BRIEF REPORT

Does third-party punishment in children aim at equality?

Young-eun Lee and Felix Warneken

Department of Psychology, University of Michigan

East Hall, 530 Church St., Ann Arbor, MI, 48109, USA

Author Note

Correspondence concerning this article should be addressed to Young-eun Lee at yl4898@columbia.edu. We pre-registered our hypotheses and analyses (<https://aspredicted.org/up4q5.pdf>). All data and protocols are available through the Open Science Framework: <https://osf.io/3v8zw/>.

Acknowledgements

This research was funded by a National Science Foundation CAREER Grant (#1760238) to FW. We thank research assistants for their help with data collection. We also thank Sebastian Grüneisen for helpful comments on the previous version of this manuscript, Amy Nowack for proofreading the manuscript and CSCAR at the University of Michigan for statistical advice.

This paper has been accepted at Developmental Psychology.

Abstract

Third-party punishment has been regarded as an important mechanism to promote fairness. While previous research has shown that children aged 6 and older punish unfair behaviors at a personal cost, it is unknown whether they actually intend to establish equality or whether equality is a mere byproduct of punishment. In this pre-registered study, $N = 60$ 5-to-9-year-olds witnessed that an agent made unfair resource allocations to a peer. Children could then decide not only whether to punish but also how much to punish. We found that with age, children's intervention is more likely to equalize outcomes between third parties (e.g., turning 3:1 into 1:1). In conclusion, the egalitarian motive to reduce differences in payoffs could underlie children's punishment over development.

Keywords: fairness, third-party punishment, development

BRIEF REPORT: Does third-party punishment in children aim at equality?

Fairness norms guide how resources should be divided between individuals (Deutsch, 1975). Nonetheless, sometimes people act selfishly, making it necessary to put mechanisms into place that deter selfish acts and maintain cooperation (Fischbacher, Gächter, & Fehr, 2001; Raihani & Bshary, 2019; Trivers, 1971). One such mechanism is *third-party punishment* (TPP), i.e., the act of incurring a cost to punish others even when the agent is an unaffected third party, which has been observed frequently in adults with behavioral economic paradigms (Fehr & Fischbacher, 2004; Henrich et al., 2006; Nelissen & Zeelenberg, 2009; Yamagishi et al., 2017). TPP has been claimed to be important for the enforcement of social norms and cooperation in adults (Fehr & Fischbacher, 2003; Gurerk, Irlenbusch, & Rockenbach, 2006). In fact, it is considered as an index of people's concern for norms — such as fairness norms — because punishers pay a cost to enforce these norms even though they are an unaffected third-party (McAuliffe et al., 2017). For these reasons, TPP is important to assess children's genuine concern for fairness.

Developmental studies have begun to trace the developmental trajectory of TPP in children. Children's TPP against unfairness has been found in children aged 6 and older (Gummerum & Chu, 2014; House et al., 2020; Jordan, McAuliffe, & Warneken, 2014; McAuliffe, Jordan, & Warneken, 2015). For example, in McAuliffe et al. (2015), children were shown how an absent child (hereafter divider) allocated 6 Skittles between the self and another absent child (hereafter recipient). The divider made either fair (3 for the self, 3 for the recipient) or unfair allocations (6 for the self, 0 for the recipient). As a third-party observer, children could either accept the allocation, which meant that the Skittles were distributed the way the divider had allocated them, at no cost to the participant. The alternative was for the participant to pay

one of their own Skittles to reject the divider's allocation, in which case all 6 Skittles were thrown away and became inaccessible to everyone. Therefore, rejection serves as third-party punishment, resulting in a 0:0 for the divider and the recipient. McAuliffe et al. (2015) found that 6-year-olds rejected unfair allocations more often than fair allocations, whereas 5-year-olds showed a similar, but less reliable pattern of punishment. This study suggests that TPP against unfairness develops around age 6.

Although these earlier studies provide insight into the development of TPP, one critical question remains. In previous studies, children's punishment was binary (House et al., 2020; Jordan et al., 2014; McAuliffe et al., 2015). For example, children could stay with the divider's original allocation by accepting it or remove everything by rejecting it. Consequently, punishment automatically resulted in equality between two individuals (0:0). Therefore, it is unclear whether children genuinely aimed for equality when they chose to punish or equality was a byproduct of their punishment decisions. For example, it is possible that children punish selfish dividers to see the person suffer (Marshall, Yudkin, & Crockett, 2020; Twardawski & Hilbig, 2020) or to avenge the victim without an intention to restore equality. If this was the case, the equality that resulted from punishment was not the main intention, but only a side-effect of the child's goal to punish.

