Young Children Police In-Group Members at Personal Cost

Daniel A. Yudkin*, Jay J. Van Bavel[§†], & Marjorie Rhodes[†]

*Department of Psychology, Yale University

[§]Center for Neural Science, New York University

[†]Department of Psychology, New York University

6 Washington Place, New York, NY 10003

Correspondence and requests for materials should be addressed to Daniel Yudkin (day235@nyu.edu) or Marjorie Rhodes (marjorie.rhodes@nyu.edu).

Keywords: costly third-party punishment, reputation, cooperation, fairness, development

Humans' evolutionary success has depended in part on their willingness to punish, at personal cost, bad actors who have not harmed them directly—a behavior known as *costly third party punishment*[1]. Though this behavior has been widely observed in adults[2], important questions remain as to its underlying psychology. We approached these questions from a developmental perspective, using a novel, naturalistic experiment to study costly punishment in young children (age 3-6). Results showed that even the youngest children in our sample (age 3-4) enacted costly punishment. In addition, younger (age 3-4) children policed members of their own group when placed in a position of authority. These effects of group membership and authority, along with evidence of the emergence of costly punishment at an age prior to the development of reputational concerns[3], indicate that costly punishment is promoted in part by the desire to regulate the behavior of potential cooperative partners, not just by spite[4] or reputation management[5]. Overall, the results shed new light on a behavior critical for cooperation, with implications for theories of human development, altruism, and justice.

Humans have a distinct capacity to maintain long-term cooperative arrangements among groups of genetically unrelated individuals—a feat which may have contributed to their remarkable social and technological success[6]. This ability is supported in part by their willingness to pay a cost to punish those who violate group norms[7,8]. Because such behavior provides a deterrent for would-be cheaters and free-riders, it can help sustain cooperation over time[6].

While costly punishment has been observed in a variety of human societies, the motivations underlying it remain unclear. While some evidence suggests that people only punish

when they have external incentives to do so[5,9,10,11], other work suggests that such behavior may stem in part from their intrinsic motivation to sanction bad behavior[12,13,14].

In the current research, we tested costly punishment in young children (age 3-6) in an effort to better understand the psychological processes underlying it. Past work has shown that children engage in a variety of prosocial behaviors by age 3, including reacting negatively to norm violations[15,16], restoring stolen objects to their original owners[17], and protesting violations of group norms[18]. However, no work has found direct evidence of costly punishment in children younger than 6[19,20,21]. This is notable because reputational concerns develop around age 5[3] (but do not appear to operate earlier); therefore, evidence that children before this age engage in costly punishment would be consistent with the possibility that costly punishment is motivated in part by more intrinsic desires to regulate the behavior of potential cooperative partners (a possibility that we also test with via our experimental manipulations).

In order to render our experimental method more accessible to young children, we developed a new, naturalistic method for assessing costly punishment behavior in very young children that avoided the need for quantitative comparisons, which might otherwise have proved challenging for younger samples (see Materials and Methods; Fig. 1). Our procedure was as follows. Children were brought into a room in a museum, with a large red slide. They were given the opportunity to test out the slide by going down it (all did). Then they were shown a video of another child (the "transgressor"), who had ostensibly been visiting the museum earlier that day. In the video, the transgressor was asked to hold a third child's drawing but instead crumples it up and throws it on the ground. Participants were told that the transgressor had mentioned that he or she would be returning to the classroom later in the day, and were given the chance to close the

slide to punish the transgressor—an action rendered costly insofar as it would deprive participants of the slide.

We conducted two experiments. The first experiment was intended as an initial test of whether young children would engage in costly punishment at all in our experiment. We randomly assigned participants to either a "transgression" (actor crumples drawing) or a "benign" (actor holds drawing) condition. We predicted that if children were willing to engage in costly punishment, then they would punish more in the transgression than the benign condition.

In the second experiment, we examined the motivations underlying costly punishment. Specifically, we were interested in the effect that group membership would have on children's willingness to punish transgressors. Past research presents divergent evidence as to how group membership impacts punishment decisions. Some work suggests that out-group members are subjected to higher levels of punishment than in-group[20,22,23]. On this account, punishment may be seen as a tribal means of inflicting damage on potential competitors. By contrast, other evidence suggests that in-group members are subject to harsher punishment[18,24,25], a practice called "in-group policing." This is consistent with the notion that punishment serves to promote good behavior within a cooperative community.

Our theorizing led us to suspect that differences in punishment according to group membership may be caused by a moderating factor: namely, the punisher's sense of responsibility in the group. Past research suggests that hierarchical arrangements evolved in part to solve coordinated-action dilemmas and enhance cooperation[26,27]. As a result, when individuals are placed in a position of authority, they may use punishment to promote cooperation within their group, and thus subject in-group members to higher punishment. In the absence of such

cues, they may view punishment merely as a tool to damage out-group members, and thus subject them to higher punishment.

