Not All Punishment Is Equal:

The Effect of Punishment Severity on Children's Social Evaluations

Young-eun Lee[1,2] and Larisa Heiphetz Solomon[1]

[1]Department of Psychology, Columbia University

1190 Amsterdam Ave., New York, NY 10027, United States

[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

43 Vassar St., Cambridge, MA 02139, United States

Lee, Y., & Solomon, L. H. (2025). Not all punishment is equal: The effect of punishment severity on children's social evaluations. *Developmental Psychology*, *61*, 311-322. doi: 10.1037/dev0001845

Author Note

Correspondence concerning this article should be addressed to Young-eun Lee at yelee@mit.edu.

Acknowledgments

This project was made possible through the support of grant #61808 from the John Templeton Foundation and National Science Foundation CAREER Award #2044360 to the last author. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation or the National Science Foundation. Materials, data and analysis code are available at https://osf.io/pxtqn/?view_only=e622768e7c154da387449167a9f8ba03. Preregistrations are available at the following URL: https://aspredicted.org/blind.php?x=VTW_BH1 (Study 1) and https://aspredicted.org/C9Y_KKM (Study 2).

Abstract

Punishment is a key mechanism to regulate selfish behaviors and maintain cooperation in a society. However, children often show mixed evaluations about third-party punishment. The current work asked how punishment severity might shape children's social judgments. In two studies, 5- to 10-year-old children heard about a punisher who took different numbers of items from a transgressor and evaluated the punisher's behavior and moral character. In Study 1 (n=68), when the transgression was relatively mild (i.e., unfair sharing), children across ages evaluated taking no items from the unfair sharer ("no punishment") most positively, while evaluating taking three items ("harshest punishment") most negatively. In Study 2 (n=68), when the transgression was more serious (i.e., stealing), younger children evaluated taking two items ("equality-establishing punishment") more positively than older children, while evaluating taking none most negatively. However, children became more likely to evaluate equality-establishing punishment negatively with age. Overall, the current results show that punishment severity is a key factor underlying children's third-party punishment judgments. The current research extends work on moral development by showing how children conceptualize the severity of punishment.

Keywords: fairness, morality, third-party punishment

Public Significance Statements

What messages do children receive when they observe other people's punishment with varying severities? Our work shows that children between five and ten years of age view not punishing most positively when one person shares resources (e.g., cookies) with another person unfairly, suggesting that children do not approve of any degrees of punishment in the unfair sharing context. However, in a context in which one person steals another person's resource, children view punishment that establishes equality between the transgressor and the victim most positively (although this preference decreases with age). These findings suggest that children do not view punishment in a uniform way. They interpret punishment differently depending on punishment severity, transgression context, and age.

Not All Punishment Is Equal:

The Effect of Punishment Severity on Children's Social Evaluations

Punishment of selfish acts functions as a key mechanism to sustain social norms and cooperation in a group (Fehr & Gächter, 2002). Children, as well as adults, punish those who transgress even when they themselves are an unaffected third party (Fehr & Fischbacher, 2004; McAuliffe et al., 2015). For example, children in middle childhood punish an individual who shares resources in an unfair manner between themselves and another person by taking resources away from the unfair individual (Gummerum & Chu, 2014; House et al., 2020; McAuliffe et al., 2015). Also, children as young as 3 years of age punish those who destroy another person's belonging by preventing the transgressor from enjoying a fun activity (Marshall et al., 2020; Yudkin et al., 2020).

While the prior studies described above provide important insight into how children enact punishment behaviorally, less is known about how children evaluate other people's punishment—an important topic of study because children are frequently in a position to observe others' punishment (Marshall & McAuliffe, 2022; Morris et al., 1973). Thus, probing their evaluations of punishment provides information about the social consequences that punishers may face (e.g., positive versus negative responses from others).

The work that does exist on this topic suggests that punishment receives mixed evaluations from children. On the one hand, children seem to evaluate punishment positively and prefer those who punish transgressors. Piaget's classic work on this topic suggested that children in early elementary school believe that violating rules leads to severe punishment and endorse particularly harsh punishments for transgressions that they perceive to be particularly wrong (Piaget, 1932; see also Robinson & Kurzban, 2006, for a more recent view on punishment that is

proportional to the eliciting offense). Another finding indicate that 5-year-olds prefer those who enforce norms on a transgressor (for example by saying, "You aren't allowed to do that. Don't ever do that again") over those who do not enforce norms (Vaish et al., 2016). Additionally, with increasing age, children between 5 and 9 years old prefer punishers over neutral third parties (Lee & Warneken, 2020). These findings suggest that punishers receive positive evaluations from children when compared with non-punishers. In addition to viewing punishers positively in some circumstances, children also endorse punishment as an important mechanism to regulate selfish behaviors. For instance, between 5 and 8 years old, children increasingly report that they prefer to live in a hypothetical world with, rather than without, punishment (Bregant et al., 2016). Thus, elementary-schoolers appear to endorse the necessity of punishment in a society. This result may occur because children view punishment as serving important social functions. For example, 6- to 8-year-olds report that mean people will become nicer after receiving severe punishments such as incarceration as well as minor punishments such as time-out (Dunlea & Heiphetz, 2021). That is, children appear to conceptualize punishment as one factor that can help people attain moral improvement. Together, these results show that children view punishment and punishers positively.

On the other hand, children do not *always* evaluate punishment positively. For example, 5- to 9-year-olds prefer helpers who compensate a victim's loss over punishers (Lee & Warneken, 2020; Liu et al., 2021). Additionally, children perceive punishers to be more aggressive and less prosocial than compensators (Lee & Warneken, 2020), showing that children are more likely to attribute negative traits and less likely to attribute positive traits to punishers compared to helpers. Overall, these results suggest that children's evaluations of punishment are nuanced and mixed. They do not always prefer punishment and often regard it negatively.

The current work identified punishment severity as one factor that might shape how children evaluate punishment and punishers. A recent study (Lee & Warneken, 2022) supports the possibility that children may show sensitivity to punishment severity. In this work, 5- to 9-year-olds observed an individual allocate resources in an unfair manner (e.g., 3 for herself and 1 for her partner). With increasing age, children calibrated the exact amount of punishment needed to restore equality between the two individuals (e.g., turning a 3:1 distribution into a 1:1 split). The finding hints that children do not always exact a high degree of punishment. Instead, they develop an understanding about what the right amount of punishment is and adjust their punishment accordingly. Therefore, it is possible that children apply their developing sense of the right punishment amount when evaluating it. If this is the case, then children would be expected to regard punishment that establishes equality between a transgressor and a victim most positively, while viewing punishment that is too mild or too harsh in less positive terms.

The current work examined how children evaluate punishment of different severity levels, ranging from no punishment to harsh punishment. Children heard a story in which one person divided valued resources in an unfair manner by keeping three items for herself and giving one item to another person. Then, children learned that a third party who watched what happened took zero, one, two, or three resources from the unfair divider and threw them into a trash can, showing "no punishment" (for taking zero items) to "harsh punishment" (for taking three items). Children evaluated how good or bad the third-party punisher was ("person evaluation") as well as how good or bad the punishment itself was ("behavior evaluation").

The current work examined three competing possibilities. First, there may be a positive relation between punishment severity and children’s evaluations of punishment and punishers. That is, the more items punishers take away, the more positive children's evaluations may be.

This prediction stems from prior work showing that children readily enact punishment, even at a personal cost (Gummerum & Chu, 2014; House et al., 2020; Marshall et al., 2020; McAuliffe et al., 2015; Yudkin et al., 2020). Because children are willing to be punishers themselves, and because children generally see themselves as good people (Heiphetz, 2019; Kochanska et al., 2010; Krettenauer, 2022), they may view other punishers positively as well. Further, they may reason that if some punishment is good, then more punishment must be better. Work with adults has suggested that punishment could signal the punisher's dominance (Gordon et al., 2014; Raihani & Bshary, 2019). Also, in some contexts, young children prefer dominant over subordinate individuals (see Thomsen, 2020 for review). Therefore, children may evaluate those who display dominance through punishment positively, and the more items the punisher takes away from the wrongdoer, the more dominance the punisher might signal to children. If this is the case, the data would reveal a positive relationship between punishment severity and children's evaluations of that punishment.

