Received: 12 November 2019 | Revised: 18 December 2020 | Accepted: 13 January 2021

DOI: 10.1111/desc.13093

Developmental Science WILEY

PAPER

Children favor punishment over restoration

Katherine McAuliffe[1,2] | Yarrow Dunham[2]

[1]Department of Psychology, Boston College, Chestnut Hill, MA, USA

[2]Department of Psychology, Yale University, New Haven, CT, USA

Correspondence
Katherine McAuliffe, Department of Psychology, Boston College, Chestnut Hill, MA, USA.
Email: mcaulikg@bc.edu

Funding information
John Templeton Foundation, Grant/Award Number: 56036

Abstract

Why do people punish selfish behavior? Are they motivated to punish perpetrators of selfishness (retribution) or to compensate the victims of selfishness (restoration)? Developmental data can provide important insight into these questions by revealing whether punishment of selfishness is more retributive or restorative when it first emerges. Across two studies, we examined costly third-party intervention in 6- to 9-year-olds. In Study 1, children learned about a selfish actor who refused to share with a recipient. Children then chose to (1) punish the selfish actor by rejecting their payoff (*retribution*); (2) compensate the victim of selfishness by equalizing payoffs between the perpetrator and victim (*restoration*); or (3) do nothing. We found that children were more likely to punish than compensate in response to selfishness, suggesting that intervention in this context is more retributive than restorative. In Study 2, we tested third-party intervention in the face of generosity which, like selfishness, can lead to unequal outcomes. As in Study 1, children in this context could reject unequal payoffs, thereby depriving the recipient of the advantageous payoff but having no effect on the actor. Children could also use compensation in this context, equalizing the payoffs between actor and recipient. We found that children did not punish inequality that stemmed from generosity, suggesting that the retributive punishment in Study 1 was specifically targeting selfishness rather than inequality more generally. These results contribute to the debate on the function of third-party punishment in humans, suggesting that retributive motives toward selfish transgressors are privileged during ontogeny.

KEYWORDS
cooperation, punishment, restoration, retribution, social-cognitive development

Research Highlights

- From the age of 6, children show costly third-party punishment of selfishness. However, their motivations for punishment are unknown.
- We test whether children are driven to punish by retributive or restorative motives.
- We find that children retributively punish selfishness, even when restoration is an option.
- These effects do not generalize to unequal but generous allocations, suggesting these motives are specific to intervention against selfishness.

1 | INTRODUCTION

Across societies people are willing to pay to punish those who behave unfairly (Henrich, Ensminger, et al., 2010; Henrich et al., 2006). Although punishment is widely acknowledged to be an effective means of promoting cooperative behavior (Boyd & Richerson, 1992; Fehr & Gächter, 2002), the psychological motivations supporting punishment are not well understood. Why do people punish those who violate fairness norms, especially when the norm violation does not affect them directly (so-called *third-party punishment*)? One possibility is that they do so to penalize the perpetrator of unfairness (referred to here as a *retributive* motive) and perhaps, in doing so, increase future compliance to fairness norms.[1] Alternatively, people may willingly inflict costs on those who have behaved unfairly in order to decrease inequalities between others. That is, punishers intervene because punishment makes the victim of bad behavior *relatively* better off (referred to here as a *restorative* motive). These two possibilities are often conflated in standard third-party punishment tasks (e.g., Fehr & Fischbacher, 2004), wherein punishment negatively affects those who have been selfish while simultaneously decreasing the inequality between the selfish individual and the victim of selfishness.

Recently, studies have begun to disentangle retributive and restorative motives in adults by giving participants the choice to punish perpetrators—an option more clearly in line with retributive motives—or to compensate victims—an option more clearly in line with restorative motives. For instance, FeldmanHall and colleagues found that when people were themselves the victims of selfishness, they preferred compensation for their low payoff as opposed to punishment of the selfish individual (FeldmanHall et al., 2014). However, when third parties witnessed unfair behavior from one individual toward another, they preferred an option that compensated the victim of unfairness while also punishing the unfair individual, thereby creating inequality that disadvantaged the perpetrator of selfishness. These data suggest that victims of unfairness focus on restoration while witnesses of unfairness are, at least in part, driven by retributive motives. Another recent study suggests that third parties readily punish unfairness when other options are not available (Chavez & Bicchieri, 2013). However, when third parties are able to choose between rewarding fairness, compensating victims of unfairness, and punishing unfair behavior, they showed a preference for positive rewards and compensation over punishment (Chavez & Bicchieri, 2013). Together, these studies have importantly advanced our understanding of what motivates third parties to intervene in response to unfair behavior in adults. Moreover, they suggest that motivations in adult interveners are mixed or at least more complex than previously thought.

One reason why motivations underlying punishment decisions may be mixed in adults is that, by adulthood, many people have been exposed to formal ideas and institutions of justice which endorse, to varying degrees, both retributive and restorative reasons for punishment (Carlsmith et al., 2002; Darley & Pittman, 2003). Under this influence, individual punishers may be motivated by differing norms, or simultaneously motivated by multiple, not entirely consistent norms, producing conflicting results. Studying the motivations underlying third-party intervention, when exposure to these socializing forces is considerably more limited, can provide insight into whether retributive or restorative motivations are privileged when children are first beginning to intervene against unfairness.