To test this hypothesis, we manipulated both the group membership of the transgressor and the authority role of the punisher. To manipulate the group membership of the transgressor, children were told that the transgressor was a member of "this museum" ("in-group" condition) or the "Boston museum" ("out-group" condition). To manipulate authority, children were assigned to a "badge" condition or not. Those in the "badge" condition were given a "Sheriff's Badge" at the beginning of the punishment phase; those in the "no badge" condition were not given a badge.

In order to ensure the findings generalized across contents of the video, four transgression stimuli were used, including different actions (crumpling versus tearing the drawing) and actors (female and male transgressors). In both experiments, we also included a measure of non-costly punishment: children were offered a sticker, and were asked whether they also wanted the experimenter to give one to the transgressor following the primary punishment phase of the experiment.

In addition, following the punishment phase of the experiment, we asked participants a number of questions about their thoughts and feelings regarding the transgression (see Materials and Methods for more information). Specifically, we asked children the extent to which they felt in charge, as well as how much they thought it was their responsibility to make sure everyone was following the rules, expecting these to vary in accordance with the authority manipulation. We also asked a number of questions about the moral status of the transgressor, including whether how bad the action was and whether the transgressor should be reprimanded.

In the first experiment, there were no effects of video version, participant gender, transgressor gender, or interactions with these variables, all *P*s > .25, and all participants stated they enjoyed going down the slide. Confirming that children understood the nature of the punishment they were administering, a higher proportion of children who closed the slide (*n* = 13/13) compared to the proportion who left it open (*n* = 5/22) subsequently indicated that they could not use the slide again, χ^2 (1, N = 35) = 19.53, *P* < .001, *V* = .747.

Next, we tested the prediction that there would be differences in punishment according to whether the behavior children viewed was transgressive or benign. In line with predictions, more children closed the slide in the transgression condition (57%) than the benign condition (0%), χ^2 (1, N = 35) = 10.79, *P* = .001, *V* = .55. These findings were echoed in the patterns of non-costly punishment, where more children punished in the transgression condition (77%), compared to the benign condition (8%), χ2(1, N = 34) = 14.8, *p* < .001, *V* = .66.

In addition, showing that children judged the transgression as wrong, more children in the transgression stated the actor's behavior was impermissible (87%), compared to the benign condition (9%), χ2(1, N = 34) = 19.1, *p* < .001, *V* = .75, and more children stated that the transgressor should be reprimanded (77%), compared to the benign condition (0%), χ2(1, N = 34) = 18.5, *p* < .001, *V* = .74. Moreover, these judgments of permissibility and reprimand were both associated with a willingness to enact costly punishment, χ2(1, N = 34) = 13.0, *p* < .001, *V* = .61, and χ2(1, N = 34) = 12.9, *p* < .001, *V* = .61, respectively.

In the second experiment, as anticipated, there were no main effects of video version, participant gender, transgressor gender, or interactions, all *P*s > .15. Almost all (97%) of participants rated the slide as enjoyable. Children viewed the transgressor's behavior as at least somewhat wrong (95%), understood that they could not go down the slide again if it was closed

(91%), and believed it was important for the transgressor change his or her behavior (87%). Most participants (83%) responded correctly on the first try to a manipulation check of transgressor group membership; the rest were prompted until they gave the correct response.

We next tested the manipulation checks of authority and a sense of responsibility, expecting these to be higher in the badge than the no badge condition. As expected, children in the badge condition (54%) were more likely to say that they felt like the boss than those in the no badge condition (34%), $\chi 2(1, N = 175) = 6.96$, $P = .008$, $V = .2$, and feeling like the boss was subsequently associated with participants' saying they felt it was their job to make sure everyone was following the rules $\chi 2(1, N = 97) = 13.4$, $P < .001$, $V = .37$. Participants were also more likely to say that they felt "in charge" when they were in the badge condition (77%) than the no-badge condition (54%), $\chi^2 (1, N = 188) = 11.20$, $p < .001$, $V = .244$. None of these effects were moderated by age, gender, or condition, Ps > .25, except that older children were more likely to say it was important for the transgressor to change his or her behavior, $P = .005$.

Next, we examined our primary dependent variable: rates of costly punishment according to group membership and authority. In order to examine whether the effects differed by age, we included age as a factor in the analysis. Overall, 48% of children engaged in costly punishment, and, while the likelihood of punishment increased linearly with age ($B = .73$, $SE = .190$, Wald $\chi 2 = 15.05$, $P < .001$, $OR = 2.08$), each unitary age group punished at a rate significantly different from zero, including the youngest cohorts of 3- and 4-year olds (all Ps < .001; see Figure 2).