An alternative hypothesis is that a negative relation between punishment severity and evaluations of punishment and punishers would emerge. In other words, the fewer items a third party takes away, the more positively children would evaluate the punishment. This result, suggesting that children judge punishment negatively, would be consistent with prior work showing that children prefer helpers over punishers and attribute more negative personality traits to punishers as compared to helpers (Lee & Warneken, 2020; Liu et al., 2021). Additionally, this possibility is in line with prior work (Eriksson et al., 2016, 2017) showing that adults disapprove of peer punishment and that they disapprove of the punishment more when it is more severe.

Finally, a third possibility is that we would observe an inverted U-shaped relation between punishment severity and its evaluation. That is, children would evaluate the type of

punishment that restores equality between individuals most positively, while evaluating no punishment or harsh punishment most negatively. This finding would be consistent with a recent finding that elementary schoolers become adept at fine-tuning the right amount of punishment to establish exact equality between individuals (Lee & Warneken, 2022). Another way of thinking about this prediction is that children may view equality-establishing punishment as proportional to a transgression that results in unequal outcomes, such as an unfair distribution of resources. This possibility is also consistent with work showing that with increasing age, children between 4 and 10 years of age assign more punitive tasks to those who acted in a more blameworthy way (Smith & Warneken, 2016), suggesting that children's punishment is proportionate to the blameworthiness of their behaviors. A finding consistent with this hypothesis would suggest that fairness concerns (i.e., equalizing resources between people) might motivate children's punishment decisions.

To test among these competing predictions, we recruited children between 5 and 10 years of age. This age range captures a developmental window during which children both enact punishment (Arini et al., 2021; Gummerum & Chu, 2014; House et al., 2020; Lee & Warneken, 2022; Marshall et al., 2020; McAuliffe et al., 2015) and show mixed evaluations of others' punishment (Bregant et al., 2016; Dunlea & Heiphetz, 2021; Lee & Warneken, 2020; Liu et al., 2021; Vaish et al., 2016). Thus, this was a particularly important developmental window to investigate the question of whether differences in punishment severity underlie differing evaluations of others' punishment.

Overview of Current Research

The present work examined how children evaluate punishers and their punishment depending on the punishment's severity (ranging from no punishment to harsh punishment).

Study 1 investigated 5- to 10-year-olds' inferences about individuals who punished one person who distributed resources unfairly to a second person, while Study 2 probed children's inferences in the context of a more severe transgression (stealing). Together, these studies shed light on how children view punishment based on punishment severity across contexts.

Study 1

Study 1 examined how children view punishment and punishers based on the severity of the punishment they delivered. Children saw an unfair resource allocation scenario in which one person kept three valued resources for herself and gave one to another person. Then, a third party took zero, one, two, or three items from the unfair individual. Children then provided moral evaluations of the punisher and the punisher's behavior.

Method

Participants

Our final sample included 68 5- to 10-year-old children (M_{age}=7.60 years, SD_{age}=1.65 years; 38 male, 29 female, 1 non-binary). Parents identified their child as White or European American (57%), Multiracial (16%), Asian or Asian American (9%), Black or African American (3%), or some other race (5%); the remaining children's parents did not report their race. Additionally, 13% of parents identified their children as Hispanic or Latina/o. An a priori power analysis indicated that this sample size was sufficient to detect a medium-sized effect (d=.5) with an alpha of .05 and 80% power (see our pre-registration at https://aspredicted.org/blind.php?x=VTW_BH1).

We recruited participants from a laboratory database based in the United States. We invited families to the current online research by sending a link, and anyone who landed on the page with a child in our targeted age range was eligible to start the testing session. After we

verified webcam recordings from testing sessions, parents of the child participants received a $5 Amazon gift card. Parent-reported zip code indicated that we recruited our sample from multiple states within the United States (see SOM). Based on pre-registered criteria, we excluded 13 additional children, not included in the sample description above, because they were inattentive (2) or because their parents did not provide a recording indicating their consent to participate (11).[1]

Experimental Design and Procedure

In this and the subsequent study, we used unmoderated remote research methods that have been established for data collection with children (Leshin et al., 2021; Rhodes et al., 2020). To begin the study, parents selected the study link, which allowed them to activate their webcam on their laptop or desktop computer. At the beginning of the study, parents provided verbal informed consent and children provided verbal assent. To confirm that participants could see and hear the stimuli presented on their screen, we played a sound of an animal and asked participants to choose an animal picture that matched the sound they heard. This procedure confirmed that all participants were able to see and hear the stimuli. Importantly, we instructed parents not to interfere during the study. After the instructions, children began the study, which consisted of a series of short videos and questions presented via Qualtrics.

Each child watched a total of four trials with varying severity of punishment ranging from taking zero to three resources from an unfair divider; this was a within-subjects variable. At the beginning of each trial, children heard a story about two actors who were gender-matched to the child participant. In the story, one person (hereafter, the divider) had four items (e.g.,

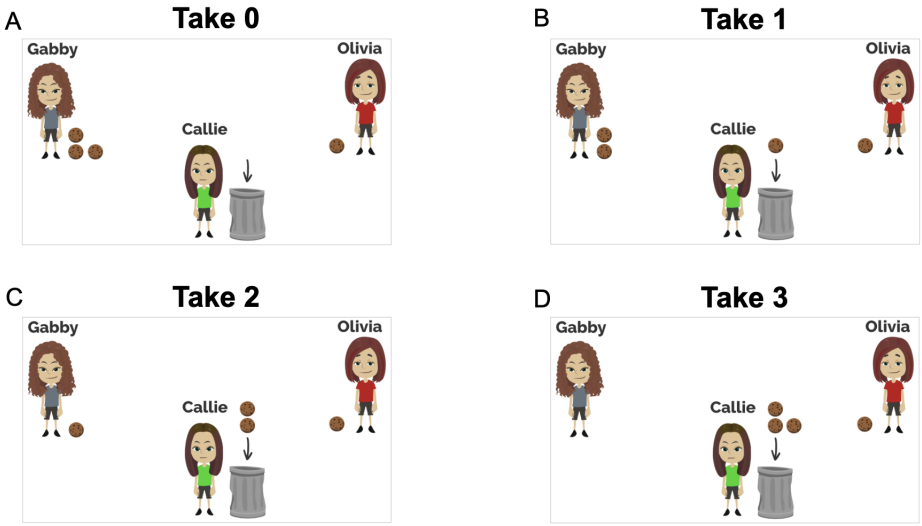
[1] In this and the subsequent study, the patterns of results were similar even after including respondents who were excluded from the main analyses. We did not analyze the data from participants who were excluded for no recording of consent because they did not provide consent for us to do so.

cookies) and could decide how to divide these items between herself and another person. She kept three items for herself and gave one item to another person, resulting in an unfair resource allocation between the two actors. We presented this unfair sharing scenario because even very young children expect fair sharing between individuals (Smith et al., 2013; Sommerville et al., 2013) and because children of the age tested in the current work have avoided receiving more than their peers in prior studies (Blake & McAuliffe, 2011; Shaw & Olson, 2012). Further, children aged six years and older reliably punish those who share unfairly (Arini et al., 2021; Gummerum & Chu, 2014; House et al., 2020; Lee & Warneken, 2022; McAuliffe et al., 2015).

During the test phase, children learned that another actor (the punisher) who watched what happened could decide whether to take items away from the divider and throw them into a trash can (for similar methods, see Lee & Warneken, 2020; Liu et al., 2021). Depending on the type of test trial, the punisher chose to take zero, one, two, or three items from the divider (see Figure 1). We used peer punishment in our scenarios because children might view any degree of punishment positively when an authority figure (e.g., adult) enacted it. For instance, 6- to 7-year-olds believe that compared with peers, adults were more likely to have positive punishment motives (e.g., rehabilitation; Marshall et al., 2022). Therefore, we used peer punishment scenarios to more precisely probe the effect of punishment severity on children's moral evaluations.