Third-party intervention emerges early in development, with children as young as 3, intervening to tattle on puppets who have broken a rule (Vaish et al., 2011), to punish theft (Riedl et al., 2015), and to punish property destruction in peers (Yudkin et al., 2019). In some cases, children will even use corporal punishment to inflict harm on antisocial others (Kenward & Östh, 2015), yet there are constraints on their willingness to do so (Marshall et al., 2019). Third-party punishment of fairness norm violations, on the other hand, emerges somewhat later. From the age of 6, children show costly third-party punishment in response to distributive norm violations (Gummerum & Chu, 2014; Jordan et al., 2014; McAuliffe et al., 2015; Salali et al., 2015). For instance, 6- but not 5-year-olds will sacrifice their own rewards to prevent a selfish allocation of resources from being enacted between others (McAuliffe et al., 2015). Other work has shown that punishment by uninvolved third parties is an effective deterrent in children: the threat of third-party punishment increases children's cooperation in the Prisoner's Dilemma (Lergetporer et al., 2014). However, past work on children's third-party punishment of unfair behavior has been unable to distinguish between restorative and retributive motivations. For example, in McAuliffe et al. (2015), children may have been intervening for retributive reasons, that is, because they wanted to inflict costs on a selfish transgressor. Alternatively, children may have been driven by restorative motives, that is, they wanted to ensure that the victim ended up in equal standing to the perpetrator. Because of the structure of the task, the only means of equalizing payoffs was to take away the selfish individual's rewards, leaving both parties with nothing. Thus, we currently do not understand why children intervene to punish distributive norm violations when they are uninvolved third parties.

A recent study investigating children's third-party punishment of ownership violations (theft) took an important first step towards identifying children's motivations for punishment (Riedl et al., 2015). In this task, 3- and 5-year-olds children saw one puppet steal an object from another. Children were then given a choice between (1) taking the object away from the thief and making it inaccessible to both parties (punishment) or (2) taking it away from the thief and giving it back to the original owner (punishment and restoration). They found that children typically chose the latter option, showing that they would forgo pure retributive punishment when they could restore the stolen object to the victim while simultaneously penalizing the thief. This study demonstrated that, at least in the context of ownership norm violations, children are driven partly by restorative motives. However, it left unclear whether children were motivated to inflict costs on the transgressor or whether this was simply a necessary byproduct of restoring the stolen object to its rightful owner (Van de Vondervoort & Hamlin, 2015). Moreover, it raises the

question of whether restorative motives are specific to ownership norm violations or whether they generalize to other types of norm violations.

Here, we investigate the motivations underlying third-party intervention against fairness norm violations during the period when costly third-party punishment is first emerging in development (McAuliffe et al., 2015). Across two studies, we ask whether retributive or restorative motives are privileged during ontogeny in 6- to 9-year-old children. This age range was selected because previous work has shown that children in the USA begin to show costly third-party punishment of selfishness around the age of 6 (Bernhard et al., 2020; McAuliffe et al., 2015) and are even more likely to punish by age 8 (Jordan et al., 2014). While we had no specific age-related predictions, we expected that this relatively broad age range would provide an ideal window into children's motivations for third-party intervention.

In Study 1, children played a third-party intervention game in which the third party learned about a divider who offered fair or selfish divisions of rewards to a recipient. Children then had three options: (1) they could accept the divider's allocation at no personal cost; (2) they could sacrifice their own rewards to reject the divider's payoff, thereby punishing the selfish divider; or (3) they could sacrifice their own rewards to equalize the payoffs between perpetrator and victim, thereby compensating the victim of selfishness. In Study 2, we asked whether the pattern of behavior we saw in Study 1 was a specific response to selfishness or a more general response to inequality. We did this by reversing the direction of inequality such that the divider was generous rather than selfish (as in McAuliffe et al., 2015). Children again had three options, but here, rejection would mean removing the payoff from the recipient of generosity—giving us a strong test of the idea that rejections are directed toward inequality as opposed to selfishness specifically.

2 | STUDY 1: WHY DO CHILDREN INTERVENE AGAINST SELFISHNESS?

2.1 | Materials and methods

2.1.1 | Participants

We tested N = 38 children between the ages of 6- to 9-years old (Mean age = 96.0 months; SD = 12.7; 19 females; see Table S1 for breakdown). Seven children were tested but excluded because they could not eat the candy rewards (2), we were missing a consent form (1), the session stopped early due to participant agitation (1), the participant refused to follow instructions (1), or we did not have a video recording of their session (2). Our sample size is consistent with previous work that used this third-party punishment task to study punishment in children in the USA (Jordan et al., 2014; McAuliffe et al., 2015). Both these past studies targeted N = 32 per cell. Note that these previous studies tested age effects so their cells were N = 32 per age group. We had no between-participant variables and did not test hypotheses about age in our main analyses, thus our specific aim was to recruit 40 children for this study, with our sample spread roughly evenly across boys and girls and 6- and 7-year-olds and 8- and 9-year-olds. Due to exclusions, our final cell number was not exact. We received written consent from participants' parents and verbal assent from subjects. This study was approved by the Yale Institutional Review Board.

2.1.2 | Design

Children were presented with 12 trials: six equal trials and six selfish trials. In the equal trials, the divider kept two candies and gave two to the recipient (2-2). In the selfish trials, the divider kept all four candies, giving none to the recipient (4-0). In two cases, the number of fair and unfair trials was not balanced within session because a fair trial was inadvertently run instead of an unfair trial. Within participant, trials were randomized with the constraint that no more than two of the same trial type could be presented consecutively.

2.1.3 | Procedure

Third-party intervention game

We used a modified version of a previously validated Third Party Punishment game (Jordan et al., 2014; McAuliffe et al., 2015). Participants were brought in to the testing area and introduced to Skittles (a small fruit-flavored sweet). They were asked to select one of two Skittle flavors, and their choice was the only flavor used throughout the game. They were then shown how to use the experimental apparatus (Figure 1; see Figure S1 for photograph). Using the apparatus, Skittles could be moved back and forth across small grooves in a platform. When the Skittles were in place at the farthest edges of the platform, a handle could be pulled in the direction of a green arrow, distributing the Skittles to two side trays (*accepting a distribution*). Alternatively, the handle could be moved in the direction of a red arrow, causing the Skittles to disappear underneath the apparatus (*rejecting a distribution*). Participants were asked to demonstrate their understanding of the handles and received experience pulling the handle in both directions. If children did not spontaneously demonstrate their understanding of the apparatus, the handles were re-explained (see Table S3 for detailed information about comprehension checks).