Here we examine whether children punish with the aim to create equality. We start with the notion that at least in adults, TPP has been identified as a potential mechanism to enforce norms. Therefore, if children's TPP is motivated to create equality, children should punish in a way that reduces inequality among third parties. Alternatively, if their TPP is driven by a self-centered motive such as spite, competition (Raihani & Bshary, 2019) or a desire to watch the deserved punishment enacted (Mendes et al., 2018), the focus would be on inflicting costs on

others, without consideration on whether punishment reduces inequality. To date, whether children use TPP to promote equality has not been measured directly. The current study is a test of children's norm-based punishment by assessing the distributional end-state their intervention pursues.

To examine whether children intervene with the goal to create equality, we presented participants with three (pre-programmed) allocations between two peers that were represented as two avatars on a computer screen: fair allocations (2:2), mildly unfair allocations (3:1) and extremely unfair allocations (4:0). Critically, our participants were in the role of a third-party and were free to choose how many coins they wanted to take away from which individual. Children therefore decided not only whether to punish but also on the degree of punishment.

We tested the hypothesis that children use punishment to establish equality against several other possible outcomes. Here we focus on how children would respond to 3:1 allocations which are the critical trials that allow us to disentangle different hypotheses¹. Based on findings that by school-age, children from the US gravitate towards equal sharing (Blake & McAuliffe, 2011; Shaw & Olson, 2012), we predicted that between 7 and 8 years of age, children would use their punishment to balance the scales between two third parties (e.g., turning 3:1 into 1:1). However, several alternative outcomes are plausible. For example, children might punish unfair allocations more often than fair allocations, but not yet be able to use punishment to restore equality (by for example turning 3:1 into 2:1). Another possibility is that children might be motivated to avenge the recipient by over-punishing the selfish divider (e.g., turning 3:1 into

¹ We included 4:0 allocations to replicate prior research (e.g., McAuliffe et al., 2015), but these trials cannot distinguish the different hypotheses that are the focus of our study. In fact, the hypothesis that TPP is motivated by fairness concerns as well as the non-fairness hypotheses (the over-punishment motivated by moral outrage without concerning fairness, punishment motivated to win against other players specified in the introduction) predict the same outcome in 4:0 trials, namely that children will turn 4:0 into 0:0.

0:1) without considering how their punishment tilts the scales in the opposite direction. Another possibility is that children do not actually punish others for their unfairness, but want to deprive others of rewards with the competitive goal of ending up with more resources than others. If this is the case, children should punish fair as well as unfair allocations, and should take all coins away from the recipient as well as the divider (e.g., turning 2:2 or 3:1 into 0:0). Our study was designed to assess these different possibilities.

Method

Participants. Our final sample were $N = 60$ 5- to 9-year-old children ($M = 88.47$ months, range = 61 - 119 months, $n = 12$ in each age group, 30 male, 30 female). Children were tested at a museum in the Midwest of the US. Demographic information such as race, education and income could not be obtained as per the rules of the museum. Four additional children were excluded because of failure to correctly answer at least one of the comprehension checks (2), parental interference (1), or parental report of their child having autism (1).

With a final sample of $N = 60$, we conducted $n = 1000$ simulations using the package ‘simr’ (Green & MacLeod, 2016) with alpha set at 0.05, allowing us to estimate power to detect each fixed effect from our final model. We ran separate power analyses for each fixed effect we were interested in. The results indicated that we had enough power (0.995) to detect an interaction between age and allocation on the rejection rate. We also had enough power to detect the main effects of age and allocation on the rate of creating equality, resulting in values of 0.998 and 1.00, respectively.

This study was approved by the institutional review board at the University of Michigan (IRB protocol number: HUM00139220, IRB protocol title: Third-Party Interventions).

Figure 1. (A) Computer screen display of the coin game during the test phase.

This example shows a 3:1 allocation made by the divider (Pearl on the top left corner) to the recipient (Callie on the top right corner). Children as a third-party observer at the bottom could push either their own coin above the black basket (rejection) or the gray button (acceptance). (B) Children's cost and possible outcome by allocation type and their intervention decision.

Experimental design and procedure. After parents gave written consent, children sat at a table with the study apparatus while the parents watched passively from a few steps away. A female experimenter introduced the computer game referred to as the "coin game" and explained that players could collect virtual coins to later exchange for prizes. During a *prize introduction*, children learned that the more coins they have during the coin game, the more and the better

prizes they would be able to choose afterwards (see section 1.1. on p.1 in supplementary materials).