We then tested the effects of age, group manipulation, and authority manipulation on punishment, looking for evidence of in-group policing in the authority condition. The analysis revealed a significant 3-way interaction between group, role, and age ($B = .376$, $SE = .171$, Wald $\chi 2 = 4.84$, $P = .028$, $OR = 1.46$), as well as a two-way interaction between group and role among

younger children, B = -.629, SE = .275, Wald $\chi 2$ = 5.22, P = .022, OR = 1.87. Pairwise contrasts within condition among the younger cohort suggested this interaction was driven by the different treatment of in-group and out-group members according to authority role. Specifically, children assigned to the authority role punished in-group members (43%) at a significantly higher rate than out-group members, (17%), P = .038, whereas children not assigned to this role did not show an effect of group membership (indeed, the pattern of means was in the opposite direction, though not significantly so, P = .109; see Figure 3). Meanwhile, older children demonstrated no main effects, simple effects, or interactions of transgressor group membership or authority role on punishment (all Ps > .25). In sum, younger, but not older, children demonstrated different patterns of punishment according to group membership and authority role.

In order to see whether these behavioral tendencies were reflected in children's attitudes, we performed a parallel series of analyses on children's sense of responsibility for regulating others' behavior. Results showed that the extent to which children felt "in charge" echoed the pattern of effects in punishment, with a 3-way interaction between age, badge condition, and group condition, B = .38, SE = .14, Wald $\chi 2$ = 6.78, P = .009. Replicating the punishment effects, while older children's sense of responsibility did not differ as a function of their group or leadership condition, p > .25, younger children's did, B = -.55, SE = .21, Wald $\chi 2$ = 6.37, P = .012. Simple effects tests showed that, when punishing in-group members, participants in an authority role felt significantly more responsible than those in the control, B = .75, SE = .30, Wald $\chi 2$ = 6.39, P = .011; no significant effects of authority emerged in the out-group transgressor condition, P > .25. Finally, there was a positive correlation between costly punishment and participants' believing it was important for the transgressor to change his or her

behavior, $r = .32$, $P < .001$, and believing it was important for them to teach the transgressor a lesson, $r = .23$, $P = .001$.

Examining the rates of non-costly punishment, 63.6% of participants enacted non-costly punishment, a significantly higher proportion than those who delivered costly punishment (47.6%), $z = 3.12$, $P = .002$, providing further evidence that the punishment was indeed seen as costly. Rates of non-costly punishment significantly increased with age, $B = .49$, $SE = .19$, Wald $\chi 2 = 2.82$, $P = .010$, $OR = 1.63$. There was no age-by-punishment-type interaction, $P = .37$, and no significant 3-way interaction of age, role, and group, $P = .61$. Non-costly punishment was predicted by participants' sense of authority (feeling like "the boss"), $\chi 2(1, N = 183) = 7.64$, $P = .006$, $V = .21$, and by their belief that the transgressor should be reprimanded for his or her behavior, $\chi 2(1, N = 178) = 15.03$, $P < .001$, $V = .29$.

These findings speak to ongoing debates regarding the motivations underlying costly punishment. Some scholars have argued that costly punishment is born of people's desire to enhance their own reputations[5,9]. Consistent with this view, people are more likely to punish when observed[5], and those who do punish are subsequently considered more trustworthy[11]. By contrast, other research suggests costly punishment derives from inherent motivations to uphold social norms[12,13,14]. Supporting this view, research shows people find enacting punishment intrinsically rewarding[14,28], and will punish transgressors even when their decisions are anonymous[12,12]. In addition, groups in which costly punishment is prevalent have been shown to perform better in intergroup conflict[29], demonstrating how this tendency may prove adaptive in group-based contexts.

Our findings inform this debate when considered light of research on the emergence of reputational concerns in children. A growing body of evidence suggests that children begin

caring about their reputations at age 5, but not earlier[3]. For instance, the presence of observers increases helping behavior in children age 5 and older[30,31,32], but has no effect on younger children[33]. And being reminded of reputational concerns diminished the rate of cheating among 5-year-olds, but had no effect on children aged 3 and 4[34,35]. Thus observations of costly punishment before age 5 (at an earlier age than reputational concerns have ever been found) is consistent with the possibility that reputation-based processes are not the sole motivator of costly punishment. This is precisely what we observed; whereas past work has found evidence of costly punishment at age 6 (but not younger[19,19,20]), we observe this behavior in children as young as 3-4 years of age. This constitutes the youngest instance of costly punishment ever observed, and, together with our experimental effects of group membership and authority, is consistent with the possibility costly punishment behavior is motivated at least in part by the desire to regulate the behavior of potential cooperative behaviors, and does not always depend on reputational concerns. Of course, we cannot rule out the possibility that reputation-based processes promote costly punishment once they become active; nor can we specify the motivational factors that drive punishment later in life.