Figure 1

Example of Punishment Scenarios in Study 1



Note. In this example, children heard that Gabby kept three cookies for herself, while giving one cookie to Olivia. Then, they heard that a third party, Callie, decided to take zero, one, two, or three cookies away from Gabby and throw them into the trash can.

After hearing the story in each trial, to check if children conceptualized the unfair allocation scenario as a violation of a norm, we asked participants whether it was okay or not okay for the divider to keep three items for herself and give one item to another person. Children reported that the distribution was not okay 52% of the time (for further discussion of this result, see the Results and Discussion sections for this study). We did not provide any feedback on children's response to this question to prevent any biases in subsequent test questions.

In addition, we asked several comprehension check questions. The results from these questions confirmed that a vast majority of the child participants correctly identified the divider (97%), the number of items the divider gave to another person (97%), the punisher (99%), and the number of items the punisher took from the divider (97%). In the minority of trials in which children answered incorrectly, we provided corrective feedback by playing a video with the

correct answer. In line with our pre-registration, we did not exclude children on the basis of incorrect initial answers.

Our dependent measures asked how children evaluated (1) the punisher themselves (i.e., person evaluation) and (2) the punisher's behavior (i.e., behavior evaluation).[2] We measured both person and behavior evaluations because in some contexts, children rate punishers positively but rate their punishment behavior negatively or neutrally (e.g., Lee & Warneken, 2020; see also Uhlmann et al., 2015). Hence, by asking both evaluations, we aimed to disambiguate the extent to which person judgments map on to behavior judgments in the scenarios we tested. Concretely, we assessed person evaluations with two questions—the person's goodness and niceness—and averaged them into one composite score. For example, we phrased the goodness question as follows: "What do you think about [punisher's name]? Do you think [punisher's name] is a good person or a bad person? How [good/bad] is [she/he]? Just a little [good/bad], kind of [good/bad], or very [good/bad]?" Similarly, we assessed behavioral evaluations with two questions—the behavior's goodness and niceness—and averaged responses to these two questions into one composite score. For example, we phrased the goodness question as follows: "What do you think about what [punisher's name] did? Do you think it is a good thing to do or a bad thing to do? How [good/bad] is it? Just a little [good/bad], kind of [good/bad], or very [good/bad]?" In both person and behavior evaluations, children's responses ranged from 0 ("very bad" or "very mean") to 5 ("very good" or "very nice"). The correlations between goodness and niceness measures were high: $r(195)=.86$, $p<.001$ in person evaluations and $r(195)=.88$, $p<.001$ in

[2] In addition to the behavior and person evaluations, we pre-registered and measured children's judgments of the likelihood that the unfair divider would change his/her behavior in the future. Overall, children reported that the unfair divider would change his/her behavior regardless of punishment severity, showing their general tendency to perceive that people can change for the better (i.e., act better in the future than they did in the past). See SOM for more details about these results.

behavior evaluations.[3] Exploratory analyses revealed that children overall evaluated punishers more positively than punishment; see SOM.

In this and the subsequent study, we counterbalanced the punisher's identity across scenarios (e.g., Callie was a punisher in the Take 1 scenario for some participants and in the Take 2 scenario for other participants) and the type of item (cookies, candies, chocolates, donuts) across scenarios. We also coded our survey to randomize the order of punishment scenarios and test questions, with one block consisting of person evaluation questions and another block consisting of behavior evaluation questions. We randomized the order of the two blocks as well as the order of questions within each block.

Data Coding and Analyses

Qualtrics automatically recorded children's responses, and we conducted all statistical analyses with R statistical software (R version 4.1.2; R Core Team, 2021). As pre-registered, here and in Study 2, we analyzed the two dependent measures (person evaluations and behavior evaluations) using Linear Mixed Models (LMM) with the R package "lme4" (Bates et al., 2015). All of our full models included children's age in months as a continuous variable, punishment severity (Take 0, 1, 2, or 3) as a continuous variable[4], and an interaction between age and punishment severity as fixed effects, plus subject identity (ID) as a random effect to account for the within-participant nature of the data. 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 assessed changes

[3] We accounted for repeated measured using the R package "rmcorr" (Bakdash & Marusich, 2023) for all correlation analyses in this and the subsequent study.

[4] For exploratory purposes, we tested the same linear mixed models with punishment severity as a categorical variable. This analysis revealed similar patterns as those reported in the main text.

in model fit using Likelihood Ratio Tests (LRTs). All of our full models provided a significantly better fit to the data than the null models (*p*s<.001). All materials, data and analysis code for Studies 1 and 2 are available at https://osf.io/pxtqn/?view_only=e622768e7c154da387449167a9f8ba03.

Results

We examined whether age and punishment severity shaped children's person evaluations.[5] This analysis revealed a significant main effect of punishment severity, $\chi^2(1)=67.12$, $b=-0.73$, $SE=0.08$, $p<.001$ (see Figure 2). To probe this effect, we compared each severity level with each other severity level. Because this analysis included six tests, *p* values needed to be .008 or lower to pass the Bonferroni-corrected significance threshold. Across age, children evaluated the third parties who took zero items ($M=3.90$, $SD=1.46$) more positively than those who took any other number of items, *p*s<.001 for each comparison (Take 0 versus Take 1, Take 0 versus Take 2, and Take 0 versus Take 3). Additionally, punishers who took one item ($M=2.75$, $SD=1.61$) received more positive evaluations compared to those who took three items ($M=1.58$, $SD=1.68$, $t(65)=5.03$, $p<.001$). Children evaluated punishers who took two items ($M=2.42$, $SD=1.77$) more positively than those who took three items, $t(64)=4.00$, $p<.001$. All pairwise comparisons reached significance except for the comparison between Take 1 and Take 2, $t(65)=1.65$, $p=.103$. We did not observe main or interaction effects involving age (*p*s>.688).

[5] As post-hoc analyses, we also tested whether adding punishment severity as a quadratic variable (versus monotonic variable) provided a better fit to the data. A model with main and interaction effects involving age as a *quadratic* variable as well as punishment severity as fixed effects and subject ID as a random intercept did not provide a significantly better fit to the data than a model with main and interaction effects involving age as a *monotonic* variable as well as punishment severity as fixed effects and subject ID as a random intercept. This was true for both behavior evaluations, $\chi^2(2)=5.62$, $p=.060$, and person evaluations, $\chi^2(2)=3.47$, $p=.176$.

We also examined if adding a main effect of gender to the final model provided a significantly better fit to the data compared to our final model without the gender effect. However, in both person and behavior evaluations across studies, we found no evidence indicating gender effects, all *p*s>.294.

Similarly, we assessed whether age and punishment severity shaped children's behavior evaluations. As in the person evaluations, this analysis revealed a significant main effect of punishment severity, $\chi^2(1)=83.18$, $b=-0.85$, $SE=0.08$, $p<.001$. Across age, taking zero items ($M=3.73$, $SD=1.61$) received more positive evaluations than taking any other number of items, ps<.001 for each comparison. Additionally, taking one item ($M=2.06$, $SD=1.51$) received more positive evaluations than taking three items ($M=0.93$, $SD=1.36$, $t(65)=5.30$, $p<.001$). Taking two items ($M=1.95$, $SD=1.71$) also received more positive evaluations than taking three items, $t(63)=5.16$, $p<.001$. That is, as in the person evaluations, all pairwise comparisons reached significance except for the comparison between Take 1 and Take 2, $t(64)=0.71$, $p=.483$. Additionally, as in the person evaluations, we did not observe any significant main effects or interactions involving age (ps>.163).

We initially expected that children would report unfair sharing as unacceptable (e.g., Smith et al., 2013). However, children reported that the distribution was not okay only 52% of the time. It is possible that the negative evaluations of punishment were due to the fact that children did not view unfair sharing as a transgression. To address this possibility, we examined whether the pattern of results among children who viewed unfair sharing as not okay was similar to the pattern of results across all participants.

