UNIVERSITY of York

This is a repository copy of *Social norms and cultural diversity in the development of third-party punishment*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/165620/

Version: Accepted Version

Article:
House, Bailey R orcid.org/0000-0002-4023-9724, Kanngiesser, Patricia, Barrett, H Clark et al. (5 more authors) (2020) Social norms and cultural diversity in the development of third-party punishment. Proceedings of the Royal Society B: Biological Sciences. p. 20192794. ISSN: 1471-2954

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

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

FINAL ACCEPTED MANUSCRIPT

Social norms and cultural diversity in the development of third-party punishment

Bailey R. House[1,2]*, Patricia Kanngiesser[4], H. Clark Barrett[5], Süheyla Yilmaz[6], Andrew Marcus Smith[5], Carla Sebastian-Enesco[7], Alejandro Erut[5], Joan B. Silk[2,3]

[1] University of York, Department of Psychology.

[2] Arizona State University, Institute of Human Origins.

[3] Arizona State University, School of Human Evolution and Social Change.

[4] Freie Universität Berlin, Faculty of Education and Psychology.

[5] University of California, Los Angeles, Department of Anthropology.

[6]Leipzig University, Faculty of Education.

[7] Universidad Complutense de Madrid, Faculty of Psychology.

* Correspondence to: bailey.house@york.ac.uk

Abstract

Human cooperation is likely supported by our tendency to punish selfishness in others. Social norms play an important role in motivating third-party punishment, and also in explaining societal differences in prosocial behavior. However, there has been little work directly linking social norms to the development of third-party punishment across societies. In this study, we explored the impact of normative information on the development of third-party punishment in 603 children aged 4-14 years, across six diverse societies. Children began to perform third-party punishment during middle childhood, and the developmental trajectories of this behavior were similar across societies. We also found that social norms began to influence the likelihood of performing third-party punishment during middle childhood in some of these societies. Norms specifying the punishment of selfishness were generally more influential than norms specifying the punishment of prosocial behavior. These findings support the view that third-party punishment of selfishness is important in all societies, and its development is shaped by a shared psychology for responding to normative information. Yet, the results also highlight the important role that children's prior knowledge of local norms may play in explaining societal variation in the development of both third-party punishment and prosociality.

Keywords

Third-Party Punishment; Prosocial Behavior; Social Norm; Cross-Cultural; Altruistic Punishment; Antisocial Punishment

Social norms and cultural diversity in the development of third-party punishment

Background

As a species, humans are unusual because we regularly cooperate in large groups of unrelated individuals and incur costs to punish violators of social norms. Both theory and data suggest that cooperation and punishment are tightly linked [1]. In large groups, the temptation to gain the benefits of cooperation without contributing (sometimes called "free-riding") can destabilize cooperation. Selective punishment of third-party individuals (Third-Party Punishment) who free-ride can promote cooperation by reducing the benefits of free-riding without harming cooperators, making cooperation a relatively more attractive choice to individuals [2–4]. The prospect of punishment increases rates of cooperation in experimental games [5–7] and plays an important role in sustaining costly forms of altruism in societies without coercive institutions [8,9]. Third-Party Punishment (TPP) of non-cooperators likely plays a stronger role in maintaining cooperation in large groups, relative to other mechanisms such as dyadic or indirect reciprocity [10]. Yet, TPP is itself costly to individuals and there is a temptation to free-ride on it as well, which is why costly TPP of non-cooperators/free-riders is often referred to as 'altruistic punishment'. This means that to understand human cooperation it is important to understand the factors that influence individuals' tendency to engage in TPP [6].

Some researchers argue that TPP is at least partly motivated by culturally-specific norms and institutions (organized systems of norms) [1,8,9,11], which are acquired through a universal human psychology for learning and conforming to norms [12]. Norms are behavioral standards shared and enforced by a community [13], and they are based on expectations about what others in your community *do* (or think *you* should do) which

motivate your own behavior [14]. There is evidence that culturally-specific norms at least partly motivate punishment of non-cooperative or selfish third parties (i.e. altruistic punishment) [8,9], but also punishment of prosocial third parties (i.e. antisocial punishment) [7,15]. This is important because the tendency to punish selfish third parties increases group-level cooperation, but the tendency to punish prosocial third parties decreases group-level cooperation [7]. The influence of culturally-varying norms for punishment could explain the considerable variation across societies in rates of cooperation, likelihood of imposing sanctions, and the kinds of behaviors that are sanctioned [7,16–19]. It could also account for why levels of prosociality and TPP are positively associated across societies [17].

Developmental data provide a means to examine the influence of norms on social behavior. Experimental evidence from a diverse range of societies suggests that children become sensitive to normative information about prosocial behavior during middle childhood (about 6-12 years of age), and this precedes or co-occurs with the development of culturally-variable adult behavior [12,20,21]. Evidence that children's sensitivity to normative information about punishment is linked to the development of TPP in children would support the hypothesis that TPP is at least partly motivated by learned cultural norms.

This evidence would provide important insights into the factors that underlie the development of TPP. There is evidence that from a young age children in urban North America and Europe protest rule violations and punish certain forms of deviant third-party behavior [22–27], including self-maximizing resource allocations by third parties in cooperative experimental tasks [28–32]. However, it is not clear whether children's punishment is motivated by awareness of relevant social norms. Moreover, even if children are aware of norms at a young age, they may not always be motivated to *conform* to norms.

In particular, it is likely that in situations in which following norms incurs a personal cost, children's *conformity* to norms may lag behind their knowledge of norms [12]. A good example of this comes from a study of children's behavior in the Dictator Game (DG) conducted in Boston, USA [33]. In the DG, a subject is given an endowment (e.g. cash, stickers, candy), and is able to allocate a fraction of the endowment to another anonymous participant. On average, 3-4 year-old children thought they should allocate about 43% of their original endowment of stickers, but on average they actually gave up only 13% of the endowment. By 7-8 years of age, children's allocations matched what they reported they were 'supposed' to do.