The finding that a sense of authority affects young children's punishment of group members corroborates other work regarding the effects of group membership and leadership on social norm enforcement. Existing theories suggest that hierarchical social structures evolved to help coordinate collective action within groups[26,26]. Given this conceptualization, a sense of authority might be expected to engender greater punishment, particularly of in-group members, since punishment can help encourage positive behavior among future cooperation partners. The patterns of punishment among the 3-4 year olds fit precisely this prediction. By contrast, when young children were not given an authority role, such effects were erased or even reversed (as

indicated by the non-significant out-group derogation effect in the "no badge" condition). In this way, these findings show how in-group policing may emerge when people have a sense of responsibility to regulate group behavior.

While some work has examined the adaptive role that leadership can play in promoting cooperation in group contexts, the current research presents a number of advances in this regard. First, to our knowledge this is the first study to directly demonstrate how a greater sense of authority can increase people's willingness to enact costly punishment of in-group members. In addition, our work shows how young children may be sensitive to cues indicating their authority position. The fact that such authority cues impacted the behavior of children age 3-4 suggests that the human mind may be primed to assume the punitive responsibilities that leadership confers.

Whereas younger children punished differently according to group membership and authority, older children punished in-group and out-group members almost equally. One possible explanation for this finding has to do with the reputational concerns alluded to above. Specifically, is possible that the sorts of reputational concerns that emerge around age 5 promoted greater egalitarianism. Consistent with this possibility, children aged 5 and older act more fairly when they are observed[3]. Another possible explanation is the development of higher-order reasoning capacities[36]. Research with adults has found that deliberation mitigates intergroup bias in punishment[22]; thus, children age 5 and older may have engaged in more egalitarian deliberation than younger children in this context.

There are several important limitations to note in this set of experiments. First, the sample of 3-year-olds in Experiment 2 was small and no children age 3 were tested in Experiment 1. For this reason we grouped the 3- and 4-year-olds together (note that age 4 is still prior to the

emergence of either costly punishment or reputation management found in previous work), but future work should examine the development of costly punishment prior to age 4 in more detail. Second, it is difficult to tell from the present data precisely why children showed different patterns for costly and non-costly punishment. Costly punishment sends a particularly strong signal that the punisher disapproved of the transgression, since the punisher was willing to incur a personal cost to sanction it. As people may be sensitive to punishment's communicative intent[13,28] , costly punishment may serve as a more effective means of group regulation (because it sends such a strong signal of disapproval) —thus making it more sensitive to the manipulations of group membership and authority. This may also explain why costly punishment is more effective than other forms of punishment for promoting cooperation[37]. The possibility that different psychological processes underlie decisions to engage in costly vs. non-costly processes is an important area for future work.

Because costly punishment serves to regulate the behavior of individuals not directly related to the self, it is believed to underlie more abstract notions of justice and fairness. For instance, the fact that members of a society will endure considerable inconvenience and effort to ensure just desserts for transgressors with whom they have had no direct involvement is the basis of the modern legal system. And a willingness to intervene against injustice, even at considerable personal sacrifice, animates many of the organizations fighting for a more equitable society around the world. Here we find that a variety of factors influence punishment decisions, including group membership, authority, and non-reputational concerns. The fact that young children police group members at personal cost suggests that the underpinnings of such vital human institutions emerges early in life.

Materials and Methods

Participants: Experiment 1. We anticipated a large effect size (V = .5, power = 80%) and so sought a sample of approximately 35 participants. The sample was 9 white, 9 black, 2 Asian, 2 Hispanic, 9 mixed race/other, 4 unreported, *m* age = 5.4, range 4-6 years.

Participants: Experiment 2. A power analysis performed in G*Power 3 suggested a sample of approximately 200 would be capable of detecting the small-to-medium main effects and interactions that we anticipated (Cohen's f^2 = .06, number of predictors = 4, power = 80%). Our initial sample consisted of 204 participants, of which thirteen were excluded from all analysis on the basis of *a priori* exclusion criteria: 5 for inattention or a learning disability, 3 for experimenter error, and 5 for parental interference (further information on excluded participants can be found at osf.io/c62t5). The core sample was 53 white, 15 black, 25 Asian, 17 Hispanic, 60 mixed race/other, 21 unreported, *m* age 5.2, range 3-6 years.