As post-hoc analyses, we analyzed responses only from the 26 children (out of 68 participants) who consistently reported across all four trials that it was not okay for the divider to keep three items for herself and give one item to another person. In response to person evaluations, these 26 children evaluated the Take 0 scenarios ($M=3.38$, $SD=1.80$) most positively, followed by Take 2 ($M=3.17$, $SD=1.81$), Take 1 ($M=3.14$, $SD=1.76$), and Take 3 ($M=2.14$, $SD=1.99$). We found a significant main effect of punishment severity, $\chi^2(1)=6.21$, $b=-$

0.37, *SE*=0.14, *p*=.013. Pairwise comparisons involving Take 3 all reached significance: *t*(24)=2.41, *p*=.024 for the comparison between Take 0 and Take 3; *t*(24)=2.57, *p*=.017 for the comparison between Take 1 and Take 3; *t*(24)=2.72, *p*=.012 for the comparison between Take 2 and Take 3. Other pairwise comparisons did not reach significance, *p*s>.673. We did not observe significant main effects or interactions involving age (*p*s>.249). In sum, these children evaluated Take 3 significantly more negatively than other scenarios, but we did not observe significant differences among Take 0, Take 1, and Take 2.

We found similar patterns for behavior evaluations in these children. They rated the Take 0 scenarios (*M*=3.00, *SD*=1.90) most positively, followed by Take 2 (*M*=2.87, *SD*=1.89), Take 1 (*M*=2.54, *SD*=1.73), and Take 3 (*M*=1.40, *SD*=1.81). We found a significant main effect of punishment severity, $\chi^2(1)=7.95$, *b*=-0.44, *SE*=0.15, *p*=.005. Pairwise comparisons involving Take 3 all reached significance: *t*(24)=3.06, *p*=.005 for the comparison between Take 0 and Take 3; *t*(24)=2.97, *p*=.007 for the comparison between Take 1 and Take 3; *t*(24)=3.90, *p*<.001 for the comparison between Take 2 and Take 3. Other pairwise comparisons did not reach significance, *p*s>.139. We did not observe significant main effects or interactions involving age (*p*s>.121). Hence, as in person evaluations, these children evaluated Take 3 significantly more negatively than other scenarios, but there were no significant differences among Take 0, Take 1, and Take 2.

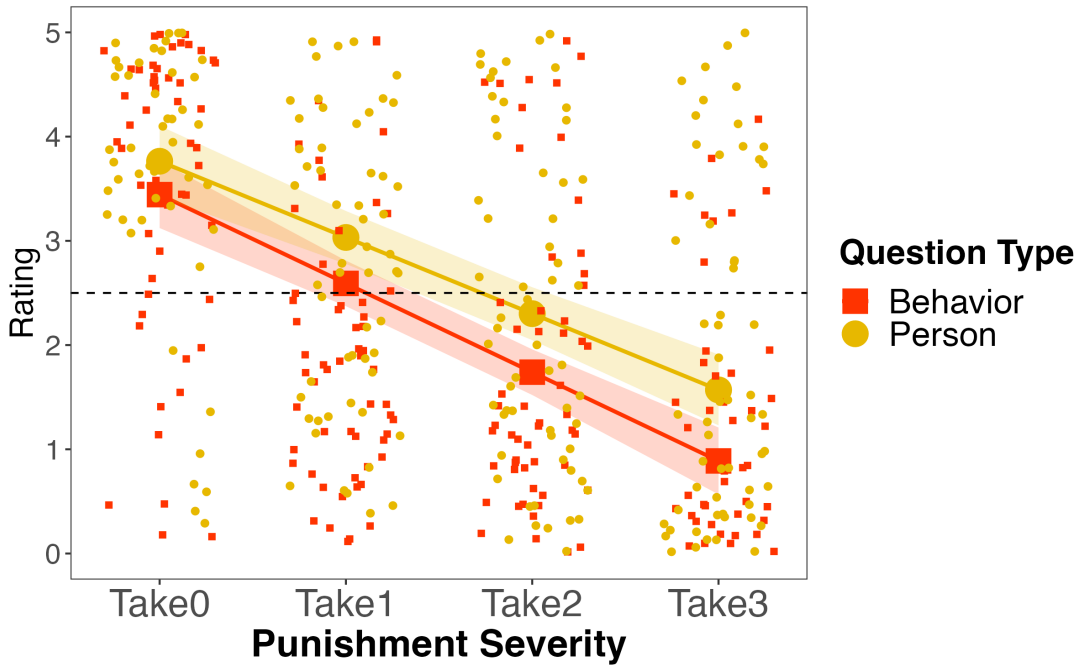
Overall, children who consistently perceived the unfair resource allocation as a transgression viewed no punishment (the Take 0 condition) relatively positively and severe punishment (the Take 3 condition) relatively negatively. These results align with the patterns obtained among the full sample, although the significance levels were closer to the .05 threshold, perhaps due to the smaller sample size in these analyses. Although the patterns of results were

similar in the full sample and the sub-sample of children who consistently reported that the actor's behavior was a transgression, these results require caution because the sub-sample analyses did not include the number of children required by our power analysis and because we did not pre-register these tests.

To examine the degree to which behavior and person evaluations were similar to each other, we computed a correlation between the two measures among the full sample of 68 participants. We observed a high degree of association, $r(196)=.84$, $p<.001$, suggesting that behavior and person evaluations tended to converge.

Figure 2

Children's Social Evaluations by Punishment Severity and Question Type in Study 1



Note. We plotted data points using jittering to provide a better view of overlapping data points. On the Y-axis, 0 indicates the most negative evaluations (e.g., "very bad", "very mean"), while 5 indicates the most positive evaluations (e.g., "very good", "very nice"). The dashed line indicates a neutral valence of 2.5. The confidence bands represent 95% CI.

Discussion

The goal of Study 1 was to assess how children evaluate punishers and punishment based on punishment severity in the context of unfair sharing. This study found a negative relation between punishment severity and its evaluations regardless of age. Specifically, children viewed the third parties who did not punish at all most positively, followed by those who took one item, two items, and three items. The more items the punisher took away from the unfair individual, the less positively children evaluated that punisher. This result indicates that in the current context, punishment led to negative social evaluations.

Children's favorable evaluations of people who did not punish at all are surprising in light of work suggesting that children actively enact punishment against unfair sharing (e.g., House et al., 2020; Lee & Warneken, 2022; McAuliffe et al., 2015). One possibility is that children only evaluate punishment favorably when the transgression that elicited the punishment is severe enough. For instance, children as young as three years of age view moral violations (e.g., hitting another person) as more deserving of punishment than violations of social conventions (e.g., wearing pajamas in class; Smetana, 2013). Children in Study 1 might not have perceived violations of fairness norms as rising to that required level of severity (Yucel et al., 2022). This possibility is consistent with our finding in Study 1 that about a half of the time, children reported that it was okay for the divider to keep three items for herself. Further, children might have thought that the divider has possession over the food items and does not have an obligation to share them equally with another person. The nuanced difference in unfair sharing scenarios between Study 1 and prior work on children's punishment against unfair allocations (e.g., House et al., 2020; Lee & Warneken, 2022; McAuliffe et al., 2015) might have contributed to children's perceptions about the divider's responsibility to allocate resources equally. Study 2

addressed this possibility by presenting a more serious moral violation: Taking another person's food without asking its owner.

Study 2

The goal of Study 2 was to examine potential boundary conditions to the effect observed in Study 1. Specifically, we tested whether two manipulations would increase children's positive inferences about punishers. First, instead of the unfair allocation scenario, children saw a stealing scenario. We made this change because 4- to 8-year-olds consider stealing to be a more serious moral violation than unfairness (Yucel et al., 2022). Second, we introduced a cost for punishment to create a stronger test of whether punishment genuinely leads to negative social evaluations in children. Social psychological research with adults suggests that costly punishment signals the punisher's positive traits, such as cooperativeness and trustworthiness (Jordan et al., 2016). Specifically, when a person incurs a cost to deliver punishment, people are more likely to perceive that the punisher is trustworthy and thus evaluate the punisher positively (Balliet et al., 2011; Nelissen, 2008; Raihani & Bshary, 2019). Therefore, introducing costly punishment in Study 2 provided a stricter test of whether children view punishment negatively in general or whether, instead, they may view punishment positively in some contexts.