This suggests that to investigate the link between social norms and TPP, we need to investigate how children respond to information about social norms that directs them to punish third parties. The age at which children modify their TPP in response to this information would mark a developmental increase in children's tendency to strongly *conform* to social norms for TPP. Given the theoretical connections between TPP and prosociality [6], we would expect that norms will begin to motivate TPP at the same age that norms begin to influence prosocial behavior: middle childhood [12,21].

To explore whether the link between social norms and TPP is a product of a universal psychology for conforming to social norms, we should study how children in different societies respond to social norms for punishing both selfish and prosocial third parties. If children's responsiveness to information about social norms develops similarly across societies, this would support the hypothesis that there is a universal human psychology for conforming to norms. A universal psychology for conforming to norms is also expected to lead to the emergence of societal *variation* in TPP because the tendencies to perform altruistic punishment (i.e. punishing selfish third parties) and antisocial punishment (i.e.

punishing prosocial third parties) vary greatly across societies. Evidence suggests that this variation (largely documented through studies of adults) is likely due to the fact that the content of local punishment norms differs across societies [7].

We conducted a cross-cultural experimental study of the development of personally-costly TPP in children. Specifically, we addressed four Research Questions: **RQ1**: Does TPP develop in a uniform way among children across societies? **RQ2**: At what age during development do children become responsive to experimentally-manipulated normative information about punishment? **RQ3**: Does the content of norms influence children's responsiveness to them (i.e. are children equally responsive to information about norms that directs the punishment of selfishness or prosociality)? **RQ4**: Does the development of children's responsiveness to normative information about punishment resemble the development of their responsiveness to normative information about prosocial behavior?

Methods

Participants: 603 children aged 4-14 years from six societies (Table 1, Supplementary Table 1). 87 children were excluded for failing at least one comprehension question about the procedure (out of seven sets of questions; see ESM for details).

Overview: The Third-Party Punishment Game (TPPG) is often used to approximate social interactions in which punishment could be applied [28,34]. We used a version of the TPPG which was designed to be both child-appropriate and comparable across field sites. Subjects were told about the choices of third parties in a discrete form of the Dictator Game (DG), in which the third party could make either (i) a prosocial choice or a (ii) self-maximizing choice. Subjects were then given an opportunity to allocate some of their own endowment

(i.e. their payoff) to punish the third party. Additional Methodological Details are available in the ESM.

Procedure: Subjects were introduced to uniquely colored people-shaped figurines and told that each represented an unknown person from their community. Third parties and Recipients were represented by these figurines. Subjects were told that the third parties "play the game at a different time", and this was true: the third party was a participant in a parallel study in which children made DG choices knowing that they could be punished by someone else. Subjects' punishment decisions in the current study were used to modify the third party's payoffs in this parallel study.

Subjects were told that third parties could select either one of two laminated paper trays, examples of which were placed in front of the Subject (Figure 1). Each tray was marked with a red and a blue circle, and tokens were placed in the circles: any tokens placed in the red circle would go to the third party, and any tokens placed in the blue circle would go to the Recipient. One tray corresponded to the prosocial outcome (1 reward for the third party and 1 reward for a Recipient; 1/1), and the second corresponded to the self-maximizing ("selfish") outcome (2 rewards for the third party and 0 rewards for a Recipient; 2/0). By selecting a tray, third parties selected either the 1/1 or 2/0 outcome in the DG.

Subjects were next told that, if they wanted to, they could give the experimenter one of their own tokens (out of a supply of three provided at the start of each trial) and in exchange the experimenter would take one token from the amount that would be obtained by the third party (based on the third party's choice in the DG). Both the given and taken tokens would be removed from the game (meaning that punishment was costly, as it reduced the payoffs of both the subject and the third party). If children chose to keep all

their tokens, no tokens would be taken from the third party (meaning that the third party would not be punished).

Subjects then viewed a short video in which an adult model verbalized normative information about the different possible choices subjects could make in the TPPG (we refer to this normative information as the 'norm prime'). Videos used a standardized script, but were recorded separately at each field site using local translations of the script and local adults as models (models were selected to be unfamiliar to the subjects, but were not matched by age or sex across sites). Subjects saw one of three norm prime videos:

Punish-Selfish: a directed norm prime that indicated that 2/0 was 'wrong' and 'it is good to take a token from someone who chooses this', and also that 1/1 was 'right' and 'it is bad to take a token from someone who chooses this'. Here, the norm prime specifies punishing selfish third parties.

Punish-Prosocial: a directed norm prime that 2/0 was 'right' and 'it is bad to take a token from someone who chooses this', and also that 1/1 was 'wrong' and 'it is good to take a token from someone who chooses this'. Here, the norm prime specifies punishing prosocial third parties.

Punish-Either: an undirected norm prime that indicated that 2/0 was 'ok' and 'it is ok to take a token from someone who chooses this', and that 1/1 was 'also ok' and 'it is also ok to take a token from someone who chooses this'. This provides an ideal reference condition because the model's actions in the video (with respect to 2/0 and 1/1) are closely matched to both the 'Punish-Selfish' and 'Punish-Prosocial' videos, but the content of the norm prime does not specify punishing any behavior by third parties as it does in the two 'directed' conditions.

After viewing the video, the two primary test trials were conducted. At the start of each trial, subjects were allocated three tokens and asked whether they wanted to (i) give one token to the experimenter (i.e. punish the third party), or (ii) keep all of their tokens (i.e. not punish). In the 'Selfish Third-Party Trial', subjects chose whether to punish a third party who had previously chosen 2/0 in the DG (the self-maximizing outcome) . In the 'Prosocial Third-Party Trial', subjects chose whether to punish a third party who had previously chosen 1/1 in the DG (the prosocial outcome). The order of these trials was randomized across subjects by a tablet computer used to record the data.