In the subsequent *practice phase*, the experimenter introduced the two other players in the game by stating that they were children of the same age and gender at another museum, who are currently connected online. In reality, the decisions of the two other players were computer-programmed. The experimenter introduced the role of the *divider* and the *recipient*. Children were told that the divider has four coins and can give any number of coins to the recipient. The divider then made one of three allocations: (a) 2 for the self and 2 for the recipient, (b) 3 for the self and 1 for the recipient and (c) 4 for the self and 0 for the recipient. The recipient was a passive player who could only accept the divider's allocation.

After introducing the roles, children watched on the screen how the divider produced different allocations and practiced their role as a third-party punisher. After the divider made an allocation, children could press either the gray button or their own coin above their basket (Figure 1). If they pushed the gray button (acceptance), the four coins went into each player's basket just the way the divider allocated the coins, and the child's own coin went back into their own basket. That is, acceptances incurred no cost to the child.

Alternatively, if children pushed their own coin above their own basket (rejection), they could decide which other coins they wanted to take away from either the divider or the recipient or both players. When children pushed a coin, a vacuum appeared at the top of the screen and sucked up the coin, such that no one could keep it. Any remaining coins on the screen that children choose not to take away from players went into each player's basket. Therefore, in our game, children could be flexible about who to punish (divider, recipient or both) and the number of coins they want to take away. Critically, regardless of allocation, children had to pay their one

coin first to take coins away from others. Rejection rate refers to how often children paid their cost to take coins away from other players, which is the binary measure used in prior studies as well (House et al., 2020; McAuliffe et al., 2015). Unlike the prior studies, however, a new feature in the current study was that children could decide the number of coins they would like to take away from other players, making punishment a continuous measure.

There were four practice trials in total. Children practiced four possible outcomes of each button (accept vs. reject) in each allocation (fair vs. unfair). The experimenter asked comprehension checks about the consequence of each button and whether each button required the payment of the child participant's coin or not. All children included in the data analysis passed these comprehension checks.

After the practice trials, to make children believe that the other players were real, the experimenter pretended to call the other players on speakerphone and checked if they were ready to play the game. In reality, a confederate answered the phone call. Upon the completion of the study, the experimenter left and a secondary experimenter asked children whether they thought the players were real or pretend. We found that 78% of children (47 out of 60) said the players were real.

During the subsequent *test phase*, children played 6 rounds as a third-party observer in total. Children received 20 coins as their initial endowment (coins that dropped into their basket on the screen). Children's initial endowment was chosen based on prior research that used a costly third-party punishment task (McAuliffe et al., 2015), giving children an endowment that is substantially larger than the coins that are at play in a given trial. Each child played 2 rounds of 2:2 allocations, 2 rounds of 3:1 allocations and 2 rounds of 4:0 allocations, presented in a pseudo-random order with the restriction that two identical allocations were not presented

consecutively. The divider and recipient were different from those in the practice trials (with different names). The role of each player remained the same throughout the test trials.

After the test phase, to assess whether children's numeric understanding affected their performances in the coin game, we administered a non-social numeric task. Children saw four triangles on the computer screen that were identical to the three allocations in the coin game and were asked to make both sides equal (see section 1.2. on p. 2 in supplementary materials). Children were able to create equal numbers of triangles in 96% of the trials. Thus, any potential age-related sensitivity to equality in the coin game cannot be attributed to their numeric inability to make both sides equal.

We counterbalanced the order of test trials, practice trials, deception check questions, numerical task trials, the other player's identity, and the type of unfair allocation during practice.

Data coding and analyses. Children's responses were recorded by GameMaker Studio (<https://www.yoyogames.com>) and later entered into a spreadsheet by independent coders. All statistical analyses were conducted with R statistical software (R version 3.5.2; R Core Team, 2018).

We pre-registered our hypotheses and analyses before data collection (<https://aspredicted.org/up4q5.pdf>)². All data and protocols are available through the Open Science Framework: <https://osf.io/3v8zw/>.

We analyzed the rejection rate and the rate of creating exact equality with Generalized Linear Mixed Models (GLMM). Our full model included age (in months), allocation (2:2, 3:1 and 4:0) and the interaction of age and allocation as fixed effects, as well as subject ID as a

² To account for the complexity of the data and compare different hypotheses that were suggested by reviewers, we ran several additional statistics beyond the pre-registered analyses. These analyses proved to be highly informative and were thus included in the main text while the pre-registered analyses are included in sections 2.4., 2.5., 2.6. on p. 10-13 in the supplementary materials.

random effect. We also analyzed gender, but did not find any systematic differences (see section 2.12. on p. 25 in supplementary materials). We first compared the full model with a null model that included only subject ID as a random intercept. If the full model provided a significantly better fit to the data than the null model, we then conducted hypothesis-driven tests to examine the role of individual predictors by sequentially dropping them from the full model and assess changes in model fit using Likelihood Ratio Tests (LRTs).