Method

Participants

Our final sample included 68 5- to 10-year-old children (M_{age}=7.56 years, SD_{age}=1.62 years; 36 male, 32 female). Parents identified their child as White or European American (56%), Multiracial (15%), Asian or Asian American (7%), Black or African American (6%), Native American or Pacific Islander (3%), or some other race (4%); the remaining children's parents did not report their race. Additionally, 10% of parents identified their children as Hispanic or Latino.

We tested the same sample size as in Study 1 (see our pre-registration at https://aspredicted.org/C9Y_KKM), and the recruitment procedure was identical to that earlier study. Based on pre-registered criteria, we excluded 22 additional children, not included in the sample description above, because they participated in Study 1 (6), because their sibling interfered with the session (1), or because their parents did not provide a recording indicating their consent to participate (15).

Experimental Design and Procedure

The experimental procedure in Study 2 was identical to Study 1 except for two changes. First, rather than learning about transgressions involving unfairness as in Study 1, participants in Study 2 saw a more serious transgression in which one person took one item away from another person without asking. In this stealing scenario, two actors both began with two items each. However, one actor (the transgressor) took one item from the other actor (the victim) without asking, resulting in three items for the transgressor and one item for the victim. Thus, the outcomes of the unfair sharing scenario (Study 1) and of the stealing scenario (Study 2) were identical, i.e., both scenarios resulted in a 3:1 split between the two individuals. The only difference was in how the scenarios resulted in this unequal allocation: in Study 1, one person shared unfairly; in Study 2, one person stole from another.

As in Study 1, we asked a norm conceptualization question to confirm if children conceptualized the stealing scenario as a norm violation. Unlike Study 1 where we asked the norm conceptualization question in each trial, in Study 2, we asked the norm conceptualization question only once at the beginning of the first trial to make the study shorter. All children reported that it was not okay to take an item from another person without asking, showing that children viewed the stealing scenario as a norm violation.

Second, children saw two blocks in counterbalanced order (within-subject): (1) a non-costly punishment block and (2) a costly punishment block. As in Study 1, each block consisted of four trials (i.e., Take 0, Take 1, Take 2, and Take 3). The critical difference between the two blocks was whether the punisher had to pay a sticker in order to throw items into a trash can. To introduce the concept of paying a cost, at the beginning of the study, we presented two rooms with different colors. For example, in a costly punishment block, children heard that "in the green classroom, there is a special rule about throwing things into trash can. If someone wants to open a trash can and throw something into it, the person has to pay their sticker into a red box. When the person pays their sticker into the red box, they cannot get it back." In contrast, in a non-costly punishment block, children heard that in the blue classroom, there is no such rule and people do not have to pay a sticker to throw something into the trash can. Additionally, in each punishment trial, children received reminders about the classroom the actors were in and the classroom's rule about whether people had to pay a sticker to enact punishment. We used stickers as a cost to punish because children as young as five years old understood a similar set-up in prior work (McAuliffe et al., 2015).

As a minor change, we assessed both person and behavior evaluations by asking only about their goodness. Because Study 2 included more test trials than Study 1, we dropped the questions about niceness to prevent children from becoming inattentive. In Study 1, there were high correlations between goodness and niceness measures: $r(195)=.86$, $p<.001$ in person evaluations and $r(195)=.88$, $p<.001$ in behavior evaluations.

A majority of participants correctly identified the transgressor (95%), the punisher (96%), the number of items that the punisher took away from the transgressor (95%), and whether the punisher paid a sticker in a given trial (98%). As in Study 1, in a minority of trials in which

children answered incorrectly, we provided corrective feedback and did not exclude participants on the basis of incorrect initial responses. We counterbalanced the order of costly versus non-costly punishment blocks and the room color of these blocks (costly in the blue room versus costly in the green room).

Results

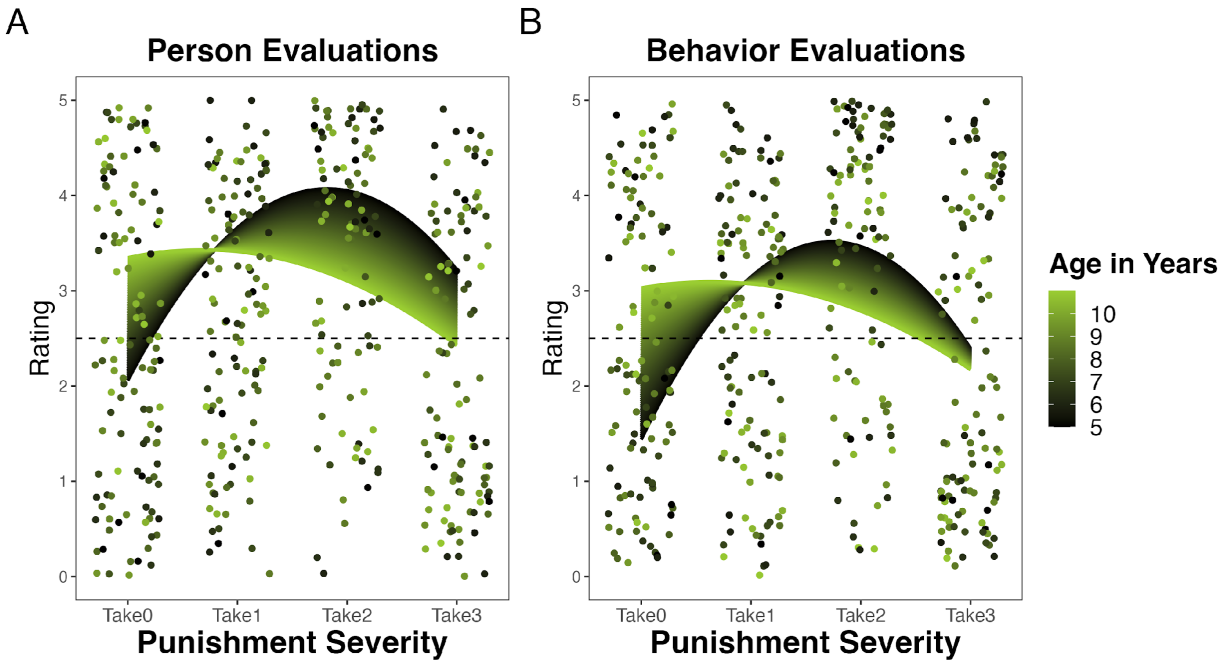
We assessed whether age, punishment severity, and cost (costly punishment versus non-costly punishment) shaped children's person evaluations.[6] This analysis revealed a significant interaction effect between age and punishment severity, χ^2 (2)=15.31, p<.001 (see Figure 3). The interaction effect was significant because with increasing age, children evaluated those who took zero items more positively with marginal significance, χ^2 (1)=3.65, b=0.02, SE=0.01, p=.057, while they viewed those who took two items more negatively, again with marginal significance, χ^2 (1)=3.40, b=-0.01, SE=0.01, p=.065. Evaluations of taking one item or three items did not change significantly with age, ps>.292. That is, in our youngest group of children, i.e., 5-year-olds, Take 2 (M=4.26, SD=1.41) elicited the most positive judgments, while Take 0 (M=2.50, SD=1.85) elicited the most negative judgments. However, in the oldest group of children, i.e.,

[6] In Study 2, although not pre-registered, we entered punishment severity as a quadratic variable because visual inspection of the data suggested that conceptualizing punishment severity as a quadratic variable might be a better fit to the data. In fact, we found that entering punishment severity as a quadratic variable provided a better fit to the data than entering punishment severity as a monotonic variable in person evaluations, LRT, χ^2(4)=45.92, p<.001, and behavior evaluations, LRT, χ^2(4)=45.25, p<.001. However, when we looked at the parameters of the final model with punishment severity as a quadratic variable in person evaluations, the interaction between age and punishment severity as a quadratic variable was marginally significant, b=0.14, SE=0.07, p=.061, while the interaction between age and punishment severity as a monotonic variable was significant, b=-0.25, SE=0.07, p<.001. The results do not affect the way we unpacked the interaction effect reported in the main text but do affect the way we represent the results in Figure 3A. After consulting with a statistician, we decided to plot person evaluations with punishment severity as a quadratic variable because (i) the full model with punishment severity as a quadratic variable provided a significantly better fit to the data than the model with punishment severity as a monotonic variable and (ii) the interaction between age and punishment severity as a quadratic variable was marginally significant, b=0.14, SE=0.07, p=.061, and the current plot could be helpful for readers to understand the pattern of responses. There were no similar issues in behavior evaluations in Study 2.