After learning about the apparatus, children were introduced to an absent *divider* and *recipient* who allegedly had played the game the day before. These absent children were gender matched to participant gender. Participants were told that they would make decisions that would affect the payoffs of both the divider and recipient, who were represented by drawings on paper bags. The participant learned that the divider had been given the opportunity to divide up four Skittles between themselves and the recipient. The divider's allocations were illustrated on paper cards. Participants were introduced to two example allocation cards and asked to demonstrate

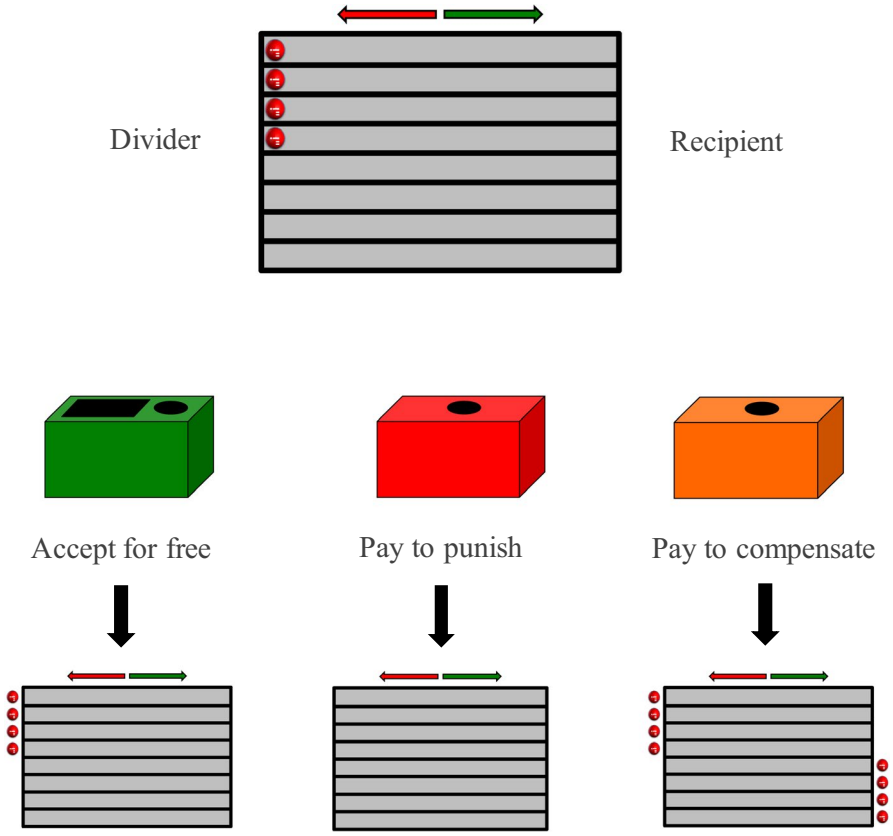

FIGURE 1 Participants played a Third-Party Intervention game in which they learned about an absent divider who had four sweets that could be shared with an absent recipient. Children learned how many sweets the divider shared and then decided whether to accept (green box), punish (red box), or compensate (orange box) the distribution. Accepting distributions did not involve sacrificing any rewards. Children simply moved a sweet from the green box back into the same box through the small hole and then pulled the handle toward the green arrow, thereby distributing the allocated sweets to the divider and recipient. If children wanted to punish or compensate, they moved one sweet from the green box to either the red or orange box, respectively. After a punishment decision, a handle was pulled towards the red arrow and all sweets disappeared from the board. After a compensation decision, the experimenter equalized payoffs between the divider and recipient and the equal payoff was distributed by pulling the handle toward the green arrow

their understanding of the recorded decisions. Participants were led to believe that the absent divider and recipient were real children whose ultimate payoffs would be decided by their decisions.

Costly punishment and compensation

Before playing the game, participants were given an endowment of 27 Skittles, a number chosen based on past work on costly punishment in children (McAuliffe et al., 2015). Our reason for giving children such a large endowment was so that, on any given trial, they would have more Skittles than the absent divider and recipient, thereby controlling for the possibility that the child's feelings of envy toward those with more would drive intervention (Jordan et al., 2016; Leibbrandt & López-Pérez, 2012).

Children were then shown three boxes: a green box, a red box, and an orange box (Figure 1). Their Skittles were placed into the green box. They were taught that before making a decision they would have to remove one Skittle from their endowment, which was stored in the green box, and place into one of three holes.

If they wanted to *accept* a distribution, they would take a Skittle from the large hole in green box and place it through a smaller hole in the same box causing it to fall back into their endowment pile. They could then pull the handle to the green zone, distributing the Skittles between the divider and recipient. At the end of the game, they could take all the Skittles in the green box home. Thus, accepting a distribution was free but still involved engaging in a physical action.

If children wanted to *punish* a distribution, they would take a Skittle from the green box and place it into the red box. They could then pull the handle to the red zone, causing the Skittles to disappear under the apparatus so that neither the divider nor recipient would receive them. At the end of the game, the Skittles in the red box would be thrown away. Thus, punishment was costly.

If children wanted to *compensate* a distribution, they would take a Skittle from the green box and place it into an orange box. When they did this, the experimenter would equalize the number of Skittles between the divider and recipient. For example, if the divider kept all four Skittles, the experimenter would put four on the recipient's side. If the divider kept two and gave two, the experimenter would give two additional Skittles to the divider and two to the recipient. Thus, compensation in response to both selfish and equal trials resulted in four candies for the recipient and four candies for the divider. We ensured

that it was possible to 'compensate' on equal trials to control for the possibility that children would enjoy creating more value in the game. Once the distributions had been equalized, the participant could pull the handle to the green zone, distributing the Skittles to the divider and recipient. Just like the red box, the Skittles in the orange box would be thrown away at the end of the game and thus compensation was costly. Children were asked to demonstrate their understanding of all three boxes and had experience moving Skittles into each box and pulling the handle in the correct direction.