Results

When do children punish? We assessed whether children's decision to intervene was influenced by age and allocation. A full GLMM on children's rejection (0 = acceptance, 1 = rejection) provided a better fit to the data than the null model (LRT, $\chi^2(5) = 36.11, p < .001$), showing that the predictors combined significantly affected children's intervention decisions. A model comparison between the full model and a more parsimonious model showed a significant interaction between allocation and age (LRT, $\chi^2(2) = 22.46, p < .001$). Specifically, as shown in Figure 2A, estimates for children's rejection rate of unfair allocations increased with age from around 20% of trials at age 5 to over 70% of trials by age 9 (LRT, $\chi^2(1) = 9.20, b = 0.04, SE = 0.01, p < .01$ for 3:1 and $\chi^2(1) = 13.15, b = 0.05, SE = 0.02, p < .001$ for 4:0), whereas rejection of fair allocations remained low at around 20% of trials overall (and even tended to slightly decrease with age; LRT, $\chi^2(1) = 3.32, b = -0.03, SE = 0.02, p = .07$). In fact, pairwise comparisons showed that the trajectories of responding to unfair allocations differ both between 3:1 and 2:2 ($b = 0.07, SE = 0.02, p < .001$) and 4:0 and 2:2 ($b = 0.07, SE = 0.02, p < .001$), but not from each other (4:0 vs. 3:1, $b = 0.01, SE = 0.02, p > .62$). In sum, with age, children became more likely to intervene against unfair over fair allocations and punished allocations of 3:1 and 4:0 at similar rates. These results are consistent with the previous work (McAuliffe et al., 2015).

Figure 2. (A) Estimates of children's rejection rate with 95% confidence intervals based on the final model. (B) Estimates of the rate of creating equality with 95% confidence intervals based on the final model. (C) The average number of coins children took away from the divider by allocation type (both accepted and rejected trials included). When children accepted the trial, the number of coins children took away from either the divider or recipient was coded as 0 coins. Error bars represent 95% confidence intervals. (D) The average number of coins children took away from the recipient by allocation type (both accepted and rejected trials included). 4:0 allocations are not included because in these trials the recipient had no coins that could be taken away. Error bars represent 95% confidence intervals. For the average number of coins taken away in rejected trials only, see Figure S5 on p. 12 in supplementary materials.

One could speculate that 4:0 allocations would provoke more punishment than 3:1 allocations. However, we found no significant difference in the rejection rate between 3:1 and 4:0 allocations, suggesting that both types of unfair allocations elicit similar responses in children. Thus, it seems that the relatively smaller deviation from equality is sufficient to induce punishment.

How often do children create exact equality? Next, we examined the rate at which children used punishment to establish perfect equality such that both players ended up with exactly the same number of coins. Here we look at the final outcomes across all trials, aggregating across trials with and without punishment. A full GLMM on children's creation of perfect equality between two other players (0 = inequality, 1 = perfect equality) provided a significantly better fit to the data than the null model (LRT, $\chi^2(5) = 183.39, p < .001$). Model comparisons show significant main effects of allocation (LRT, $\chi^2(2) = 163.51, p < .001$) and age (LRT, $\chi^2(1) = 19.31, p < .001$), but no interaction between age and allocation (LRT, $\chi^2(2) = 2.92, p > .23$). Pairwise comparisons from the main effect of allocation revealed that exact equality occurred more often after fair allocations ($M = 0.93, SD = 0.26$) than in unfair allocations ($b = -4.29, SE = 0.52, p < .001$ for 2:2 vs. 3:1; $b = -3.99, SE = 0.51, p < .001$ for 2:2 vs. 4:0; $b = 0.30, SE = 0.32, p > .34$ for 3:1 vs. 4:0), which is not surprising given that equality already existed in fair allocations.

Importantly, we found a main effect of age, suggesting that across allocations, children became more likely to establish exact equality with age ($b = 0.05, SE = 0.01, p < .001$). As shown in Figure 2B, estimated rates of establishing perfect equality in unfair allocations increased from below 10% of trials at age 5 to over 50% of the trials at age 10. These rates of creating equality did not differ from each other between 3:1 ($M = 0.26, SD = 0.44$) and 4:0

allocations ($M = 0.31$, $SD = 0.46$; $b = 0.30$, $SE = 0.32$, $p > .34$). The effect of age on establishing equality remains even when the 2:2 allocations were not included in the analyses (see section 2.3. on p. 9 in supplementary materials). This analysis provides evidence that with age, children become more likely to create equality between third parties.