10-year-olds, Take 0 (*M*=3.45, *SD*=1.19) elicited the most positive judgments, followed by Take 2 (*M*=3.05, *SD*=1.57), Take 1 (*M*=2.90, *SD*=1.48), and Take 3 (*M*=2.50, *SD*=1.32).

Figure 3

Children's Social Evaluations About Punishment by Age and Punishment Severity in Study 2



Note. Overall, 5-year-olds (black line) showed an inverted U-shape relation between punishment severity and its evaluations, rating Take 2 most positively. In contrast, 10-year-olds (green line) tended to show a negative relation between punishment severity and its evaluations, rating Take 0 most positively. (A) Estimated ratings from person evaluations by age and punishment severity based on the model. (B) Estimated ratings from behavior evaluations by age and punishment severity based on the model. We plotted data points using jittering to provide a better view of overlapping data points. On the Y-axis, 0 indicates most negative evaluations (i.e., "very bad"), while 5 indicates most positive evaluations (i.e., "very good"). The dashed line indicates a neutral valence of 2.5. The black line indicates the youngest age we tested (60 months), while the green line indicates the oldest age we tested (132 months).

As in person evaluations, we examined whether age, punishment severity, and cost shaped children's behavior evaluations. We observed a significant interaction effect between age and punishment severity, χ^2 (2)=12.53, *p*=.002. The interaction effect occurred because with

increasing age, children evaluated taking zero items more positively with marginal significance, χ^2 (1)=2.76, *b*=0.02, *SE*=0.01, *p*=.096, while viewing taking two items more negatively with age, χ^2 (1)=4.07, *b*=-0.02, *SE*=0.01, *p*=.044. Evaluations of taking one item or three items did not change significantly with age, *p*s>.278. To put it differently, among 5-year-olds, Take 2 (*M*=4.53, *SD*=0.96) elicited the most positive judgments, while Take 0 (*M*=2.65, *SD*=2.06) elicited the most negative judgments, showing an inverted U-shaped relation between punishment severity and its evaluations. However, in 10-year-olds, Take 0 (*M*=3.15, *SD*=1.27) elicited the most positive judgments, followed by Take 2 (*M*=2.90, *SD*=1.48), Take 1 (*M*=2.00, *SD*=1.26), and Take 3 (*M*=1.95, *SD*=1.47). Therefore, young children perceived punishment that established equality (i.e., Take 2) most positively. However, with age, children viewed punishment more negatively even when it established equality between the transgressor and the victim.

Additionally, we found a significant interaction effect between age and cost in behavior evaluations, *b*=-0.01, *SE*=0.01, *p*=.034. Specifically, with increasing age, children rated non-costly punishment more negatively, χ^2 (1)=5.13, *b*=-0.01, *SE*=0.01, *p*=.023, while their evaluations of costly punishment did not change significantly with age, χ^2 (1)<1, *p*=.670. This finding shows that children are increasingly likely to provide negative judgments of non-costly punishment across development.

As in Study 1, we observed a strong correlation between behavior and person evaluation measures, *r*(460)=.75, *p*<.001, indicating that children showed relatively similar judgments of punishments and of the individuals who delivered those punishments. Thus, as in Study 1, person and behavior evaluations converged.

Discussion

The goal of Study 2 was to test whether children evaluated punishment negatively even in a context involving a relatively serious moral violation and a cost for punishment. This study found an inverted U-shaped relation between punishment severity and its evaluations in younger children. For example, 5-year-olds rated punishment that established equality between individuals (i.e., Take 2) most positively, while rating taking no action against transgressions (i.e., Take 0) most negatively. However, with age, children considered Take 0 more positively, while viewing Take 2 more negatively. These results imply that equality-establishing punishment elicits positive judgments about punishers and the punishments they deliver but that these positive inferences decrease with age. Further, we found that with age, children became more likely to judge non-costly punishment negatively in behavioral evaluations, but not in person evaluations (see General Discussion for a potential explanation for this finding). Overall, our results show that how children view others' punishment changes depending on punishment severity, punishment cost, and children's age.

General Discussion

The current work tested whether children's evaluations of punishment, and the people who deliver it, might vary in part as a function of the punishment's severity. Using an unfair resource allocation scenario, Study 1 found that across age, children showed the most positive evaluations when no punishment occurred and the most negative evaluations when people enacted harsh punishment. Study 2 built on Study 1 by introducing a more serious transgression (stealing) and a cost for punishment. This study revealed that younger children showed the most positive evaluations of equality-establishing punishment and the most negative evaluations of no punishment. However, positive evaluations of equality-establishing punishment declined by the age of ten years. In addition to punishment severity, punishment cost also shaped children's

evaluations. Specifically, with increasing age, children evaluated non-costly punishment more negatively.

Children often punish transgressions (Arini et al., 2021; Gummerum & Chu, 2014; House et al., 2020; Lee & Warneken, 2022; Marshall et al., 2020; McAuliffe et al., 2015), but prior work has revealed mixed evidence about whether children view others' punishment positively (Bregant et al., 2016; Dunlea & Heiphetz, 2021) or relatively negatively compared with compensation (Lee & Warneken, 2020; Liu et al., 2021). Our research revealed that to children, not all punishments are equal. In fact, their evaluations of punishment and punishers were nuanced and varied significantly depending on the severity of punishment. The current research is one of the first studies that examined the role of punishment severity on children's judgments by manipulating severity systematically. This work sheds light on potential motivations underlying children's punishment by showing that the motive to equalize resources between people might drive the development of third-party punishment in young children, especially in a stealing context.

One may wonder whether our measures assessed children's evaluations about equality rather than their evaluations about punishment. However, it is unlikely that children's social evaluations merely reflect their preference for equality. If this alternative explanation accounted for the results of the present work, we should have observed the most favorable evaluations of behaviors that established equality in Study 1; however, this was not the pattern that emerged. Further, older elementary-schoolers endorse and follow norms of fair sharing (Blake & McAuliffe, 2011; Shaw & Olson, 2012) but nevertheless preferred no punishment to equality-establishing punishment in Study 2. Therefore, it does not appear that the results of the current work support an interpretation focusing only on children's preferences for equality.

The present work advances research on moral development by investigating how children conceptualize punishers and punishment. Five- to 10-year-olds in the current research did not appear to view punishment as necessary in an unfair allocation context; rather, they evaluated all degrees of punishment more negatively than no punishment. This result is in line with a recent finding that children around this age consider unfair resource allocations to be less serious than typical moral harms (e.g., hitting; Yucel et al., 2022). However, when the wrongdoing became more serious (i.e., stealing), children—especially younger children—perceived punishment more positively compared to no punishment. In particular, they judged punishment that restored equality between the wrongdoer and the victim most positively, suggesting that children as early as five years of age may conceptualize punishment as a way to rectify existing inequality between two other people (see also Arini et al., 2021; Lee & Warneken, 2022).