Addressing Research Questions: To answer RQ1, a sample of children in each society was presented with the undirected norm prime: Punish-Either. We used results from this experimental condition to compare the probability of punishing selfish and prosocial third parties, and to assess ontogenetic changes in the likelihood of punishing selfish and prosocial third parties within and across societies.

To answer RQ2, in each society we presented two additional samples of children with one of the two directed norm primes: Punish-Selfish or Punish-Prosocial. We compared the impact of each of the directed primes on the likelihood of punishing selfish and prosocial third parties. We also explored how the effects of these norm primes changed with age, whether the effects of the primes varied across societies, and whether the primes influenced punishment of selfish and prosocial third parties to the same extent.

To answer RQ3, we determined whether the Punish-Selfish prime increased children's tendency to punish selfish third parties more than prosocial third parties (relative to children given the Punish-Either prime), and also whether the Punish-Prosocial prime decreased children's tendency to punish selfish third parties more than prosocial third parties (relative to children given the Punish-Either prime). We then compared the

magnitude of these effects, to determine whether either the Punish-Selfish or Punish-Prosocial primes influenced children to a greater degree.

To answer RQ4, we compared the effects of the norm primes on children's punishment behavior in the present study (i.e. RQ2) to the effects of similar norm primes on the same children's prosocial behavior in a prior published study [12].

Task comprehension: Subjects understood third parties' decisions in the DG, because they had previously made their own decisions in a DG at the start of the testing session, before being presented with the two TPPG trials. This DG also involved viewing a norm prime video. Norm primes in the DG and TPPG were always congruent in content (i.e. they always stated that the same choices were 'good' and 'bad'). Analysis of children's choices in this DG can be found in [12] (we do not analyze the relationship between children's choices in the DG and in the TPPG in this paper). Children also passed seven sets of Comprehension Questions (CQs) which asked children to explain the task procedures (see ESM for details on children who were excluded for answering CQs incorrectly).

Rewards: Subjects were told that "the more tokens they received, the more rewards they would receive", but the precise nature of the rewards or the exchange rate was not communicated to subjects so as to reduce differences in the perceived value of the payoffs across sites. After the study, tokens were exchanged for rewards (locally sourced small food items, or prizes such as stickers).

Statistical modeling approach: We used statistical analyses which are well-suited to deriving inferences from the complex hierarchical data produced by this study. All data were binary choices taking the form of "1" (subject chose to punish) or "0" (subject chose NOT to punish). We modeled subjects' choices using regression with a binomial link function. The posterior distribution of the model was estimated using Markov Chain Monte Carlo, where

we generated model predictions by processing many samples from the posterior distribution of the model. Data were analyzed in the R Environment for Statistical Computing version 3.5.1 [35], with most models specified using the function 'map2stan' (R package 'rethinking' version 1.59), a convenience tool for fitting different regression models [36]. Multilevel models were run using a variant of Hamiltonian Monte Carlo (an algorithm particularly good with high dimension models) implemented in RStan version 2.18.2 [37]. Models were specified using weakly informative priors, which reduce overfitting and also help the Markov chain to converge to the posterior distribution more effectively than flat priors. The posterior distribution we present here is based on 15000 samples from three chains (after 3000 adaptation steps), for a total of 36000 samples. These samples were sufficient to establish convergence to the target posterior distribution. We assessed convergence through (a) visual inspection of the chains, (b) the R-hat Gelman and Rubin statistic (~1.00 for all parameters, R-hat values greater than 1.01 can indicate that the chain did not converge), and (c) the effective number of samples for all parameters were reasonable (though the chains were inefficient, requiring large numbers of overall samples).

Results

We evaluated five models to predict the likelihood that children will give up a token to punish a third party, each capturing a different hypothesis about the factors that influence TPP (Table 2). The model which receives the most WAIC weight, Model 5, indicates that the third party's behavior (selfish, prosocial), norm content (Punish-Prosocial, Punish-Selfish, or Punish-Either), subject age, and site/society all influence punishment outcomes. We answer all four research questions using the results of Model 5, which we present

graphically because the complexity of the model makes numerical results difficult to interpret (see the ESM for some numerical results).

RQ1: Does TPP develop in a uniform way among children across societies?

The Punish-Either norm prime condition captures children's tendency to punish selfish and prosocial third parties in the absence of normative information specifying the punishment of either. For this reason, the model's estimates for the Punish-Either condition provide the clearest view on the development of TPP (additional analyses in the ESM show similar developmental patterns for the Punish-Selfish and Punish-Prosocial conditions).

Across the six societies, children were more likely to punish selfish third parties than prosocial third parties, with the 95% confidence intervals indicating that this pattern increases with age and becomes reliable by at least age 9 (Figure 2A). Qualitatively similar developmental patterns characterized all six societies (Figure 2B-2G), although the estimates for punishment of selfish third parties and prosocial third parties were only reliably different for Berlin, La Plata, Phoenix, and Pune (additional analyses in the ESM show that punishment develops similarly across societies for both selfish and prosocial third parties). For the Shuar and the Wichí the developmental pattern for the estimates also looks similar to the other societies, but the wide confidence intervals mean we cannot be confident that this pattern is reliable. We note that the Shuar and Wichí samples were small; we include them for completeness, but interpret them with caution.

RQ2: At what age do children become responsive to normative information about punishment?