Before test trials commenced, children were asked several comprehension questions. They were asked (1) whether the divider and recipient had taken home Skittles when they came in previously; (2) to identify the bags belonging to the divider and recipient; and (3) what would happen to the Skittles in the green/red/orange boxes at the end of the game. If children did not answer these questions spontaneously correctly, answers were re-explained. Following these comprehension questions, test trials were commenced. On the first three trials, children were asked to explain the divider's decision based on the distribution card. For the next nine trials, the experimenter stated the distribution. If needed, children were reminded that they needed to move a Skittle into a box before pulling the handle and which way to pull the handle once they had chosen a given box.

The following variables were counterbalanced between subjects: (1) the name of the divider and recipient (Jane/James and Annie/Andy); (2) the side of the divider and recipient with regards to the apparatus; (3) whether children first learned how to pull the handle to the green or red zone; (4) whether children were first introduced to an equal or unequal practice distribution card; and (5) whether children were first introduced to the red or orange box.

After making 12 decisions, children were again asked whether Skittles had already been taken home by the divider and recipient, to identify the bags belonging to the divider and recipient, and what would happen to the Skittles in each of the boxes. They were additionally asked to explain why they decided to move Skittles into each of the boxes (they were not asked about boxes they did not use). Following these explanations, children were asked whether they thought the absent divider and recipient were real or pretend. About half the children (47%) reported that they thought the other children were real, which we expect is due to the fact that they either misunderstood pretend as meaning 'not present' and/or because they were confused that the experimenter who introduced the absent peers was now questioning their existence. Importantly, children overwhelmingly answered all other questions about the absent children correctly (Table S3) and our results are robust to the inclusion of a belief term in our models (Table S4). At the end of the study, children were debriefed and told that the absent children were in fact not real children.

2.1.4 | Coding and analysis

Data were coded from videos except in one case in which the camera turned off mid-session and thus data for non-recorded trials were taken from live coding. Reliability between live and video coding was very good (κ = 0.92). All analyses were run in R version 3.6.3 (R Core Team, 2020). Our dependent measure consisted of three levels (accept, punish, or compensate). Analysis of a three-level dependent measure would typically be done using multinomial logistic regression. However, to our knowledge, current mixed effects modeling packages are unable to accommodate multinomial logistic models. Consequently, following recommendations from Dobson and Barnett (2008), we parameterized the multinomial model by fitting a series of mixed effects binomial models. To assess whether the divider's decision to be fair or unfair influenced participants' decisions to accept versus to punish or compensate, we fit two generalized linear mixed models with a common reference category (punish (coded as 1) vs. accept (coded as 0); compensate (coded as 1) vs. accept (coded as 0)). We then ran a third model that compared children's decisions to punish (coded as 1) to their decisions to compensate (coded as 0). Note that each model examined a specific subset of children's decision-making data: the first model excluded compensation decisions, the second model excluded punishment decisions, and the third model excluded acceptance decisions. By running these three models, we were able to contrast all levels of our dependent measure in line with multinomial regression, while controlling for participant-level variation in each regression. Note that our results hold across an alternative analytic approach in which we first analyze whether children are more likely to accept versus not accept (i.e., to compensate *or* punish) and second, compare punishment and compensation directly (see Figure S3; Table S5).

Our models included distribution (equal or unequal), participant gender, and a standardized age term (months). Age standardization was done using the 'scale' function, which divides each value by subtracting the mean of the vector and dividing by the standard deviation. Age was scaled to facilitate model convergence. In supplemental models, we confirmed that our main results hold when trial number (1–12) is included as a predictor (Table S7). In exploratory analyses, we examined the interactions between distribution and age and between distribution and trial number (see Table S8; Figure S6).

Participant ID was included as a random effect (intercepts) to control for repeated measures within individual. We assessed whether inclusion of particular terms improved model fit by dropping them from the model and comparing models with and without a given term using Likelihood Ratio Tests (LRTs; conducted using the 'drop1' command).

Finally, we coded the explanations given by children in the post-test question period to explore whether the concept of fairness was evoked when children were asked to explain why they chose to put their Skittles in the different boxes. Responses were double coded and inconsistencies were rare (<4% of trials) and resolved through consensus.

2.2 | Results

Children mostly accepted the dividers' allocations, presumably because they were themselves unaffected by the divider's behavior

and because intervention was costly (Figure 2). Despite this, distribution was a clear predictor of punishment but not compensation (Table 1). As shown in Figure 2, children were more likely to punish selfish divisions than equal divisions. Our model predicting punishment relative to acceptance showed that whether the divider shared selfishly or equally was a strong predictor of punishment (Table 1; distribution: $\chi^2_1 = 29.19$, $p < 0.001$). Age did not predict children's punishment behavior relative to acceptance ($\chi^2_1 = 0.43$, $p = 0.51$).

Compensation behavior relative to acceptance was not predicted by any of our terms (Table 1), although we found marginally significant effects of age ($\chi^2_1 = 3.27$, $p = 0.07$) and distribution ($\chi^2_1 = 3.36$, $p = 0.07$), which suggested that children were slightly more likely to compensate equal compared to unequal allocations and that older children were slightly more likely to compensate overall compared to younger children. However, because these effects were weak, we do not richly interpret them here.