Additionally, the current studies have implications for philosophy and legal theory. Scholars in these fields have proposed that punishment can serve multiple functions, including retribution (harming people because they have harmed others; Bradley, 2003), incapacitation (putting people in a situation where they can no longer transgress in the same ways, such as imprisoning them and therefore removing the opportunity to commit further transgressions against non-incarcerated individuals; Zimring & Hawkins, 1995), deterrence (preventing people from transgressing in the future, even after the punishment has ended; Gibbs, 1968), and rehabilitation (improving people's moral character so that they do not transgress in the future; Vitiello, 1990). More recent work has argued that punishment is also expressive, i.e., that another goal of punishment is to communicate information (e.g., Feinberg, 2014; Glasgow, 2015; Primoratz, 1989). This information can focus on the transgressor (e.g., that she is a bad person, Dunlea & Heiphetz, 2020; Kleinfeld, 2016; Yankah, 2004) or on social norms more broadly

(e.g., that the punished behavior is unacceptable, Bregant et al., 2016; Sarin et al., 2021). The current work speaks most directly to this latter theory of punishment by showing that punishment can communicate information about those who deliver it. Our data suggest that in some contexts, children view punishment, especially non-equalizing punishment, as communicating negative information about punishers (e.g., that they are less good people than individuals who do not punish). These results thus extend expressivist theories of punishment by showing that punishment communicates information not just about transgressors and social norms but about punishers as well.

Lastly, our work has translational implications regarding children's experience of punishment by showing that children—especially older children—tend to evaluate punishment negatively. To the extent that punishment elicits any positive responses, those tend to come from punitive acts that establish equality rather than those that may appear too lenient or too severe. These findings may inform adults' responses to children's transgressions. For instance, one child may steal another child's cupcake during a classroom birthday celebration. A teacher may be motivated to take away all snacks from the first child or, in contrast, ignore the behavior and let the two children figure it out for themselves. The current work suggests that the remaining classmates may respond most positively if the teacher instead takes a third approach—namely, returning the stolen cupcake to its rightful owner and establishing equality between the cupcake enthusiasts. By responding in this way, the teacher may communicate that stealing is unacceptable without "going overboard." Additionally, children may perceive punishers and their behaviors as more legitimate if those behaviors align with children's own intuitions (Tyler, 2006). Therefore, responding in a way that sends positive messages to observers while also remedying the harm that occurred may help children see the teacher's actions as fair and to adjust

their future behavior to comply with the norms her punishment established (e.g., to avoid stealing in the future).

Limitations and Future Directions

While the current research extends scientific understanding of punishment, like all research programs, it cannot possibly answer all questions. One open question concerns the extent to which the findings obtained here speak to children's everyday experiences. As discussed above, we presented participants with scenarios in which peers punished each other because children could judge punishment enacted by authority figures (e.g., adults) positively *because* it is enacted by an authority figure, regardless of its severity. Since children sometimes directly punish each other during the course of everyday life (Lamb et al., 1980; Masters & Furman, 1981), this approach also reflects children's lived experiences to some degree. However, children's peer punishment may often be indirect, such as telling adults that a transgression has occurred (Ingram & Bering, 2010; Yucel & Vaish, 2018). Further, children may perceive adults as more prototypical punishers than peers, especially for direct punishment such as removing resources from a transgressor. Thus, future work can investigate how children evaluate direct versus indirect punishment as a function of its severity and how these evaluations may depend on who is delivering the punishment.

Future work can also examine other ways to intervene against transgressions. To build on the existing literature (e.g., Gummerum & Chu, 2014; House et al., 2020; Lee & Warneken, 2020, 2022; Liu et al., 2021; McAuliffe et al., 2015) and allow for systematic manipulations of punishment severity, we presented material punishment. Yet, this design choice might have prevented children from viewing punishment more positively, particularly with older children in Study 2. For instance, children might perceive the act of throwing cookies into a trash can

negatively because of its aggressive nature or because they perceive throwing away food to be wasteful. Also, with age, children may develop an understanding of alternative ways to intervene and come to believe that a peer throwing food away is not an ideal way to rectify the situation. Yet, past research has shown that children themselves are willing to throw material resources such as erasers into a trash can to establish equality between peers (Shaw & Olson, 2012), suggesting that the scenarios in the current work may not have seemed unnatural to children, since children are willing to engage in similar behaviors themselves. However, children could perceive it as more wasteful to throw away food rather than erasers. While the current studies always relied on throwing food away as a measure of punishment, and therefore showed that children evaluate this type of punishment differently depending on its severity, future work can build on this result by investigating other forms of punishment.

Relatedly, future work can manipulate the transgression that resulted in the punishment. Because our main question concerned the role of punishment severity in evaluations of punishment and the people who deliver it, we controlled the underlying transgression. For instance, in Study 2, the transgression always involved stealing, as the degree to which participants perceived a given punishment as severe could vary depending on the underlying transgression (e.g., throwing away someone's resources might appear to be a particularly severe response to a minor accidental transgression and a response that is not severe enough for a particularly damaging offense, like purposefully hitting another person). In fact, we created scenarios where the punishment was aligned with the offense, in the sense that the offense in both studies involved distributions of material resources and the punishment involved controlling access to the same types of resources. Because proportionality is an important component of punishment-related decisions and judgments (Cushman, 2015; Piaget, 1932; Robinson &

Kurzban, 2006), future work can vary the underlying offense to determine effects on evaluations of punishment. Additionally, children's evaluations could vary depending on their perceived transgression severity as well as transgression type. Although children were more likely to report that it is not okay to steal (Study 2) than to share unfairly (Study 1), future work could probe their perceptions of transgression severity in a more fine-grained way (e.g., by asking *how* wrong each behavior is) and test whether responses to this question predict evaluations of punishment.

A final suggestion for future research is to examine the extent to which the current findings generalize to other contexts that are less quantifiable in nature. Based on past research on children's punishment judgments, which has typically involved tangible items (Gummerum & Chu, 2014; House et al., 2020; Lee & Warneken, 2020, 2022; Liu et al., 2021; McAuliffe et al., 2015), the current work used unfairness and stealing scenarios involving quantifiable resources. This approach allowed us to vary punishment severity in a systematic manner. However, other transgression types (e.g., physical or emotional harm) and punishment types (e.g., inflicting emotional suffering) are less tangible and less quantifiable. Given our findings that children—especially younger children—prefer to equalize between the transgressor and victim, it is possible that children would endorse a similar approach in a less quantifiable context as well. However, it may be difficult to determine exactly how to create equality in this context. Future research could examine this possibility by adapting our task with less quantifiable harm and probing children's perceptions of whether it is possible to create equality and, if so, how.

Conclusion

The current work examined how 5- to 10-year-old children evaluate punishers and their behaviors. In Study 1, in the context of a relatively mild transgression (distributing resources unfairly), 5- to 10-year-olds evaluated all degrees of punishment more negatively than no

punishment. Study 2 examined potential boundary conditions to the effect observed in Study 1. Specifically, we tested whether a more serious transgression (stealing) and an introduction of a cost for punishment could make children's evaluations of punishment more positive. In Study 2, younger children showed positive evaluations of equality-establishing punishment. Also, with age, children viewed non-costly punishment negatively. This research extends work on moral cognition and development by revealing that children's evaluations of punishment change dynamically across age depending on the type of transgression, punishment severity, and punishment cost.

References

Arini, R. L., Wiggs, L., & Kenward, B. (2021). Moral duty and equalization concerns motivate children's third-party punishment. *Developmental Psychology, 57*, 1325-1341. https://doi.org/10.1037/dev0001191

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*, 594-615. https://doi.org/10.1037/a0023489

Bakdash, J. Z., & Marusich, L. R. (2023). rmcorr: Repeated measures correlation. R package version 0.6.0. https://CRAN.R-project.org/package=rmcorr

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48. https://doi.org/10.18637/jss.v067.i01

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. https://doi.org/10.1016/j.cognition.2011.04.006

Boseovski, J. J. (2010). Evidence for "rose-colored glasses": An examination of the positivity bias in young children's personality judgments. *Child Development Perspectives, 4*, 212–218. https://doi.org/10.1111/j.1750-8606.2010.00149.x

Bradley, G. V. (2003). Retribution: The central aim of punishment. *Harvard Journal of Law & Public Policy*, *27*, 19-31.