To assess this we examined the impact of the directed norm primes (Punish-Selfish, Punish-Prosocial) on children of different ages. In La Plata, Phoenix, and Pune children who viewed the Punish-Selfish prime are more likely to punish selfish third parties than are children who viewed the Punish-Prosocial prime (Figure 3C,E,G). This effect emerges by about 8 years of age (Figure 3C,E,G). However, the norm primes did not affect the likelihood of punishing prosocial third parties in these societies (Figure 3D,F,H).

In Berlin, Shuar, and Wichí there is no evidence that the directed norm primes influenced children's punishment at any age. In these societies, there was no reliable difference in children's likelihood of punishing selfish third parties or prosocial third parties based on whether they were given the Punish-Selfish (Figure 3A,I,K) or Punish-Prosocial prime (Figure 3B,J,L).

RQ3: Does the content of norms influence children's responsiveness to those norms?

The analyses of RQ1 and RQ2 suggest that, in at least some societies, the Punish-Selfish prime increased children's tendency to punish selfish third parties more than prosocial third parties (hereafter "children's bias towards punishing selfish third parties"), while the Punish-Prosocial prime decreased children's bias towards punishing selfish third parties. But, are the magnitudes of these effects the same?

Two interaction terms in Model 5 capture these effects. These interactions are between the two dummy parameters for (a) the Punish-Selfish prime or (b) the Punish-Prosocial prime, and the fixed effect for Third Party Behavior (selfish third party or prosocial

third party). These interactions capture the degree to which each of the directed norm primes (Punish-Selfish and Punish-Prosocial) increase or decrease children's bias towards punishing selfish third parties, relative to the magnitude of that bias in the Punish-Either prime condition (the regression model's reference level). If the coefficient for the interaction "Punish-Selfish prime X Third Party Behavior" is greater than zero, then the Punish-Selfish prime increased children's bias towards punishing selfish third parties (relative to the Punish-Either prime condition). Similarly, if the coefficient for the interaction "Punish-Prosocial prime X Third Party Behavior" is less than zero, then the Punish-Prosocial prime decreased children's bias towards punishing selfish third parties (relative to the Punish-Either prime condition).

Figure 4 plots the magnitudes of the estimated coefficients for these interactions for children of different ages. In La Plata, Phoenix, and Pune, the "Punish-Selfish prime X Third Party Behavior" interaction increases with age, and is reliably greater than zero by about age 9-10 (Figure 4B,C,D). This indicates that children's bias towards punishing selfish third parties was reliably increased by the Punish-Selfish prime over their bias towards punishing selfish third parties when they were given the Punish-Either prime. In contrast, the "Punish-Prosocial prime X Third Party Behavior" interaction does not reliably differ from zero at any age, despite generally being somewhat lower than zero (Figure 4B,C,D). This indicates that children's bias towards punishing selfish third parties was *not* reliably *decreased* by the Punish-Prosocial prime.

In the other societies (Berlin, Shuar, Wichí), however, there is a different pattern. For Berlin, the estimated coefficients for both interactions are essentially identical, and neither are reliably different from zero at any age (Figure 4A). This indicates that in Berlin children's bias towards punishing selfish third parties was unaffected by either of the directed norm

primes. Similarly, for the Shuar and Wichí, the estimated coefficients for both norm prime interactions are not reliably different, nor reliably different from zero (Figure 4E,F). Thus, the directed norm primes did not reliably influence children's bias towards punishing selfish third parties.

RQ4: Does the development of children's responsiveness to norm primes about TPP resemble the development of their responsiveness to norm primes about prosocial behavior?

We compared the impact of the directed norm primes (Punish-Selfish, Punish-Prosocial) on children's punishment in the TPPG (i.e. the analysis of RQ2), to the impact of similar directed norm primes on the same children's prosocial behavior in the DG (conducted in the same experimental session, prior to the TPPG; these data have been published previously [12]). In the TPPG, children's punishment of selfish third parties displayed a reliable responsiveness to norm primes by about 6-8 years at the latest, but only in La Plata, Phoenix and Pune (Figure 5A,C,E,G,I,K; same as Figure 3A,C,E,G,I,K). However, in the DG, children displayed a responsiveness to norm primes in all six of the societies (Figure 5B,D,F,H,J,L). This responsiveness to norm primes in the DG emerged in all societies by age 7-8 at the latest, but it emerged much earlier in some societies, perhaps prior to age 4 in Berlin (Figure 5b) and Phoenix (Figure 5f).

Discussion and Conclusions

This study has four primary findings. (1) Children in four of the six societies (Berlin, La Plata, Phoenix, Pune) show a reliable bias towards punishing selfish third parties more than prosocial third parties by 9-10 years of age. (2) In three of the four societies in which

children show this pattern, reliable responsiveness to information about social norms for punishment emerged by about age 6-8. (3) In these three societies, norms directing the punishment of selfishness had a more substantial influence on children's behavior than did norms directing the punishment of prosociality (i.e. norms directing the punishment of selfishness increased children's bias towards punishing selfish third parties *to a greater degree* than norms directing the punishment of prosociality decreased children's bias towards punishing selfish third parties). (4) In the three societies in which norms reliably influenced TPP, this influence increased during middle childhood (about age 6-12) when norms also begin to increasingly influence prosociality. Together, these findings reveal broad similarities across societies in the development of costly punishment and the influence of social norms on punishment. However, they also indicate that the link between punishment and social norms may vary in important ways across societies, and this may be related to societal differences in the content of local norms for punishment and children's understanding of those norms.