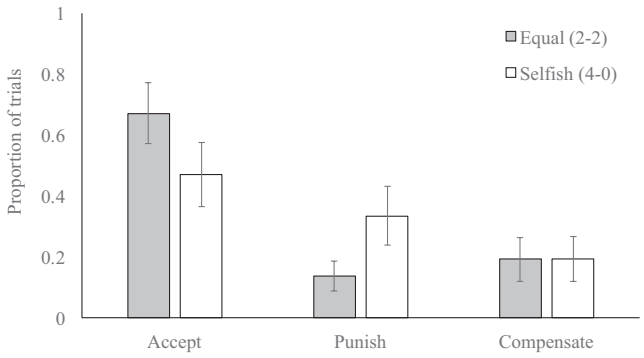

FIGURE 2 Children's decisions when confronted with distributions that had been shared with an absent recipient by an absent divider. Bars show proportion of trials in which children accepted, punished, or compensated when confronted with different distributions. The divider shared in one of two ways: they were either maximally selfish (kept 4, shared none) or shared equally (kept 2, shared 2). Children were presented with 6 selfish and 6 equal trials. Errors bars show 95% confidence intervals

When we compared punishment to compensation directly, we found that children were more likely to punish than compensate selfish offers compared to equal offers (distribution: $\chi^2_1 = 9.24$, $p = 0.002$; Table 1). We found no effect of age ($\chi^2_1 = 1.97$, $p = 0.16$).

When asked to explain their decisions, children frequently invoked the concept of fairness: 21% invoked fairness when asked why they chose the green box (acceptance), 29% when asked about the red box (punishment), and 32% when asked about the orange box (compensation).

2.3 | Discussion

Findings from Study 1 suggest that children preferentially punish rather than compensate selfishness, a response more consistent with a retributive rather than restorative motive for intervention in the face of selfishness. However, before making any strong conclusions we wanted to rule out an alternative explanation that was also consistent with this result. Namely, children may not be specifically motivated to punish *selfish* individuals but rather are motivated to intervene against those who create *inequality* in any form. Evidence in support of the idea that children strongly dislike inequality comes from past work showing that—particularly at the older end of our age range—children reject unequal allocations even when they place the child at an advantage relative to a peer (reviewed in McAuliffe et al., 2017). In our Study 1 design, selfishness was confounded with inequity, and we are thus not yet in a position to disambiguate these alternative explanations for our findings.

To disentangle these two potential accounts, we designed a follow-up study in which children were again presented with an absent divider who divided resources equally or unequally, as in Study 1. However, unlike Study 1, unequal allocations were *generous* instead of selfish. In these cases, the absent divider gave all four candies to the absent recipient, keeping none for themselves. We were interested in whether our punishment results would transfer from Study

Selfish Inequality	Punish versus accept	Compensate versus accept	Punish versus compensate
Intercept	−2.10 (0.40)***	−1.74 (0.42)***	−0.40 (0.34)
Scaled age	0.15 (0.23)	0.49 (0.28)	−0.27 (0.19)
Distribution: unequal	1.45 (0.28)***	0.53 (0.29)	0.97 (0.33)**
Gender: male	0.44 (0.47)	0.12 (0.54)	0.15 (0.37)
Akaike information criterion	395.85	365.42	265.56
Bayesian information criterion	415.39	384.70	281.92
Log likelihood	−192.93	−177.71	−127.78
Number of trials	368	349	195
Number of participants	38	38	34
Participant ID (intercept)	1.36	1.79	0.31

**$p < 0.01$.

***$p < 0.001$.

TABLE 1 Estimate and standard error of effects in mixed models predicting participants' behavior in the Third-Party Intervention Game. In the first two models, punishment (=1) and compensation (=1) are predicted relative to acceptance (=0). In the punishment versus compensation model, punishment (=1) is predicted relative to compensation (=0). Baselines for factors were as follows: distribution =equal, gender =female. Table also shows goodness-of-fit statistics. Because each model examines a different subset of decision data, the sample size depends on whether children made the target decisions. For instance, data from only 34 children are included in the third model because four children in our sample never punished or compensated

1 to Study 2. If children continue to invest in punishment, this would cast doubt on our interpretation that punishment was retributive in Study 1. Rather, it would indicate that children were punishing in response to inequality rather than selfishness. However, if children no longer punish unequal allocations when inequality is borne of generosity, this would suggest that children in Study 1 were indeed specifically targeting the perpetrator of selfishness and would lend support to our claim that punishment was retributive rather than restorative.

We retained a compensation option in Study 2 to maintain structural parallels across studies. However, we note that we were less interested in comparing compensation from Study 1 to Study 2. This is because compensation in the face of generosity (Study 2) could be interpreted as equalizing, as in Study 1 (selfishness), or it could be interpreted as rewarding generosity, something which was not possible in Study 1. Thus, our main goal in designing Study 2 was to help adjudicate between the possibility that costly third-party punishment targets selfishness specifically and the possibility that it targets inequality more generally.

3 | STUDY 2: ARE CHILDREN SIMILARLY 'RETRIBUTIVE' IN THE FACE OF GENEROSITY?

3.1 | Materials and methods

3.1.1 | Participants

We tested $N = 40$ children between the ages of 6- to 9-years old (Mean age = 93.5 months; SD = 12.4; 20 females; see Table S1 for breakdown). Five children were tested but excluded because they could not eat the candy rewards (2), the parent reported that they were not typically developing (1), the session stopped early due to participant agitation (1), or because we did not have a video recording of their session (1). As in Study 1, our aim was to recruit approximately 40 children.

3.1.2 | Design

As in Study 1, children were presented with 12 trials: six equal trials and six generous trials. The fair trials were identical to Study 1: the divider kept two candies and gave two to the recipient (2-2). In the generous trials, the divider gave all four candies to the recipient, keeping none for themselves (0-4). In one case, the number of fair and unfair trials was not balanced within participant because one unfair trial was invalid due to experimenter error. Within participant, trials were randomized with the constraint that no more than two of the same trial type could be presented consecutively.

3.1.3 | Procedure

The procedure was identical to Study 1 with the exception that generous allocations were presented instead of selfish allocations. As in Study 1, children overwhelmingly answered the comprehension questions correctly (see Table S3). A small majority of children reported that absent peers were real (60%) yet answered all questions about the peers correctly, and we thus believe this represents a problem with our phrasing of the question (see procedure section for Study 1 and see Table S4 for analyses including the belief term as a predictor).