Bregant, J., Shaw, A., & Kinzler, K. D. (2016). Intuitive jurisprudence: Early reasoning about the functions of punishment. *Journal of Empirical Legal Studies*, *13*, 693-717. https://doi.org/10.1111/jels.12130

Cushman, F. A. (2015). Punishment in humans: From intuitions to institutions. *Philosophy Compass*, *2*, 117-133. https://doi.org/10.1111/phc3.12192

Dunlea, J. P., & Heiphetz, L. (2020). Children's and adults' understanding of punishment and the criminal justice system. *Journal of Experimental Social Psychology*, *87*, 103913. https://doi.org/10.1016/j.jesp.2019.103913

Dunlea, J. P., & Heiphetz, L. (2021). Children's and adults' views of punishment as a path to redemption. *Child Development*, *92*, e398–e415. https://doi.org/10.1111/cdev.13475

Eriksson, K., Andersson, P. A., & Strimling, P. (2016). Moderators of the disapproval of peer punishment. *Group Processes & Intergroup Relations*, *19*(2), 152–168. https://doi.org/10.1177/1368430215583519

Eriksson, K., Strimling, P., Andersson, P. A., Aveyard, M., Brauer, M., Gritskov, V., … Yamagishi, T. (2017). Cultural universals and cultural differences in meta-norms about peer punishment. *Management and Organization Review*, *13*(4), 851–870. doi:10.1017/mor.2017.42

Fehr, E., & Fischbacher, U. (2004). Third party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137-140. https://doi.org/10.1038/415137a

Feinberg, J. (1965). The expressive function of punishment. *The Monist, 49*, 397-423.

Gibbs, J. P. (1968). Crime, punishment, and deterrence. *The Southwestern Social Science Quarterly*, *48*, 515-530.

Glasgow, J. (2015). The expressivist theory of punishment defended. *Law and Philosophy, 34,* 601–631.

Gordon, D. S., Madden, J. R., & Lea, S. E. G. (2014). Both loved and feared: Third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals. *PLoS ONE*, *9*(10), 1–10. https://doi.org/10.1371/journal.pone.0110045

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. https://doi.org/10.1016/j.cognition.2014.06.001

Heiphetz, L. (2019). Moral essentialism and generosity among children and adults. *Journal of Experimental Psychology: General*, *148*, 2077-2090. https://doi.org/10.1037/xge0000587

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

Ingram, G. P. D., & Bering, J. M. (2010). Children's tattling: The reporting of everyday norm violations in preschool settings. *Child Development*, *81*, 945-957. https://doi.org/10.1111/j.1467-8624.2010.01444.x

Jordan, J., Hoffman, M., Bloom, P., & Rand, D. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*, 473–476. https://doi.org/10.1038/nature16981

Kleinfeld, J. (2016). Two cultures of punishment. *Stanford Law Review, 68*, 933-1036.

Kochanska, G., Koenig, J. L., Barry, R. A., Kim, S., & Yoon, J. E. (2010). Children's conscience during toddler and preschool years, moral self, and a competent, adaptive developmental trajectory. *Developmental Psychology, 46*, 1320-1332. https://doi.org/10.1037/a0020381

Krettenauer, T. (2022). Development of moral identity: From the age of responsibility to adult maturity. *Developmental Review*, *65*, 101036. https://doi.org/10.1016/j.dr.2022.101036

Lamb, M. E., Easterbrooks, M. A., & Holden, G. W. (1980). Reinforcement and punishment among preschoolers: Characteristics, effects, and correlates. *Child Development*, *51*, 1230-1236. https://doi.org/10.2307/1129565

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. (2022). Does third-party punishment in children aim at equality? *Developmental Psychology*, *58*(5), 866–873. https://doi.org/10.1037/dev0001331

Leshin, R. A., Leslie, S. J., & Rhodes, M. (2021). Does it matter how we speak about social kinds? A large, preregistered, online experimental study of how language shapes the development of essentialist beliefs. *Child Development*, *92*(4), e531–e547. https://doi.org/10.1111/cdev.13527

Liu, X., Yang, X., & Wu, Z. (2021). To punish or to restore: how children evaluate victims' responses to immorality. *Frontiers in Psychology*, *12*, 1–15. https://doi.org/10.3389/fpsyg.2021.696160

Marshall, J., & McAuliffe, K. (2022). Children as assessors and agents of third-party punishment. *Nature Reviews Psychology*, *1*(6), 334–344. https://doi.org/10.1038/s44159-022-00046-y

Marshall, J., Gollwitzer, A., & Bloom, P. (2022). Why do children and adults think other people punish? *Developmental Psychology*, *58*, 1783-1792. https://doi.org/10.1037/dev0001378

Marshall, J., Yudkin, D. A., & Crockett, M. J. (2020). Children punish third parties to satisfy both consequentialist and retributive motives. *Nature Human Behaviour, 5*, 361-368. https://doi.org/10.1038/s41562-020-00975-9

Masters, J. C., & Furman, W. (1981). Popularity, individual friendship selection, and specific peer interaction among children. *Developmental Psychology, 17*, 344–350. https://doi.org/10.1037/0012-1649.17.3.344

McAuliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, *134*, 1–10. https://doi.org/10.1016/j.cognition.2014.08.013

Morris, W. N., Marshall, H. M., & Miller, R. S. (1973). The effect of vicarious punishment on prosocial behavior in children. *Journal of Experimental Child Psychology*, *15*, 222–236. https://doi.org/10.1016/0022-0965(73)90144-6

Nelissen, R. M. A. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, *29*, 242–248. https://doi.org/10.1016/j.evolhumbehav.2008.01.001

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

Primoratz, I. (1989). Punishment as language. *Philosophy, 64*, 187–205.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences*, *1*, 1–26. https://doi.org/10.1017/ehs.2019.12

Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., Benitez, J., & Ocampo, J. D. (2020). Advancing developmental science via unmoderated remote research

with children. *Journal of Cognition and Development*, *21*, 477–493. https://doi.org/10.1080/15248372.2020.1797751

Robinson, P. H., & Kurzban, R. (2006). Concordance and conflict in intuitions of justice. *Minnesota Law Review*, *91*, 1829-1907.

Sarin, A., Ho, M. K., Martin, J. W., & Cushman, F. A. (2021). Punishment is organized around principles of communicative inference. *Cognition*, *208*, 104544. https://doi.org/10.1016/j.cognition.2020.104544

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

Smetana, J. G. (2013). Young children's moral and conventional understanding. In M. Banaji & S. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 352-355). Oxford University Press.

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*, e59510. 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*, *February*, 1–17. https://doi.org/http://dx.doi.org/10.1037/a0040069

Sommerville, J. A., Schmidt, M. F. H., Yun, J., & Burns, M. (2013). The development of fairness expectations and prosocial behavior in the second year of life. *Infancy, 18*, 40–66. https://doi.org/10.1111/j.1532-7078.2012.00129.x

Thomsen, L. (2020). The developmental origins of social hierarchy: how infants and young children mentally represent and respond to power and status. *Current Opinion in Psychology*, *33*, 201–208. https://doi.org/10.1016/j.copsyc.2019.07.044

Tyler T. R. (2006). Psychological perspectives on legitimacy and legitimation. *Annual Review of Psychology, 57,* 375–400. https://doi.org/10.1146/annurev.psych.57.102904.190038

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. https://doi.org/10.1177/1745691614556679

Vaish, A., Herrmann, E., Markmann, C., & Tomasello, M. (2016). Preschoolers value those who sanction non-cooperators. *Cognition*, *153*, 43–51. https://doi.org/10.1016/j.cognition.2016.04.011

Vitiello, M. (1990). Reconsidering rehabilitation. *Tulane Law Review*, *65*, 1011-1054.

Yankah, E. N. (2004). Good guys and bad guys: Punishing character, equality and the irrelevance of moral character to criminal punishment. *Cardozo Law Review, 25*, 1019-1067.

Yucel, M., Drell, M. B., Jaswal, V. K., & Vaish, A. (2022). Young children do not perceive distributional fairness as a moral norm. *Developmental Psychology*, *58*, 1103–1113. https://doi.org/10.1037/dev0001349

Yucel, M., & Vaish, A. (2018). Young children tattle to enforce moral norms. *Social Development*, *27*, 924-936. https://doi.org/10.1111/sode.12290

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

Zimring, F. E., & Hawkins, G. (1995). *Incapacitation: Penal confinement and the restraint of crime*. Oxford University Press.