Procedure. Children recruited from the Children's Museum of Manhattan were led into a classroom in the corner of which was a slide. Two experimenters were used in order to limit demand effects. Experimenter 1 showed participants the "Open" sign that had been affixed to the slide, and allowed them familiarize themselves with the slide by going down it. (Experimenter 2 waited outside the classroom.) Next, participants were brought to a table to perform a drawing activity. They learned the museum was asking attendees to draw pictures for a book. After drawing a picture, they learned of two other children who had also drawn pictures earlier that day. In the In-group condition, the two other children were members of the Children's Museum and their drawings would be placed in the same book as the participants'. In the Out-group condition, the children belonged to the Boston Museum and their drawings were to be placed in the other book. Next, participants were told that one of the children had asked the other to hold her drawing while she went to the bathroom. They were told that what happened next had been

recorded and would be played on the computer monitor, and were asked to watch while the experimenter moved away to arrange some papers. In the Authority condition, they were also given a sheriff's badge and told they were in charge; in the Control condition there was no such badge. The monitor displayed a video ostensibly recorded by the camera depicting a child (the transgressor) sitting on the couch and crumpling the other child's drawing. The experimenter returned with the sign and said that the transgressor had mentioned she would be returning later in the day to play on the slide. She then wondered aloud which sign to put up, and left the room.

The other experimenter, who had been outside the classroom during the transgression entered and asked children which sign they wanted to affix to the slide, the green "Open" sign or the red "Closed." After participants made a decision and placed the sign, they were asked a series of follow-up questions, including whether they felt in charge (No—Very), like the "boss," (No—Yes), responsible for making sure other people were following the rules (No—Yes); and whether it was important to teach the transgressor a lesson (No—Very), how bad the behavior was (Not at all—Very), and whether the transgressor should get in trouble (No—A lot). They were then thanked for their participation and parents were fully debriefed. All data, experimental scripts, and protocols are available at the Open Science Framework, osf.io/c62t5.

Acknowledgements. We thank Elyana Feldman, Lisa Kaggen, Christine Tai, Rachel Vitale, and Kat Yee for help with data collection and the NYU Social Perception and Evaluation Lab for feedback on this research. This work was supported by a Research Challenge Grant for Women in Science and by a James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition – Scholar Award to Rhodes and National Science Foundation Grant 1555131 (to J.J.V.B.). The authors declare no competing financial interests.

Author Contributions DAY, JVB and MR designed the research, DAY collected the data, DAY analyzed the data with input by JVB and MR, and DAY wrote the paper, with critical edits by JVB and MR.

1. Fehr E, Fischbacher U (2004) Third-party punishment and social norms. *Evol Hum Behav* **25**(2):63–87.
2. Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J.C., Gurven, M., Gwako, E., Henrich, N. and Lesorogol, C. Costly punishment across human societies. *Science*, **312**, 1767-1770 (2006).
3. Engelmann JM, Rapp DJ (2018) The influence of reputational concerns on children's prosociality *Cur Opin Psych* **49**:1–19.
4. Jensen, K. (2010). Punishment and spite, the dark side of cooperation. *Phil Tran Royal Soc B: Biol Sci*, **365**(1553), 2635-2650.
5. Kurzban R, DeScioli P, O'Brien E (2007) Audience effects on moralistic punishment. *Ev Hum Behav*, **28**(2), 75-84.
6. Boyd, R., Gintis, H., Bowles, S., & Richerson, P. J. The evolution of altruistic punishment. *Proc Natl Acad Sci* **100**, 3531-3535 (2003).
7. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K (2007) Via freedom to coercion: the emergence of costly punishment. *Science* **316**(5833): 1905-1907.
8. Fehr E, Gächter S (2002) Altruistic punishment in humans. *Nature* **415**(6868):137–140.
9. Pedersen EJ, Kurzban R, McCullough ME (2013) Do humans really punish altruistically? A closer look. *Proc Biol Sci* **280**(1758), 20122723.
10. Krasnow MM, Cosmides L, Pedersen EJ, Tooby J (2012) What are punishment and reputation for? *PLOS One*, **7**(9), e45662.
11. Jordan JJ, Hoffman M, Bloom P, Rand DG (2016) Third-party punishment as a costly signal of trustworthiness. *Nature* **53**(7591), 473-476.
12. Piazza J, Bering JM (2008) The effects of perceived anonymity on altruistic punishment. *Evol Psych*, **6**(3).
13. Crockett MJ, Özmedir Y, Fehr E (2014) The of vengeance and the demand for deterrence. *J Exp Psych: Gen*, **143**(6):2279.
14 De Quervain DJ, Fischbacher U, Treyer V, Schellhammer M (2004) The neural basis of altruistic punishment. *Science* **305**(5688):1254-1258.
15. Vaish A, Missana M, Tomasello M (2011) Three-year-old children intervene in third-party moral transgressions *Brit Jo Dev Psych* **29**:124-130.
16. Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proc Nat Ac Sci*, **108**(50): 19931-19936.
17. Riedl K, Jensen K, Call J, Tomasello M (2015) Restorative justice in children *Cur Biol* **25**:1731-1735.
18. Schmidt MFH, Rakoczy H, Tomasello M (2012) Young children enforce social norms selectively depending on the violator's group affiliation *Cognition* **124**(3):325–33.
19. McAuliffe K, Jordan JJ, Warneken F (2015) Costly third-party punishment in young children *Cognition* **134**:1-10.
20. Jordan JJ, McAuliffe K, Warneken F (2014) Development of in-group favoritism in children's third-party punishment of selfishness *Proc Natl Acad Sci USA* **111**:12710-12715.
21. Salali GD, Juda M, Henrich J (2014) Transmission and development of costly punishment in children *Evol Hum Behav* **36**:86-94.
22. Yudkin DA, Rothmund T, Twardawski M, Thalla N, Van Bavel JJ (2016) Reflexive intergroup bias in third-party punishment *J Exp Psych: Gen* **145**:1448-1459
23. Bernhard H, Fehr E, Fischbacher U (2006). Group affiliation and altruistic norm enforcement. *Am Econ Rev* **96**(2):217-221.
24. Mendoza, SA, Lane SP, Amodio DM (2014). For members only: Ingroup punishment of fairness norm violations in the ultimatum game. *Soc Psych Pers Sci* **5**(6):1–9.
25. Shinada M, Yamagishi T, Ohmura Y (2004) False friends are worse than bitter enemies: "Altruistic" punishment of in-group members *Evol Hum Behav* **25**(6):379–393.
26. Van Vugt, M (2006) Evolutionary origins of leadership and followership. *Per Soc Psych Rev*,