3.1.4 | Coding and analysis

Coding and analysis was identical to Study 1. Again, reliability between live and video coding was very good ($\kappa = 1.00$).

3.2 | Results

As in Study 1, children in Study 2 mostly accepted dividers' allocations (Figure 3). Unlike Study 1, however, punishment relative to acceptance in Study 2 was not predicted by distribution ($\chi^2_1 = 1.89$, $p = 0.17$; Table 2). Age did not predict punishment either ($\chi^2_1 = 1.66$, $p = 0.198$; Table 2).

In exploratory analyses, we found that children's propensity to punish rather than accept generous divisions was predicted by an interaction between distribution and age ($\chi^2_1 = 8.45$, $p = 0.004$; Table S2). Inspection of the predicted effects from this model (Figure S2) suggests that this interaction was due to the fact that older children were more likely to punish (relative to accept) generous allocations compared with equal allocations, whereas younger children did not show this pattern, a pattern of results consistent with forms of advantageous inequity aversion seen in other work with children in this older age range (for a review, see McAuliffe et al., 2017).

In our model predicting children's compensation versus acceptance, distribution was a significant predictor of compensation ($\chi^2_1 = 5.28$, $p = 0.022$; Table 2; Figure 3): children were more likely to

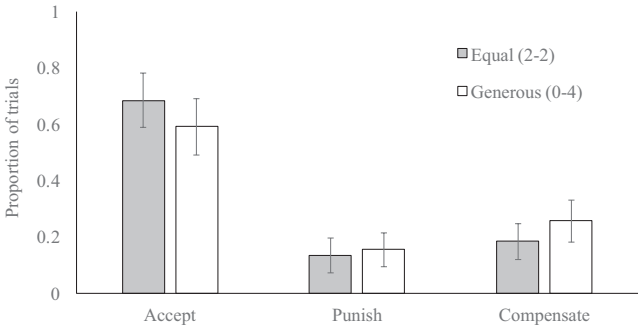

FIGURE 3 Children's decisions when confronted with distributions that had been shared with an absent recipient by an absent divider. Bars show proportion of trials in which children accepted, punished, or compensated when confronted with different distributions. The divider shared in one of two ways: they were either maximally generous (kept 0, shared 4) or shared equally (kept 2, shared 2). Children were presented with 6 generous and 6 equal trials. Errors bars show 95% confidence intervals

Generous inequality	Punish versus accept	Compensate versus accept	Punish versus compensate
Intercept	−1.96 (0.44)***	−1.32 (0.33)***	−0.38 (0.28)
Scaled age	0.37 (0.29)	0.38 (0.22)	−0.06 (0.15)
Distribution: unequal	0.42 (0.31)	0.58 (0.25)*	−0.17 (0.31)
Gender: male	−0.16 (0.57)	−0.34 (0.43)	0.13 (0.31)
Akaike information criterion	332.80	441.31	243.06
Bayesian information criterion	352.42	461.39	258.86
Log likelihood	−161.40	−215.65	−116.53
Number of trials	374	410	174
Number of participants	40	40	33
Participant ID (intercept)	2.02	1.13	0.00

*$p < 0.05$.
***$p < 0.001$.

TABLE 2 Estimate and standard error of effects in mixed models predicting participants' behavior in the Third-Party Intervention Game. In the first two models, punishment (=1) and compensation (=1) are predicted relative to acceptance (=0). In the punishment versus compensation model, punishment (=1) is predicted relative to compensation (=0). Baselines for factors were as follows: distribution =equal, gender =female. Table also shows goodness-of-fit statistics. Because each model examines a different subset of decision data, the sample size depends on whether children made the target decisions. For instance, data from only 33 children are included in the third model because seven children in our sample never punished or compensated

compensate generous divisions than equal divisions. We additionally found a marginally significant effect of age ($\chi^2_1 = 3.02$, $p = 0.082$), which suggested that older children were slightly more likely to compensate (relative to accept) than were younger children.

When punishment and compensation were compared directly, we found no effects of age ($\chi^2_1 = 0.16$, $p = 0.69$) or distribution ($\chi^2_1 = 0.29$, $p = 0.59$). In exploratory analyses, we found a marginally significant effect of the interaction between age and distribution ($\chi^2_1 = 3.44$, $p = 0.063$; Table S2). Inspection of predicted effects from this model (Figure S2) suggests that this is due to the fact that younger children were more likely to punish equal allocations than generous allocations, whereas older children showed the opposite pattern. However, this was a weak effect (Table S2).

When asked to explain their decisions, children again frequently invoked the concept of fairness: 18% invoked fairness regarding acceptance, 20% regarding punishment, and 35% regarding compensation.

3.3 | Discussion

Our results from Study 2 suggest that when children are confronted with generous but unfair behavior, they do not systematically punish, drawing a sharp distinction between punishment in the face of selfishness compared with generosity. These data help clarify and refine our understanding of children's behavior in Study 1, helping to rule out the possibility that the retributive punishment we observed in Study 1 was in fact intervention against *any* form of inequality.

Our findings from Study 2 also showed that children were more likely to compensate generous compared with equal allocations. Children in this context may thus have been motivated to offset the costs of generosity for the generous donor. This behavior could be construed as a form of reward. However, as discussed above, because reward was not possible in Study 1, we raise this tentatively as something deserving further investigation. Exploratory analyses from Study 2 also revealed a series of unexpected but interesting interactions with age, which we discuss in the General Discussion.

4 | GENERAL DISCUSSION

Findings from Study 1 demonstrated that when children intervened as third parties, they chose to punish selfish perpetrators rather than compensate the victims of selfishness. Our data support the claim that, as third parties, children seek retribution rather than restoration when confronted with a selfish distributive norm violation. Results from Study 2 showed that our punishment results from Study 1 are specific to selfish norm violations and do not generalize to the generous form of inequality. Across both studies, children frequently made explicit mention of fairness or related concepts, suggesting that they were knowingly intervening in response to fairness norm violations.

