rather than a third animal entering and moving the food items from one box to another. Further, we used only a single box such that children would not have to track the movements of two animals across two separate boxes. We also added an adult comparison group to establish a developmental endpoint, hypothesizing that 4-year-olds with higher FBU would have more difficulty assigning consequences than 5-year-olds with higher FBU and adults. Finally, we believed that consequence judgments would again be more difficult than intent judgments, particularly for 4- and 5-year-olds.

We found that when processing demands were reduced, the developmental lag for intent judgments was no longer evident: 4-year-olds, as well as 5-year-olds with higher FBU now performed better than chance in attributing agents' intentions. Nonetheless, children of both ages continued to have difficulty making appropriate punishment/reward judgments, indicative of a developmental lag between assigning moral intent and assigning consequences. In contrast, adults showed no such difficulty, performing near ceiling on both intent and consequence questions (when the latter were scored as correct/incorrect).

Why is it that the developmental lag in integrating false belief understanding with intent judgments largely disappeared in Study 2? In Study 2, we shortened the testing session, simplified the procedure, and only had children respond to false belief vignettes. As aforementioned, reducing information processing demands may have allowed four-year-olds to apply their FBU to their judgment of intent. Nonetheless, it is worth noting that, even though performance improved in Study 2, not all 4- and 5-year-olds with higher FBU correctly inferred the intent. In contrast, adults were at ceiling, indicating that a lag still exists before children reach adult levels.

More generally, the findings of Study 2 highlight that both conceptual changes and information processing improvements are likely implicated in the integration of ToM with moral judgment. With respect to processing improvements, we know that the development of executive function is strongly linked to ToM development (e.g., Carlson & Moses, 2001; Devine & Hughes, 2014), in particular that executive skills are necessary although not sufficient for ToM. In the same vein those skills may be necessary but again not sufficient for integrating ToM with moral reasoning. In Study 2 we reduced the processing demands (arguably executive in nature) and that appeared to generate improvements in four-year-olds' responses to intention questions. Reducing those demands, however, did little if anything to help children with their consequence judgments. The incorporation of ToM into moral consequence reasoning may thus require a further conceptual advance, as Cushman et al. (2013) have argued.

Our findings are both similar to and different from those of Killen et al. (2011) who also found that FBU is related to moral judgment. Specifically, they found that only the oldest children (7- to 8-year-olds) attributed good intentions to an accidental transgressor in a false belief context in which a good intention generated a moral violation. We extended Killen et al.'s approach by also including scenarios in which an agent had a negative intention, but due to a false belief ended up making a friend happy. Like Killen et al., we found that false belief understanding was linked to moral judgment. However, unlike Killen et al., our children made the link at an earlier age, with many 5-year-olds showing an understanding of moral intentions in the false belief context in Study 1 and many 4-year-olds doing so in Study 2.

Further, children in Killen et al. (2011) deemed punishing the accidental transgressor as less acceptable with age, whereas children in our studies did not make a clear distinction in whether to punish or reward agents in either the positive or negative intent condition. The findings across their research and ours are difficult to compare, however, because the question formats differed and because their study included older children. That said, their 5-year-olds appeared to have had similar difficulty to children of that age in our sample.

Our findings with respect to the difficulty children have in assigning reward and punishment appropriately parallel some findings from Cushman et al. (2013). They found that although children can make appropriate moral wrongness judgments based on intent, their punishment judgments are based more on outcome, especially at younger ages. Specifically, in Cushman et al.'s

research children were asked to make moral wrongness judgments (similar to our intent evaluation judgments) and punishment judgments for cases of accidental harm (benign intent, bad outcome) versus attempted harm (negative intent, neutral outcome). Children aged 5 and older assigned higher levels of punishment than wrongness in the case of accidental harm, and, conversely, higher levels of wrongness than punishment in the case of attempted harm. In contrast, four-year-olds did not clearly distinguish punishment and wrongness in either case.