In the four societies in which children show a reliable bias towards punishing selfish third parties, this pattern is clearly visible by middle childhood. The Punish-Either norm prime does not direct children towards punishing either of the third parties, so any bias towards punishing selfish third parties must come from the preferences that subjects brought into the experiment. In four of the six societies (Berlin, Phoenix, La Plata and Pune) this bias became reliable by about 9-10 years of age. In prior studies with American children, 6-yr olds were more willing to give up rewards to punish selfish third parties than prosocial third parties [29,31], and 3-yr olds were more willing to forego doing a fun activity to prevent selfish third parties from doing the same activity, and less likely to forego the activity for a prosocial third party [27]. Our estimate for the timing of the divergence

between punishing selfish and prosocial third parties may be conservative, because the Punish-Either norm prime states that punishing both selfish and prosocial third parties is "ok". This may have dampened differences in the likelihood of punishing selfish and prosocial third parties. However, this phrasing of the undirected norm prime also reduces the likelihood that subjects made choices based on what they thought the experimenter wanted them to choose. If such a demand effect would have the greatest impact on younger children, then our design may diminish these effects and provide a more accurate picture of the development of children's preferences for TPP. Future studies should explore this possibility by comparing an 'undirected' norm prime condition to a 'baseline' condition with no norm prime of any kind.

Our results are consistent with prior findings that children develop a bias towards punishing selfish third parties between ages 6 and 8 [29,32]. One study with 7-11 year-old Italian children found no developmental effects in TPP (17), but compared children's (a) willingness to incur costs to punish selfish third parties to their (b) willingness to incur costs that would not affect any third parties. It is possible that the psychological processes affecting decisions about whether or not to punish a third party develop differently from those affecting decisions about whether selfish third parties should be punished more than prosocial third parties.

Children's bias towards punishing selfish third parties increased during middle childhood, and this developmental pattern is similar across societies. Middle childhood is also when children across diverse societies become more prosocial [12,20,38,39] and more averse to advantageous inequity [40]. This raises the possibility that prosociality, advantageous inequity aversion, and TPP may be developmentally coupled. In each of these cases, individuals incur personal costs to produce fairer outcomes for others. One

explanation for this is that children become more responsive to social norms during middle childhood, leading them to become more likely to conform to social norms. This is consistent with our finding that norm primes begin to shape behavior during middle childhood (although TPP norm primes were not effective in all societies). This also matches the timing of children's responsiveness to information about social norms in the DG, in the same set of societies [12]. Children in all six of our populations were more likely to choose 1/1 after receiving a norm prime indicating that 1/1 is 'right', and less likely to choose 1/1 after receiving the norm prime indicating that 2/0 is 'right', and this pattern emerged by 7-8 years in all of these societies (Figure 5B,D,F,H,J,L).

Children's TPP did not respond to information about social norms in some societies (Shuar, Wichí, Berlin), even though children's prosocial behavior in these societies responded to nearly identical norm primes in the DG (Figure 5). For the Shuar and Wichí, differences in responsiveness to norm primes in the DG and the TPPG might reflect the content of children's prior beliefs about norms for punishing third parties. Shuar and Wichí children did not show a reliable bias towards punishing selfish third parties in the undirected prime condition, and this might reflect relatively weak prior beliefs about the appropriateness of punishing selfish third parties.

It is also possible that the *strength* of children's prior beliefs about norms for punishing third parties may affect their responsiveness to norm primes. If children in some societies were relatively *more certain* about their prior beliefs about punishing selfish third parties, then the norm primes would have had relatively *less* impact in those societies. This might be particularly important for explaining why children in Berlin were responsive to norm primes in the DG (Figure 5B), but not in the TPPG (Figure 5A). Children in Berlin showed a tendency to punish selfish third parties more than prosocial third parties in the

TPPG (Figure 2B). If children in Berlin were more *certain* that this was the correct thing to do than children in other societies (i.e. La Plata, Phoenix, Pune), then the behavior of children in Berlin would be less influenced by directed norm primes than the behavior of children in other societies.

Additional evidence that the strength of children's prior beliefs plays a role comes from the fact that the Punish-Selfish norm was more influential on children's choices than the Punish-Prosocial norm. Children's relative unwillingness to punish prosocial choices by third parties in the Punish-Prosocial condition is particularly notable because, by answering a comprehension question asking about the norm prime video, they had already explicitly reported that it is 'good' to punish a third party who is prosocial. There is no particular reason for this pattern of results to be observed unless children were interpreting the content of each norm prime through their prior normative beliefs about punishing selfish and prosocial third parties. Another way to put this is that children in all six of our societies may be certain that punishing *prosocial* third parties is *incorrect*, but children in some of the societies may be uncertain about whether punishing *selfish* third parties is *correct*. If this is true, it would make sense that the Punish-Prosocial norm prime wouldn't sway children's choices even in those societies where the Punish-Selfish norm prime did sway them (i.e. La Plata, Phoenix, Pune).

Together, these results point to broad, but not universal, similarities in the development of costly punishment across societies. They also support the idea that middle childhood is a period during which children's sense of fairness is becoming increasingly responsive to social norms, and that societal variation in different kinds of fair behavior (e.g. prosociality, inequity aversion, TPP) emerges through the development of a universal human norm psychology which motivates individuals to conform to their society's norms.

Works Cited

1. Richerson P *et al.* 2016 Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behav. Brain Sci.* **39**. (doi:10.1017/S0140525X1400106X)

2. Boyd R, Richerson PJ. 1992 Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195. (doi:10.1016/0162-3095(92)90032-Y)

3. Hauert C, Traulsen A, Brandt H, Nowak MA, Sigmund K. 2007 Via Freedom to Coercion: The Emergence of Costly Punishment. *Science* **316**, 1905–1907. (doi:10.1126/science.1141588)

4. Boyd R, Gintis H, Bowles S. 2010 Coordinated Punishment of Defectors Sustains Cooperation and Can Proliferate When Rare. *Science* **328**, 617–620. (doi:10.1126/science.1183665)