A retributive motive underlying punishment of selfish behavior in children is consistent with theories suggesting that punishment is an important means of promoting cooperative behavior in human societies (Boyd & Richerson, 1992). For punishment to work in this way, individuals must target the perpetrators of bad behavior and in doing so decrease their probability of future norm violations. By contrast, punishment for the sake of restoration may uphold a distributive norm in the short term by decreasing inequality, but is unlikely to have long-term consequences on the perpetrator's future compliance to norms because it fails to impose any costs on the perpetrator's selfish behavior.

Our finding that third parties are motivated to punish distributive norm violations aligns with a recent study on adults that showed that third- but not second-party punishment is partly motivated by retribution (FeldmanHall et al., 2014). In this study, however, participants could both punish selfish individuals while simultaneously compensating victims. It is thus not presently clear which motivation is the primary driver of adult behavior. By contrast, in our

task, we were interested in pitting these motivations against each other and seeing which was the primary driver of children's behavior when both motivations could not operate at once. In this same vein, our data are difficult to compare to the only other study, to our knowledge, that has investigated motivations underlying third-party punishment in children (Riedl et al., 2015). This study showed that children preferred to return stolen property to the victims of theft as opposed to punishing thieves without returning the stolen good. While these data suggest that retribution is not the exclusive driver of third-party intervention, it is impossible to assess its unique importance because restoration also punished the thief by taking the stolen object away. Additionally, we are hesitant to compare our data directly to these results since motivations supporting enforcement of ownership norms may be different from those supporting enforcement of distributive norms. Indeed, we view an exploration of the effects of different *kinds* of transgressions (e.g., property destruction (Yudkin et al., 2019), rule following (Vaish et al., 2011), and fairness norm violations (McAuliffe et al., 2015)) on the emergence and expression of third-party punishment and its underlying motivations as a key area for future work.

Our models predicting children's punishment and compensation relative to acceptance behavior in Study 2 (generous inequality) showed that children compensated more unequal than equal offers. As discussed above, this finding is intriguing in that it hints at the idea that children may have been using compensation as a reward for the costs that generosity can impose on a third party. Additionally, our exploratory age analyses showed that older children were more likely to punish generosity (when compared to acceptance) than were younger children. While perhaps puzzling at first glance, the finding that children punish generosity is consistent with past work on third-party punishment in children, in which children punished inequality to some degree even when it was generous (although less than when it was selfish; McAuliffe et al., 2015). Children may have punished generous offers because they expected others to adhere strictly to a norm of equality. In line with this, children in this age group show advantageous inequity aversion, refusing to accept unequal pay-off distributions that place them at an advantage relative to a peer (see McAuliffe et al., 2017, for a review). An additional possibility is that we are seeing an incipient form of antisocial punishment, which is observed in economic games with adults (Herrmann et al., 2008).

To further explore whether children's strategies varied in the face of selfishness compared with generosity, in a final exploratory analysis, we classified children into "types" based on whether they accepted all allocations (*full acceptor*), punished more than they compensated (*punisher*), compensated more than they punished (*compensator*), or showed equal levels of punishment and compensation (*equal punishment and compensation*; see Supporting Information for details; Figures S4 and S5; Table S6). We found that the frequency of types differed across studies when examining children's responses to unequal allocations: punishers were more common in Study 1 (selfish inequality) while compensators were common in Study 2 (generous inequality; $\chi^2_3 = 14.56$, $p = 0.002$). This pattern provides more evidence in support of the interpretation that children's motivations for intervention are dependent on whether inequality is borne of selfishness or generosity and, more specifically, is consistent with the interpretation that children are specifically motivated to punish perpetrators of selfishness.

It is important to highlight aspects of our study that raise questions for future work. First, while our sample size is consistent with past work on third-party intervention (e.g., McAuliffe et al., 2015), it will be important to replicate and extend these effect with a larger sample, perhaps changing the kinds of options that children have. For instance, would children's behavior change if punishment and compensation were more personally costly? Would children compensate even less if they had to do so from their own resources? Second, the resource disparities used here were extreme: children were either presented with fair allocations or maximally unequal (selfish or generous) allocations. It would be interesting to know whether children's probability of intervention depends on the degree of inequality, and future work could use this same paradigm to titrate the payoff difference between divider and recipient to understand the extent to which this affects children's intervention. Third, work with adult participants has fruitfully compared people's motives for intervening in second- versus third-party contexts and found that their motives are not uniform across these contexts (FeldmanHall et al., 2014). Importantly, recent work has successfully compared children's punishment across second- and third-party contexts in a within-subject design (Bernhard et al., 2020), finding that second-party punishment emerges before third-party punishment yet both are more sensitive to outcomes than intent. This work lays the groundwork for future inquiries into the extent to which children's motivations for punishment differ across these contexts. Fourth, while we argue that our question designed to establish whether children believed that their partners were real was somewhat flawed, it is of course important to consider the possibility that a portion of our sample did not believe that the absent peers were real. If this were the case, why did belief not affect their behavior (Table S4), particularly in a context in which intervention was costly? Future work should seek to examine children's beliefs in absent peers in a more rigorous way and to explore how these beliefs may affect their propensity to intervene against norm violations. Finally, it is essential to note that our work focused on children from a single population in the USA. Not only is it important to understand the extent to which children's punishment decisions and their underlying motivations vary across different societies (House et al., 2020) but it is also important to note that this sample is likely not representative of the majority of human populations, either contemporary or historical, in that it is Western Educated Industrialized and Rich (Amir & McAuliffe, 2020; Henrich, Heine, et al., 2010; Nielsen et al., 2017). Future work on this topic should seek to test children's motivations for third-party intervention across a wide range of cultural contexts to examine the scale and sources of variation of this aspect of children's social development.