10(4):354-371.
27. Glowacki L, von Rueden C (2015). Leadership solves collective action problems in small-scale societies. *Phil. Trans. R. Soc. B*, **370**(1683):20150010.
28. Funk, F, McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Per Soc Psych Bull*, **40**(8), 986-997.
29. Sääksvuori L, Mappes T, Puurtinen M (2011) Costly punishment prevails in intergroup conflict *Proc Ro Soc of London B: Bio Sci*, rspb20110252.
30. Piazza J, Bering JM, Ingram G (2011) "Princess Alice is watching you": children's belief in an invisible person inhibits cheating *J Exp Child Psych* **109**:311-320.
31. Buhrmester D, Goldfarb J, Cantrell D (1992) Self-Presentation when sharing with friends and nonfriends *J Early Adol* **12**:61-79.
32. Engelmann JM, Herrmann E, Tomasello M (2012) Five-year olds, but not chimpanzees, attempt to manage their reputations *PLoS ONE* 7:e48433.
33. Warneken F, Tomasello M (2013). Parental presence and encouragement do not influence helping in young children. *Infancy*, **18**(3):345-368.
34. Leimgruber KL, Shaw A, Santos, LR, Olson KR (2012) Young children are more generous when others are aware of their actions *PLoS ONE 7*:e48292.
35. Engelmann JM, Herrmann E, Tomasello M (2017). Concern for group reputation increases prosociality in young children. *Psyc Sci*, 0956797617733830.
36. Bunge SA, Zelazo PD (2006) A brain-based account of the development of rule use in childhood *Cur Dir Psych Sci* **15**:118-121.
37. Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Hum Nat*, **13**(1), 1-25