How many coins do children take away? One question is *how* children achieved exact equality. For example, exact equality could be established by e.g., turning 3:1 into 1:1 (by only taking away coins from the unfair divider) or 0:0 (by taking coins away from everyone). To assess this, we calculated the average number of coins taken away from each player across all trials, aggregating across trials with and without punishment. In terms of the number of coins children took away from the *divider*, with age, children became more likely to take coins away from the unfair divider (Figure 2C). For instance, 5-year-olds took less than 1 coin from the unfair divider in 3:1 and 4:0 allocations, whereas 9-year-olds took on average 1.25 coins and 2.75 coins in 3:1 and 4:0 allocations, respectively. In contrast, across ages, taking coins away from the fair divider was relatively rare, and the tendency to take coins away from the fair divider decreased even more by age 9. In terms of the number of coins children took away from the *recipient*, this behavior was infrequent overall and declined even more in 9-year-olds (Figure 2D). When we examined the trials in which children intervened (141 out of 360 trials), the patterns across ages were largely similar to those in Figure 2C and 2D except that a minority of 5- and 6-year-olds who rejected 3:1 and 4:0 trials already took a similar number of coins away from the unfair dividers as older children did (see Figure S5 on p. 12 in supplementary materials).

Additionally, the distributional outcomes in 3:1 allocations provide converging evidence (see Figure 3). For example, across ages, children turned 2:2 or 3:1 allocations into 0:0 in only a

minority of the trials (10 to 20%), showing that children rarely took coins away from fair dividers or disadvantaged recipients. Further, across ages, children rarely turned 3:1 into 0:1 (about 10% of the trials), suggesting that children do not over-punish the unfair divider. Rather, with increasing age, they became more likely to turn 3:1 into 1:1, restoring exact equality between the two other individuals. Specifically, at age 9, children's intervention resulted in 1:1 for about a half of the 3:1 allocations (see Figure 3).

Figure 3. (A) Percentage of trials in which children's intervention resulted in 3:1, 1:1, 0:1 or 0:0 between the divider and the recipient in 3:1 allocations. (B) Percentage of trials in which children's intervention resulted in 0:0 or 2:2 between the divider and the recipient in 2:2 allocations. Distributional outcomes of 4:0 allocations were not included because Figure 2B already shows the rate of turning 4:0 into 0:0 (i.e., exact equality). Other distributional outcomes in 2:2, 3:1 and 4:0 allocations were relatively rare and thus were not included in the graphs. See Figure S5 in supplementary materials (p. 12) for the frequency of other distributional outcomes.

Taken together, with age, children take more coins from unfair dividers, while they take fewer coins from fair dividers and recipients, suggesting that they establish equality by directing punishment towards unfair dividers specifically.

Discussion

Our findings demonstrate the development of punishment in which children use punishment to create equality. This was possible because in our task, children were able to express their distributional preference by deciding not only whether to punish, but how much to punish. We showed that with age, (a) children were willing to pay a cost to punish unfairness, and (b) became increasingly interested in creating exact equality. These findings support our hypothesis that with age, children use punishment to restore equality. Our study shows that children use punishment not only to prevent unequal outcomes, but specifically to create equality.

One concern might be that children's punishment reflects a competitive motive to win against others. Specifically, if children's punishment is driven by competition, (a) children should punish fair as well as unfair allocations indiscriminately and (b) their punishment should result in 0:0, taking coins away from the recipient as well as the divider. By contrast, if children's punishment is motivated by a concern to establish equality, (a) children should punish unfair allocations more often than fair allocations and (b) their punishment should aim at unfair dividers, not disadvantaged recipients. Our results do not support the competition account. Specifically, in both 2:2 and 3:1 allocations, creating 0:0 was observed in a minority of trials across ages and became less than 10% of trials at age 9 (see section 2.10. on p. 20 in supplementary materials). Generally, with age, children became less likely to punish fair allocations, suggesting that older children's punishment specifically targeted at unfair

allocations. Furthermore, children were more likely to take away coins from unfair dividers than disadvantaged recipients, suggesting that their punishment was targeted specifically at selfish dividers. Taken together, our results are inconsistent with the competition account and strengthen the hypothesis that with age, children punish to establish equality, providing converging evidence to the earlier findings (e.g., McAuliffe et al., 2015, Jordan et al., 2014).