To conclude, our findings highlight the motivations supporting third-party intervention against selfishness. Children appear to be relatively more motivated by retribution than restoration when they

encounter a selfish perpetrator, a motivation we do not see in the context of generosity. Older children, but not younger children, additionally punish unfairness regardless of whether it stems from selfishness or generosity, further highlighting late childhood as a period in which children show high levels of compliance with—and enforcement of—an equality norm. More broadly, our finding that children are primarily retributive in response to selfish behavior is consistent with theories suggesting that punishment is an important means of promoting good behavior in those who have done wrong.

ACKNOWLEDGEMENTS

We thank the families who participated in this study and the research assistants who helped collect and code these data. We are especially grateful to Shaina Coogan, Terri Frasca, Gorana Gonzalez, Jillian Jordan, Despoina Lioulu, Carleen Liu, and Rachel Yost-Dubrow, as well as the students in Y. Dunham's research methods course in Spring 2014 and their TA, Jonathan Kominsky. We thank Scott Mackie for his help designing the apparatus. Finally, we would like to acknowledge the generous support of the John Templeton Foundation and Florida State University's Philosophy and Science of Self Control initiative (#56036).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data and R analysis code that support the findings of this study are available from the corresponding author upon reasonable request. Upon publication, data and R code will be deposited in a public repository (e.g., OSF, Dryad; https://osf.io/4av2s/?view_only=c665ed5e83c24e609c34e45541b8afd2).

ORCID

Katherine McAuliffe https://orcid.org/0000-0002-4230-0250
Yarrow Dunham https://orcid.org/0000-0002-4265-4438

ENDNOTE

[1] We acknowledge a distinction between purely retributive motives (so called "just deserts") and deterrence motives. Here we do not make this distinction and use the term *retribution* more broadly to refer to intervention directed towards the transgressor, which stands in contrast to *restoration*, which is directed towards the victim.

REFERENCES

Amir, D., & McAuliffe, K. (2020). *Cross-cultural, developmental psychology: Integrating approaches and key insights*. Evolution and Human Behavior.

Bernhard, R., Martin, J., & Warneken, F. (2020). Why do children punish? Fair outcomes matter more than intent in children's second- and third-party punishment. *Journal of Experimental Child Psychology*. https://doi.org/10.1016/j.jecp.2020.104909

Boyd, R., & Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, *13*, 171–195.

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish?: Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*, 284–299.

Chavez, A. K., & Bicchieri, C. (2013). Third-party sanctioning and compensation behavior: Findings from the ultimatum game. *Journal of Economic Psychology*, *39*, 268–277.

Darley, J. M., & Pittman, T. S. (2003). The Psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, *7*, 324–336.

Dobson, A. J., & Barnett, A. (2008). *An introduction to generalized linear models*. CRC Press.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.

FeldmanHall, O., Sokol-Hessner, P., Van Bavel, J. J., & Phelps, E. A. (2014). Fairness violations elicit greater punishment on behalf of another than for oneself. *Nature Communications*, *5*, 1–6.

Gummerum, M., & Chu, M. T. (2014). Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition*, *133*, 97–103.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., & Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, *327*, 1480–1484.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., & Lesorogol, C. (2006). Costly punishment across human societies. *Science*, *312*, 1767–1770.

Herrmann, B., Thoni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science*, *319*, 1362–1367.

House, B. R., Kanngiesser, P., Barrett, H. C., Yilmaz, S., Smith, A. M., Sebastian-Enesco, C., Erut, A., & Silk, J. B. (2020). Social norms and cultural diversity in the development of third-party punishment. *Proceedings of the Royal Society B*, *287*(1925), 20192794.

Jordan, J., McAuliffe, K., & Rand, D. (2016). The effects of endowment size and strategy method on third party punishment. *Experimental Economics*, *19*, 741–763.

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 12710–12715.

Kenward, B., & Östh, T. (2015). Five-year-olds punish antisocial adults. *Aggressive Behavior*, *41*(5), 413–420.

Leibbrandt, A., & López-Pérez, R. (2012). An exploration of third and second party punishment in ten simple games. *Journal of Economic Behavior & Organization*, *84*(3), 753–766. https://doi.org/10.1016/j.jebo.2012.09.018

Lergetporer, P., Angerer, S., Glatzle-Rutzler, D., & Sutter, M. (2014). Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(19), 6916–6921. https://doi.org/10.1073/pnas.1320451111

Marshall, J., Gollwitzer, A., Wynn, K., & Bloom, P. (2019). The development of corporal third-party punishment. *Cognition*, *190*, 221–229.

McAuliffe, K., Blake, P. R., Steinbeis, N., & Warneken, F. (2017). The developmental foundations of human fairness. *Nature Human Behaviour*, *1*, 0042.

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, *134*, 1–10.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2015). Restorative justice in children. *Current Biology*, *25*, 1731–1735.

Salali, G. D., Juda, M., & Henrich, J. (2015). Transmission and development of costly punishment in children. *Evolution and Human Behavior*, *36*, 86–94.

Vaish, A., Missana, M., & Tomasello, M. (2011). Three-year-old children intervene in third-party moral transgressions. *British Journal of Developmental Psychology*, *29*(1), 124–130.

Van de Vondervoort, J. W., & Hamlin, J. K. (2015). Young children remedy second- and third-party ownership violations. *Trends in Cognitive Sciences*, *19*, 490–491.

Yudkin, D. A., Van Bavel, J. J., & Rhodes, M. (2019). Young children police group members at personal cost. *Journal of Experimental Psychology: General*, *149*(1), 182–191.

SUPPORTING INFORMATION

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

How to cite this article: McAuliffe K, Dunham Y. Children favor punishment over restoration. *Dev Sci*. 2021;24:e13093. https://doi.org/10.1111/desc.13093