Figures

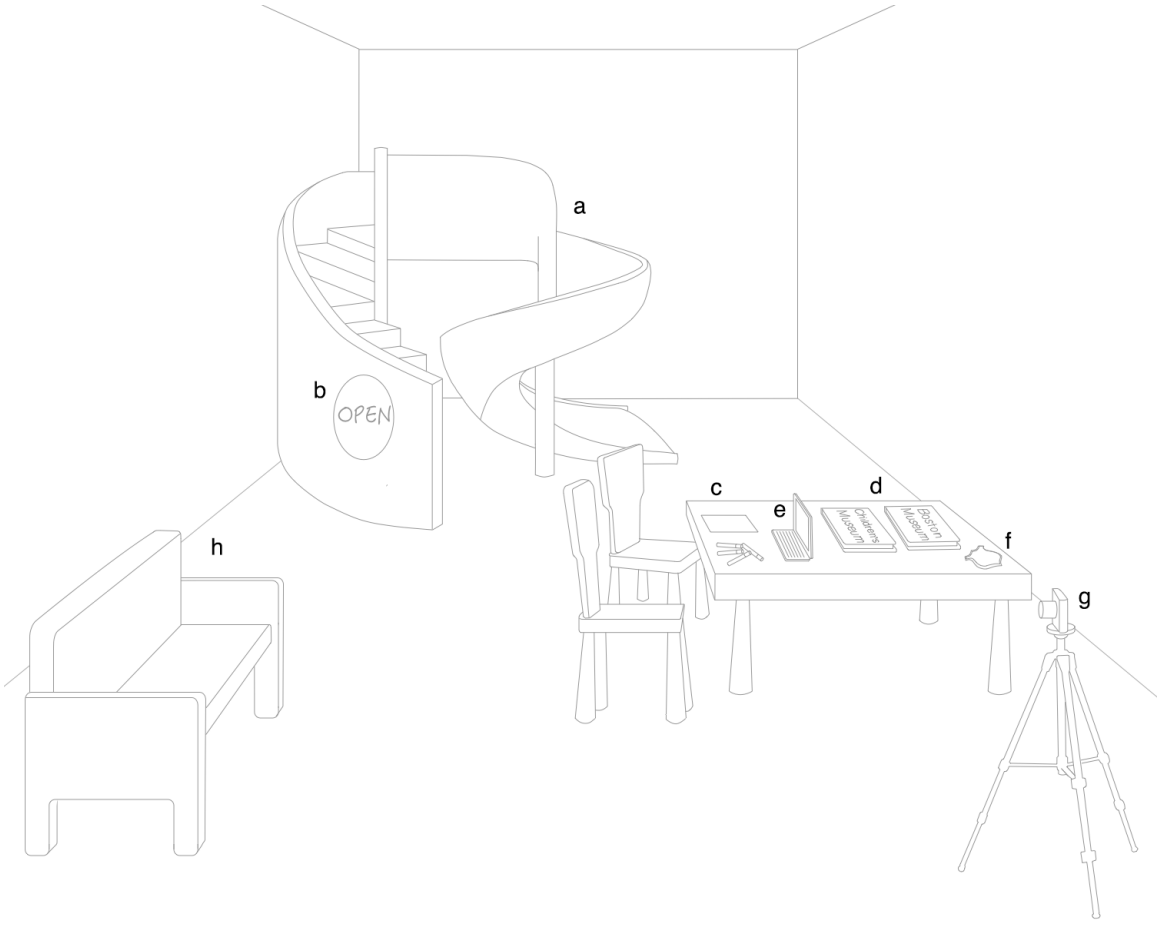

Figure 1. Schematic of experiment space. (A) Slide to be revoked as punishment. (B) Open/closed sign. (C) Paper for drawing project. (D) Books in which the art project would be stored. (E) Computer on which transgressor video was played. (F) Sheriff's badge randomly assigned to participants. (G) Video camera on which interaction was recorded. (H) Couch where transgression occurred.