5. Fehr E, Gächter S. 2000 Cooperation and Punishment in Public Goods Experiments. *Am. Econ. Rev.* **90**, 980–994.

6. Fehr E, Gachter S. 2002 Altruistic punishment in humans. *Nature* **415**, 137–140. (doi:10.1038/415137a)

7. Herrmann B, Thoni C, Gachter S. 2008 Antisocial Punishment Across Societies. *Science* **319**, 1362–1367. (doi:10.1126/science.1153808)

8. Mathew S, Boyd R. 2011 Punishment sustains large-scale cooperation in prestate warfare. *Proc. Natl. Acad. Sci.* **108**, 11375–11380. (doi:10.1073/pnas.1105604108)

9. Mathew S, Boyd R. 2014 The cost of cowardice: punitive sentiments towards free riders in Turkana raids. *Evol. Hum. Behav.* **35**, 58–64. (doi:10.1016/j.evolhumbehav.2013.10.001)

10. Marlowe FW *et al.* 2008 More 'altruistic' punishment in larger societies. *Proc. R. Soc. B Biol. Sci.* **275**, 587–592. (doi:10.1098/rspb.2007.1517)

11. Henrich J. 2006 Cooperation, Punishment, and the Evolution of Human Institutions. *Science* **312**, 60–61. (doi:10.1126/science.1126398)

12. House BR *et al.* 2020 Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nat. Hum. Behav.* , 1–9. (doi:10.1038/s41562-019-0734-z)

13. Chudek M, Henrich J. 2011 Culture–gene coevolution, norm-psychology and the emergence of human prosociality. *Trends Cogn. Sci.* **15**, 218–226. (doi:10.1016/j.tics.2011.03.003)

14. Bicchieri C. 2006 *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

15. Sylwester K, Herrmann B, Bryson JJ. 2013 Homo homini lupus? Explaining antisocial punishment. *J. Neurosci. Psychol. Econ.* **6**, 167–188. (doi:10.1037/npe0000009)

16. Henrich J *et al.* 2005 In Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies. *Behav. Brain Sci.* **28**, 795–815. (doi:10.1017/S0140525X05000142)

17. Henrich J *et al.* 2006 Costly Punishment Across Human Societies. *Science* **312**, 1767–1770. (doi:10.1126/science.1127333)

18. Henrich J *et al.* 2010 Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. *Science* **327**, 1480–1484. (doi:10.1126/science.1182238)

19. Henrich J, Boyd R, Bowles S, Fehr E, Camerer C, Gintis H. 2004 *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford University Press.

20. House BR, Silk JB, Henrich J, Barrett HC, Scelza BA, Boyette AH, Hewlett BS, McElreath R, Laurence S. 2013 Ontogeny of prosocial behavior across diverse societies. *Proc. Natl. Acad. Sci.* **110**, 14586–14591. (doi:10.1073/pnas.1221217110)

21. House BR, Tomasello M. 2018 Modeling social norms increasingly influences costly sharing in middle childhood. *J. Exp. Child Psychol.* **171**, 84–98. (doi:10.1016/j.jecp.2017.12.014)

22. Kenward B, Dahl M. 2011 Preschoolers distribute scarce resources according to the moral valence of recipients' previous actions. *Dev. Psychol.* **47**, 1054–1064. (doi:10.1037/a0023869)

23. Rakoczy H, Warneken F, Tomasello M. 2008 The Sources of Normativity: Young Children's Awareness of the Normative Structure of Games. *Dev. Psychol.* **44**, 875–881. (doi:10.1037/0012-1649.44.3.875)

24. Riedl K, Jensen K, Call J, Tomasello M. 2015 Restorative Justice in Children. *Curr. Biol.* **25**, 1731–1735. (doi:10.1016/j.cub.2015.05.014)

25. Schmidt MFH, Rakoczy H, Tomasello M. 2019 Eighteen-Month-Old Infants Correct Non-Conforming Actions by Others. *Infancy* **24**, 613–635. (doi:10.1111/infa.12292)

26. Yucel M, Vaish A. 2018 Young children tattle to enforce moral norms. *Soc. Dev.* **27**, 924–936. (doi:10.1111/sode.12290)

27. Yudkin DA, Van Bavel JJ, Rhodes M. 2019 Young children police group members at personal cost. *J. Exp. Psychol. Gen.* (doi:10.1037/xge0000613)

28. Gummerum M, Chu MT. 2014 Outcomes and intentions in children's, adolescents', and adults' second- and third-party punishment behavior. *Cognition* **133**, 97–103. (doi:10.1016/j.cognition.2014.06.001)

29. Jordan JJ, McAuliffe K, Warneken F. 2014 Development of in-group favoritism in children's third-party punishment of selfishness. *Proc. Natl. Acad. Sci.* **111**, 12710–12715. (doi:10.1073/pnas.1402280111)

30. Lergetporer P, Angerer S, Glätzle-Rützler D, Sutter M. 2014 Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proc. Natl. Acad. Sci.* , 201320451. (doi:10.1073/pnas.1320451111)

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

32. Salali GD, Juda M, Henrich J. 2015 Transmission and development of costly punishment in children. *Evol. Hum. Behav.* **36**, 86–94. (doi:10.1016/j.evolhumbehav.2014.09.004)

33. Smith CE, Blake PR, Harris PL. 2013 I Should but I Won't: Why Young Children Endorse Norms of Fair Sharing but Do Not Follow Them. *PLoS ONE* **8**, e59510. (doi:10.1371/journal.pone.0059510)

34. Fehr E, Fischbacher U. 2004 Social norms and human cooperation. *Trends Cogn. Sci.* **8**, 185–190. (doi:10.1016/j.tics.2004.02.007)

35. R Development Core Team. 2018 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. See www.R-project.org.

36. McElreath R. 2015 *Statistical Rethinking*. New York: Chapman and Hall/CRC.

37. Stan Development Team. 2018 *RStan: the R interface to Stan.* See http://mc-stan.org.

38. Cowell JM, Lee K, Malcolm-Smith S, Selcuk B, Zhou X, Decety J. 2017 The development of generosity and moral cognition across five cultures. *Dev. Sci.* **20**, e12403. (doi:10.1111/desc.12403)

39. House BR. 2017 Diverse ontogenies of reciprocal and prosocial behavior: cooperative development in Fiji and the United States. *Dev. Sci.* **20**, e12466. (doi:10.1111/desc.12466)

40. Blake PR *et al.* 2015 The ontogeny of fairness in seven societies. *Nature* **528**, 258–261. (doi:10.1038/nature15703)

Table 1: Populations sampled. For more details see Supplementary Table 1.