Another possibility was that children are focused solely on the unfair dividers without considering how their punishment impacts the balance between the two individuals. In fact, they might try to avenge the victim and over-punish in terms of taking away more from the divider than is necessary to create equality (e.g., turning 3:1 into 0:1). Prior work using punishment as a binary measure could not directly address this hypothesis. However, our study using a continuous measure showed that this account is unlikely to explain children's punishment. Children rarely took away all the coins from the unfair divider: When we look at trials in which children punished 3:1 allocations, only 15% of punishments resulted in 0:1 across ages. Especially among 9-year-olds who rejected 3:1 allocations, a vast majority of trials (71%) resulted in 1:1, while only 7% and 14% of the trials were turned into 0:0 and 0:1, respectively. Together, the results do not support the notion that children punish to avenge victims and not caring about an egalitarian outcome. In summary, with age, children's punishment reflects their fairness concern to move allocations towards equality rather than to outcompete others or simply punish unfair individuals.

Previous research also supports the notion that older children's punishment reflects their fairness concern. In Lee and Warneken (2020), 5- to 9-year-olds heard about two third parties who enacted punishment in a setup similar to the current study. They found that with age, children were more likely to evaluate a third-party punisher who *decreased* inequality more

positively than a punisher who *increased* inequality between the divider and the recipient. The finding is consistent with the account that over development, children view punishment as a way to restore fairness. Thus, this provides converging evidence that the current setup evokes fairness concerns in children.

Our study found an interesting developmental change. Five-year-olds intervened in both fair and unfair allocations indiscriminately. By age 9, however, children became highly selective: 9-year-olds punished unfair allocations in a way that could establish equality and rarely punished fair allocations. Converging evidence comes from studies using different paradigms. For instance, Sheskin, Bloom and Wynn (2014) showed that in a first-party context, 5- and 6-year-olds, but not 7- to 10-year-olds, prefer an allocation in which they are better off than their partner (1 for the self : 0 for partner) over a fair allocation (2:2) despite receiving overall less. Similarly, other studies with US children showed that children around age 8 and older give up their own resources to avoid getting more than others, suggesting that older children are averse to inequality even when it is advantageous to themselves (Blake & McAuliffe, 2011; Shaw & Olson, 2012). Together, older children's focus on inequality implies that the egalitarian motive to reduce differences in payoffs could underlie children's punishment as shown in adults (Dawes et al., 2007; Fehr & Schmidt, 1999; Johnson et al., 2009).

One remaining question is why older children did not always punish unfair allocations. Although 9-year-olds were most likely to create equality, they were far from perfect (establishing equality in about 50% of all unfair trials in which children either accepted or rejected). It is likely that the costs of punishment played a role: Neither children nor adults always punish in third-party or second-party punishment paradigms when costs are involved (Bernhard et al., 2020;

Fehr & Fischbacher, 2004; House et al., 2020; Jordan et al., 2014) and costs decrease the rate of punishment in children (McAuliffe et al., 2015).

We want to emphasize that our findings do not necessarily imply that 5-year-olds lack a sense of fairness *per se*. In contexts in which there is no cost to the self, even infants and young children expect equal allocations (Geraci & Surian, 2011; Sloane, Baillargeon, & Premack, 2012; Smith, Blake, & Harris, 2013). Moreover, when children obtain resources from collaboration, 3-year-olds share resources with their partners equally (Hamann et al., 2011; Hamann, Bender, & Tomasello, 2014; Warneken et al., 2011). These studies contrast with the current context in which children were not a recipient of resources and had to pay a cost for third-party intervention. These developmental changes highlight important developmental milestones in children moving from the detection of fairness violations and applying fairness principles in limited situations to a more generalized application of these principles in an impartial and self-sacrificial way.

We would like to address some potential limitations of our paradigm. While the computer game allows for better experimental control and anonymity, it obviously only simulates social interactions. It was encouraging to learn that the vast majority of children said that they interacted with real children, although 13 out of 60 did not. To address this, we confirmed that the results were the same when these children were excluded from analyses, suggesting that they did not bias our findings (see section 2.7. on p. 13 in supplementary materials). Another feature is that because children interacted with the same divider and recipient throughout, they might not have focused on individual allocations but the whole sequence of events. However, supplementary analyses indicated that there were no effects involving trial number or children's intervention decision in a previous trial (see section 2.8. on p. 14 in supplementary materials).