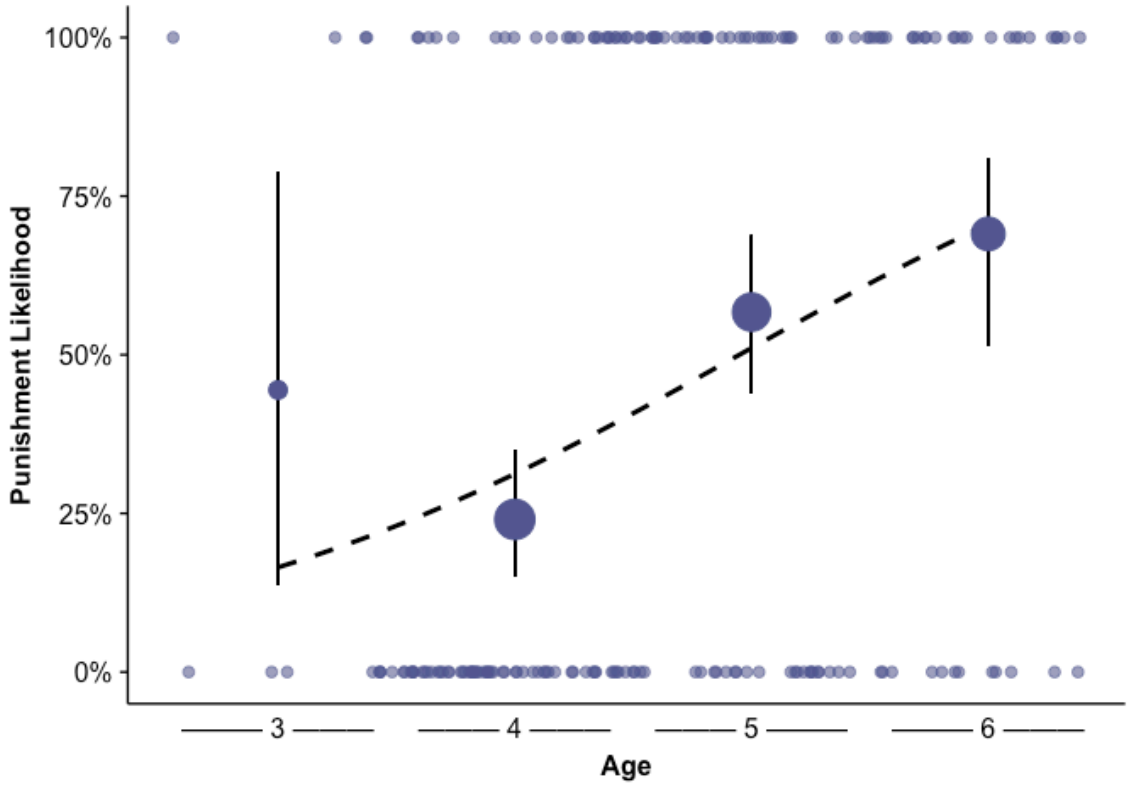

Figure 2. Likelihood of costly punishment (y-axis) predicted by age (x-axis). Large circles represent unitary age means (size reflects sample size); small dots represent individual participants by age; dashed line depicts logistic regression line; error bars show Clopper-Pearson 95% confidence intervals. X-axis labels are centered on each unitary age. Punishing increases significantly with development, and each unitary age group differs significantly different from 0, including children 3 years of age (all *P*s < .001; *n* = 191).

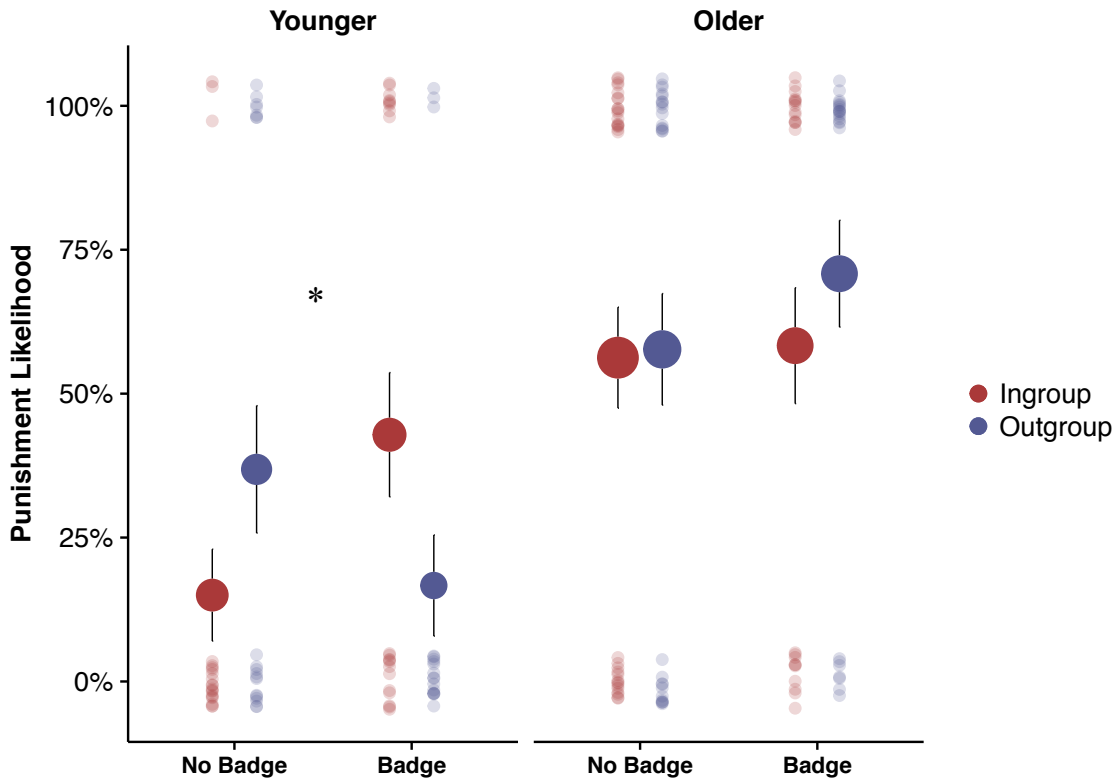

Figure 3. Costly punishment (y-axis) as a function of age, group membership and role (n = 191. Large circles represent means within the badge and group conditions (size reflects sample size); small dots reflect individual participants (jittered). Error bars = SEM. Younger (aged 3-4), but not older (aged 5-6) children punish in-group members more harshly than out-group when placed in a position of authority; older children show no such effect. *$P < .05$