Population [Location]; Description	**N (female) / age range (in years)**
German [Berlin, DEU]; Urban	110 (55) / 4.07 - 13.36
Argentinian [La Plata, ARG]; Urban	111 (53) / 5.02 - 13.86
Wichí [Misión Chaqueña, ARG]; Rural, sedentized hunter-gatherers	61 (35) / 6.47 - 13.61
American [Phoenix, USA]; Urban	140 (73) / 4.52 - 12.63
Indian [Pune, IND]; Urban	149 (72) / 4.11 - 13.92
Shuar [Amazonia, ECU]; Rural, small-scale horticulture, hunting	32 (15) / 6.59 - 12.5

Table 2: WAIC analysis. Best fit model is Model 5.

Model	**Model Parameters (DV: Subject Punished [1=yes])**	**WAIC (SE)**	**dWAIC (dSE)**	**AIC weight**
1	**Hypothesis:** Children punish Selfish Third Parties (TPs) and Prosocial TPs differently. **Fixed Effect:** Third Party Behavior (Selfish TP=1, Prosocial TP=0)**. Random Effect:** Actor ID	1441.8 (25.78)	39.5 (14.54)	<0.01
2	**Hypothesis:** Children punish Selfish TPs and Prosocial TPs differently, and this tendency varies across age. **Fixed Effects:** Third Party Behavior * Subject Age (centered)**. Random Effect:** Actor ID	1430.9 (26.80)	28.6 (12.60)	<0.01
3	**Hypothesis:** Children punish Selfish TPs and Prosocial TPs differently, and this tendency is influenced by novel normative information about who to punish. **Fixed Effects:** Third Party Behavior * Dummies for Normative Primes [2: Dummy for Punish-Selfish TP Prime, Dummy for Punish-Prosocial TP Prime; Reference level = Punish-Either Prime. **Random Effect:** Actor ID	1431.8 (27.02)	29.4 (12.32)	<0.01
4	**Hypothesis:** Children punish Selfish TPs and Prosocial TPs differently, this tendency is influenced by normative information, and these effects vary across age. **Fixed Effects:** Third Party Behavior * Subject Age * Dummies for Normative Primes [2]. **Random Effect:** Actor ID	1424.6 (28.74)	22.3 (9.76)	<0.01
5	**Hypothesis:** Children punish Selfish TPs and Prosocial TPs differently, this tendency is influenced by normative information, and these effects vary across both age and societies. **Fixed Effects:** Third Party Behavior * Subject Age * Dummies for Normative Primes [2]. **Random Effects:** Actor ID, Subject's Society	**1402.3 (31.07)**	**0.00 (NA)**	**~1.00**

Figure Captions

Figure 1: Arrangement of the apparatus and testing area (Site: Berlin, Germany).

Figure 2: Estimated probability that children in each society will punish a selfish third party and a prosocial third party, for the Punish-Either norm prime only. This captures children's tendency to punish a selfish third party more than a prosocial third party without receiving normative information directing them towards punishing either. Solid lines represent estimated probabilities for punishing selfish third parties, dashed lines represent estimated probabilities for punishing prosocial third parties. Shaded regions represent 95% Confidence Intervals for these estimates.