Lastly, our task required a basic understanding of numerical equivalence. The non-social numeric task showed that even our youngest children had no problem making both sides equal, suggesting that the lack of equality restoration in 5-year-olds cannot be attributed to a lack of numerical understanding (see section 1.2. on p. 2 in supplementary materials).

One question for future research is *why* older children showed equality-oriented TPP. It is possible that this reflects a genuine sense of fairness that emerges at this age. However, this might be partly influenced by older children wanting to signal how much they care about fairness. Future research should examine how children's reputational concerns and norms of fairness might interact in the emergence of TPP.

Another question for future research is how an alternative intervention option might affect children's punishment decisions. For example, if children have an alternative option to redistribute coins (turning 3:1 into 2:2), even young children might show an increased rate of creating equality. Thus, it would be important to investigate how having an alternative option such as redistribution or compensation compares to children's use of punishment.

Another critical question is individual differences in children's punishment. Although the current study was not designed to assess individual differences, exploratory analyses looking at individual patterns of responding indicate that some young children showed an advanced sense of fairness earlier compared to their peers (see section 2.11. on p. 22 in supplementary materials for details). Future research could use our paradigm to systematically test individual differences in fairness development.

Lastly, future research should examine children's punishment of norm-violations in contexts other than egalitarian fairness norms. This might include free-riding (Yang et al., 2018), ownership violations (Marshall et al., 2020; Riedl et al., 2015) or social rules (Yudkin, Van

Bavel, & Rhodes, 2020). It is known that in such contexts children often punish with a retributive motive or a motive to teach a lesson (e.g., Marshall et al., 2020). Based on work using hypothetical scenarios (Smith & Warneken, 2016), we would predict that children adjust the degree of punishment to the severity of norm violation. Future research should investigate whether such reasoning applies to children's own costly TPP across different contexts.

References

- Bernhard, R. M., Martin, J. W., & 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*, 200, 104909. <https://doi.org/10.1016/j.jecp.2020.104909>
- Blake, P. R., & McAuliffe, K. (2011). "I had so much it didn't seem fair" Eight-year-olds reject two forms of inequity. *Cognition*, 120, 215–224. doi:10.1016/j.cognition.2011.04.006
- Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9, 378-400. <https://doi.org/10.32614/rj-2017-066>
- Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, 446, 794–796. <https://doi.org/10.1038/nature05651>
- Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues*, 31, 137–149. <https://doi.org/10.1111/j.1540-4560.1975.tb01000.x>
- Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425, 785–791. <https://doi.org/10.1038/nature02043>
- Fehr, E., & Fischbacher, U. (2004). Third party punishment and social norms. *Evolution and Human Behavior*, 25, 63–87. doi:10.1016/S1090-5138(04)00005-4
- Fehr, E., & Schmidt, K. M. (1999). A Theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114, 817–868. <https://doi.org/10.1162/003355399556151>

- Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71, 397–404.
[https://doi.org/10.1016/s0165-1765\(01\)00394-9](https://doi.org/10.1016/s0165-1765(01)00394-9)
- Geraci, A., & Surian, L. (2011). The developmental roots of fairness: Infants' reactions to equal and unequal distributions of resources. *Developmental Science*, 14, 1012–1020.
<https://doi.org/10.1111/j.1467-7687.2011.01048.x>
- Green, P., & MacLeod, C. J. (2016). simr: an R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, 7, 493–498.
<https://doi.org/10.1111/2041-210x.12504>
- 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.
doi:10.1016/j.cognition.2014.06.001
- Gurerk, O., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312, 108–111. doi:10.1126/science.1123633
- Hamann, K., Warneken, F., Greenberg, J. R., & Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature*, 476, 328–331.
<https://doi.org/10.1038/nature10278>
- Hamann, K., Bender, J., & Tomasello, M. (2014). Meritocratic sharing is based on collaboration in 3-year-olds. *Developmental Psychology*, 50, 121–128. <https://doi.org/10.1037/a0032965>
- Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... Ziker, J. (2006). Costly punishment across human societies. *Science*, 312, 1767–1770.
<https://doi.org/10.1126/science.1127333>

House, B. R., Kanngiesser, P., Barrett, H. C., Yilmaz, S., Smith, A. M., Sebastian-Enesco, C., ...

Silk, J. B. (2020). Social norms and cultural diversity in the development of third-party punishment. *Proceedings of the Royal Society B: Biological Sciences*, 287, 20192794.

<https://doi.org/10.1098/rspb.2019.2794>

Johnson, T., Dawes, C. T., Fowler, J. H., McElreath, R., & Smirnov, O. (2009). The role of egalitarian motives in altruistic punishment. *Economics Letters*, 102, 192–194.