Thus, as in our studies, appropriately assigning consequences on the basis of intent was difficult for children in Cushman et al. (2013) research. This difficulty is manifested under at least two different circumstances. In Cushman et al., there is a conflict between intent and outcome that is driven by acts that accidentally deliver unintended outcomes (e.g., a child throws a ball to the bin in which it belongs but it accidentally breaks a mirror instead; conversely, another child deliberately throws a ball to break a mirror but it lands instead in the bin). In our studies, the conflict between intent and outcome is driven by the presence of false beliefs as we have described. Cushman et al. argue that early in development two systems operate in determining consequence judgments. One is outcome-based, the other is intent-based. They argue that a conceptual reorganization takes place such that over development the intent-based system increasingly constrains, but does not fully override, outcome based consequence judgments. Our findings are very much consistent with that view.

Future Directions and Limitations

Our findings might be followed-up in a variety of ways. First, unlike adults, even the oldest children in our sample failed to assign consequences differentially as a function of intent in the false belief context. As a result, we do not know when in development children reach adult levels of understanding. Hence, it will be important to include older age groups in future work.

Second, we found a number of interesting effects among our 4- and 5-year-olds. However, our sample sizes were not large enough to fully specify the developmental trajectory across those ages. Future work might include larger samples that would allow age to be treated continuously rather than categorically and so make possible stronger inferences about the functional relations between age and moral judgment.

Third, it is possible that our scenarios were perplexing to some children given that no explicit reason was provided as to why agents would want to make their friend happy or sad. Although making a friend happy is perhaps something of a default, it may be harder to take in why someone would want to make a friend sad without a specific reason being offered. Although children did not appear to be confused by these occurrences in the true belief context, it proved harder for them to integrate that information when a false belief was in the mix. Clarifying the reasons behind agents' actions, particularly negative intentions, may therefore be an important addition in future work.

Fourth, our scenarios involved cases of psychological harm or help (giving recipients items they either liked or did not like). It's possible that intended physical harm or help may be more familiar in children's experience and therefore easier to reason about. That said, physical harm might also have the reverse effect: If outcomes are made highly salient, children may have an especially difficult time setting aside those outcomes and reasoning in terms of less salient intentions.

Fifth, participants, particularly children, may have been influenced by the order of the test questions which were presented in a fixed order. That order was chosen to align with the natural order in which actions and their outcomes occur: Intentions are formed, agents act out those intentions framed by the beliefs they hold, outcomes are generated, and consequences follow. Nonetheless, the fact that the consequence questions always followed the belief and reality questions may have made the outcome quite salient such that it became primary in children's minds when they thought about consequences. In contrast, intention questions either came before the belief and reality questions (Study 2) or both before and after those questions (Study 1). Future research should examine the influence of question order, especially as previous research suggests question order can influence children's responses (Cushman et al., 2013; Nobes et al., 2016).

Last, asking children whether an agent's behavior is praiseworthy or blameworthy may reveal additional information regarding moral judgments. It is possible that praise and blame would be more closely linked to intention (and belief) than would reward and punishment (Malle et al., 2014). For example, although we may hesitate to give reward when outcomes are inadvertently bad, we may feel freer to offer praise for good intentions in those cases; conversely, although we may not punish when good outcomes fortuitously follow bad intentions, we may nonetheless be quite willing to blame ill-intentioned agents in such cases. In that respect, praise and blame may represent something of a way station on children's road to incorporating mental states into judgments of punishment and reward.

Conclusion

In sum, we found revealing interactions between false belief understanding and moral judgment. Although children are able to reason about intent (good and bad) in contexts in which an agent holds a true belief, it is only as children attain false belief understanding that they appropriately reason about intent in contexts in which an agent holds a false belief. Moreover, moral reasoning develops beyond the preschool years, as even 5-year-olds did not appropriately assign punishment/reward as a function of intent, and their understanding of moral intent itself lagged behind that of adults. Integrating theory of mind and moral judgment is clearly a complex developmental achievement that is vital for positive social relationships. That achievement is not, however, attained in a single step.