<https://doi.org/10.1016/j.econlet.2009.01.003>

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*, 111, 12710–12715. doi:10.1073/pnas.1402280111

Lee, Y., & Warneken, F. (2019). Is punishment for equality?

<https://aspredicted.org/blind.php?x=3iu8ci>

Lee, Y., & Warneken, F. (2020). Children's evaluations of third-party responses to unfairness: Children prefer helping over punishment. *Cognition*, 205, 104374.

<https://doi.org/10.1016/j.cognition.2020.104374>

Lee, Y., & Warneken, F. (2021). Does third-party punishment in children aim at equality?

<http://osf.io/3v8zw>

Lüdtke, D. (2018).ggeffects: Tidy data frames of marginal effects from regression models. *Journal of Open Source Software*, 3, 772. <https://doi.org/10.21105/joss.00772>

Marshall, J., Yudkin, D. A., & Crockett, M. J. (2020). Children punish third parties to satisfy both consequentialist and retributive motives. *Nature Human Behaviour*.

<https://doi.org/10.1038/s41562-020-00975-9>

- McAuliffe, K., Blake, P. R., Steinbeis, N., & Warneken, F. (2017). The developmental foundations of human fairness. *Nature Human Behaviour*, 1, 0042. <https://doi.org/10.1038/s41562-016-0042>
- McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, 134, 1–10. doi:10.1016/j.cognition.2014.08.013
- Mendes, N., Steinbeis, N., Bueno-Guerra, N., Call, J., & Singer, T. (2018). Preschool children and chimpanzees incur costs to watch punishment of antisocial others. *Nature Human Behaviour*, 2, 45–51. <https://doi.org/10.1038/s41562-017-0264-5>
- Nelissen, R. M. A., & Zeelenberg, M. (2009). Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions. *Judgment and Decision Making*, 4, 543–553.
- Raihani, N. J., & Bshary, R. (2019). Punishment: one tool, many uses. *Evolutionary Human Sciences*, 1, e12, 1–26. <https://doi.org/10.1017/ehs.2019.12>
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Riedl, K., Jensen, K., Call, J., & Tomasello, M. (2015). Restorative justice in children. *Current Biology*, 25, 1731–1735. <https://doi.org/10.1016/j.cub.2015.05.014>
- Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General*, 141, 382–395. <https://doi.org/10.1037/a0025907>
- Sheskin, M., Bloom, P., & Wynn, K. (2014). Anti-equality: Social comparison in young children. *Cognition*, 130, 152–156. <https://doi.org/10.1016/j.cognition.2013.10.008>
- Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness? *Psychological Science*, 23, 196–204. <https://doi.org/10.1177/0956797611422072>

- Smith, C. E., Blake, P. R., & Harris, P. L. (2013). I should but I won't: Why young children endorse norms of fair sharing but do not follow them. *PLoS ONE*, 8.
<https://doi.org/10.1371/journal.pone.0059510>
- Smith, C. E., & Warneken, F. (2016). Children's reasoning about distributive and retributive justice across development. *Developmental Psychology*, 52, 1–17.
- Trivers, R. L. (1971). The Evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46, 35–57. <https://doi.org/10.1086/406755>
- Twardawski, M., & Hilbig, B. E. (2020). The motivational basis of third-party punishment in children. *PLoS ONE*, 15, 1–15. <https://doi.org/10.1371/journal.pone.0241919>
- Warneken, F., Lohse, K., Melis, A. P., & Tomasello, M. (2011). Young children share the spoils after collaboration. *Psychological Science*, 22, 267–273.
<https://doi.org/10.1177/0956797610395392>
- Yamagishi, T., Li, Y., Fermin, A. S. R., Kanai, R., Takagishi, H., Matsumoto, Y., ... Sakagami, M. (2017). Behavioural differences and neural substrates of altruistic and spiteful punishment. *Scientific Reports*, 7, 1–8. <https://doi.org/10.1038/s41598-017-15188-w>
- Yang, F., Choi, Y. J., Misch, A., Yang, X., & Dunham, Y. (2018). In defense of the commons: Young children negatively evaluate and sanction free riders. *Psychological Science*, 29, 1598-1611. <https://doi.org/10.1177/0956797618779061>
- Yudkin, D.A., Van Bavel, J. J., & Rhodes, M. (2020). Young children police in-group members at personal cost. *Journal of Experimental Psychology: General*, 149, 182-191.
<https://doi.org/10.1037/xge0000613>