References

Astington, J. W. (2003). Sometimes necessary, never sufficient: False-belief understanding and social competence. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind: Implications for typical and atypical development* (pp. 13–38). Psychology Press.

Baird, J. A., & Astington, J. W. (2005). The development of the intention concept: From the observable world to the unobservable mind. *The New Unconscious* (pp. 256–276). https://doi.org/10.1093/acprof:oso/9780195307696.003.0011

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*(4), 1032–1053. https://doi.org/10.1111/1467-8624.00333

Christensen, R. H. B. (2019). "ordinal—Regression models for ordinal data." R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21. https://doi.org/10.1016/j.cognition.2012.11.008

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089. https://doi.org/10.1111/j.1467-9280.2006.01834.x

De Rosnay, M., Pons, F., Harris, P. L., & Morrell, J. M. (2004). A lag between understanding false belief and emotion attribution in young children: Relationships with linguistic ability and mothers' mental-state language. *British Journal of Developmental Psychology*, *22*(2), 197–218. https://doi.org/10.1348/026151004323044573

Devine, R. T., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development*, *85*(5), 1777–1794. https://doi.org/10.1111/cdev.12237

Hamlin, J. K. (2013). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition*, *128*(3), 451–474. https://doi.org/10.1016/j.cognition.2013.04.004

Killen, M., Lynn Mulvey, K., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition*, *119*(2), 197–215. https://doi.org/10.1016/j.cognition.2011.01.006

Lagattuta, K., & Weller, D. (2014). Interrelations between theory of mind and morality. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 385–407). Psychology Press. https://doi.org/10.4324/9780203581957

Lane, J. D., Wellman, H. M., Olson, S. L., LaBounty, J., & Kerr, D. C. (2010). Theory of mind and emotion understanding predict moral development in early childhood. *British Journal of Developmental Psychology*, *28*(4, Pt 4), 871–889. https://doi.org/10.1348/026151009X483056

Leslie, A. M. (1988). Some implications of pretense for mechanisms underlying the child's theory of mind. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing theories of mind* (pp. 19–46). Cambridge University Press.

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, *50*(1), 45–85. https://doi.org/10.1016/j.cogpsych.2004.06.002

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Nobes, G., Panagiotaki, G., & Bartholomew, K. J. (2016). The influence of intention, outcome and question-wording on children's and adults' moral judgments. *Cognition*, *157*, 190–204. https://doi.org/10.1016/j.cognition.2016.08.019

Ochoa, K. D., Rodini, J. F., & Moses, L. J. (2020). False belief understanding and moral judgment in young children. *PsyArXiv*. https://osf.io/3p5m9/

Piaget, J. (1965). *The moral judgment of the child*. Free Press. (Original work published 1932)

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Smetana, J. G. (2006). Social domain theory: Consistencies and variations in children's moral and social judgments. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 119–153). Erlbaum.

Smetana, J. G., Jambon, M., Conry-Murray, C., & Sturge-Apple, M. L. (2012). Reciprocal associations between young children's developing moral judgments and theory of mind. *Developmental Psychology*, *48*(4), 1144–1155. https://doi.org/10.1037/a0025891

Wainryb, C., & Ford, S. (1998). Young children's evaluations of acts based on beliefs different from their own. *Merrill-Palmer Quarterly*, *44*(4), 484–503.

Wellman, H. M. (2020). *Reading minds: How childhood teaches us to understand people*. Oxford University Press. https://doi.org/10.1093/oso/9780190878672.001.0001

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer.

Wimmer, H., Gruber, S., & Perner, J. (1984). Young children's conception of lying: Lexical realism—moral subjectivism. *Journal of Experimental Child Psychology*, *37*(1), 1–30. https://doi.org/10.1016/0022-0965(84)90055-9

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology*, *24*(3), 358–365. https://doi.org/10.1037/0012-1649.24.3.358

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Development*, *67*(5), 2478–2492. https://doi.org/10.2307/1131635

Received October 29, 2020
Revision received March 19, 2022
Accepted May 9, 2022 ■