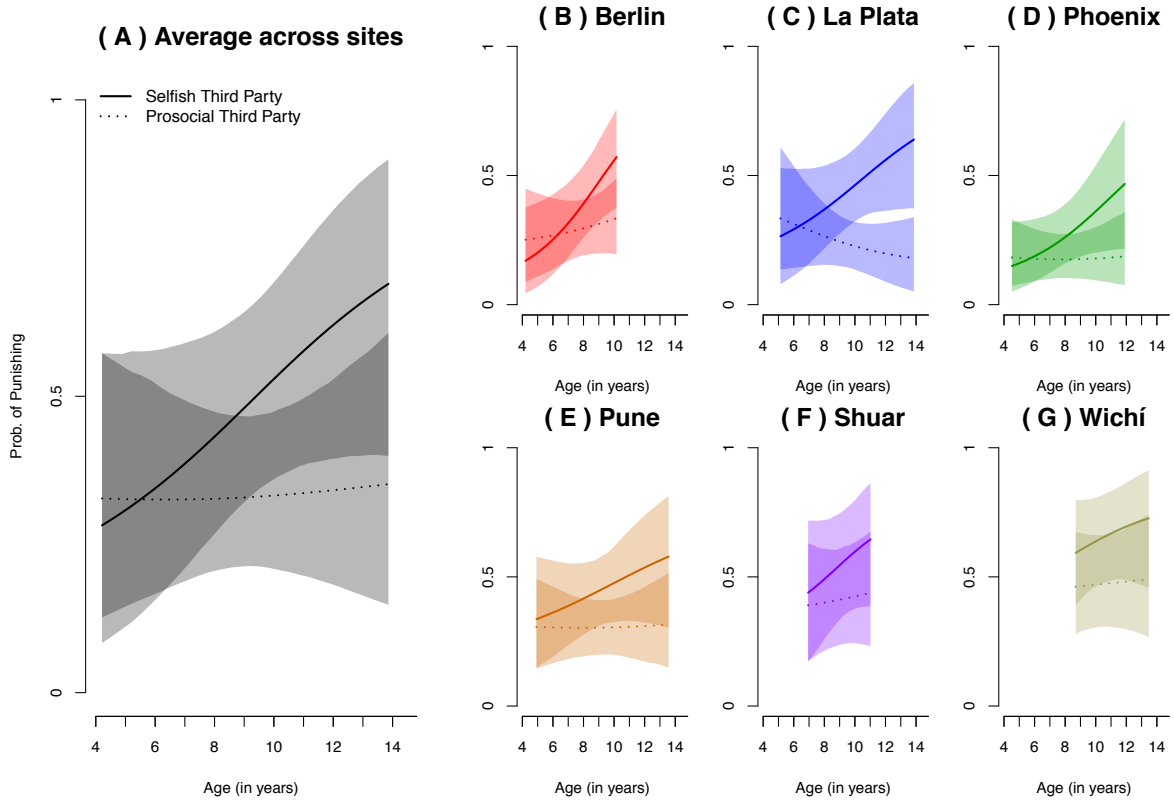

Figure 3: Estimated probability that children in each society will punish selfish third parties (3A,C,E,G,I,K) and prosocial third parties (3B,D,F,H,J,L), across the three norm primes. Solid lines represent estimated probabilities for Punish-Either, dotted lines represent estimated probabilities for Punish-Selfish, and short/long dashed lines represent estimated probabilities for Punish-Prosocial. Shaded regions represent 95% Confidence Intervals for these estimates.

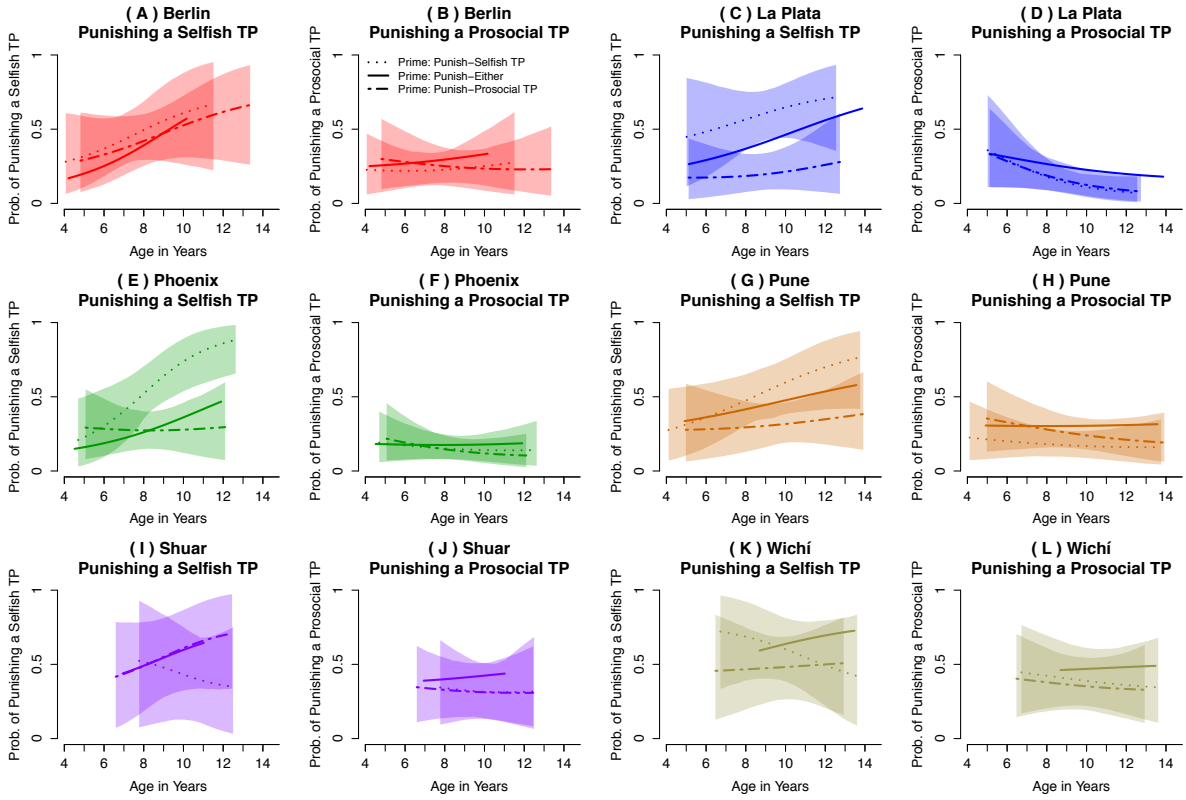

Figure 4: Effect sizes of regression coefficients for the interactions “Punish-Selfish prime X Third Party Behavior” and “Punish-Prosocial prime X Third Party Behavior”, plotted as a function of child age. These capture the degree to which the two directed norm primes (Punish-Selfish and Punish-Prosocial) increase or decrease children’s bias towards punishing selfish third parties more than prosocial third parties, relative to that bias in the Punish-Either prime condition. Solid lines represent estimated probabilities for the Punish-Selfish prime (i.e. “Punish-Selfish X Third Party Behavior” interaction), dashed lines represent estimated probabilities for the Punish-Prosocial prime (i.e. “Punish-Prosocial X Third Party Behavior” interaction). Shaded regions represent 95% Confidence Intervals for these estimates. Where the lines are reliably above zero (4B,C,D), this indicates that the prime (either Punish-Selfish or Punish-Prosocial) has reliably *increased* children’s bias towards punishing selfish third parties more than prosocial third parties, above children’s bias towards this in the Punish-Either prime condition.

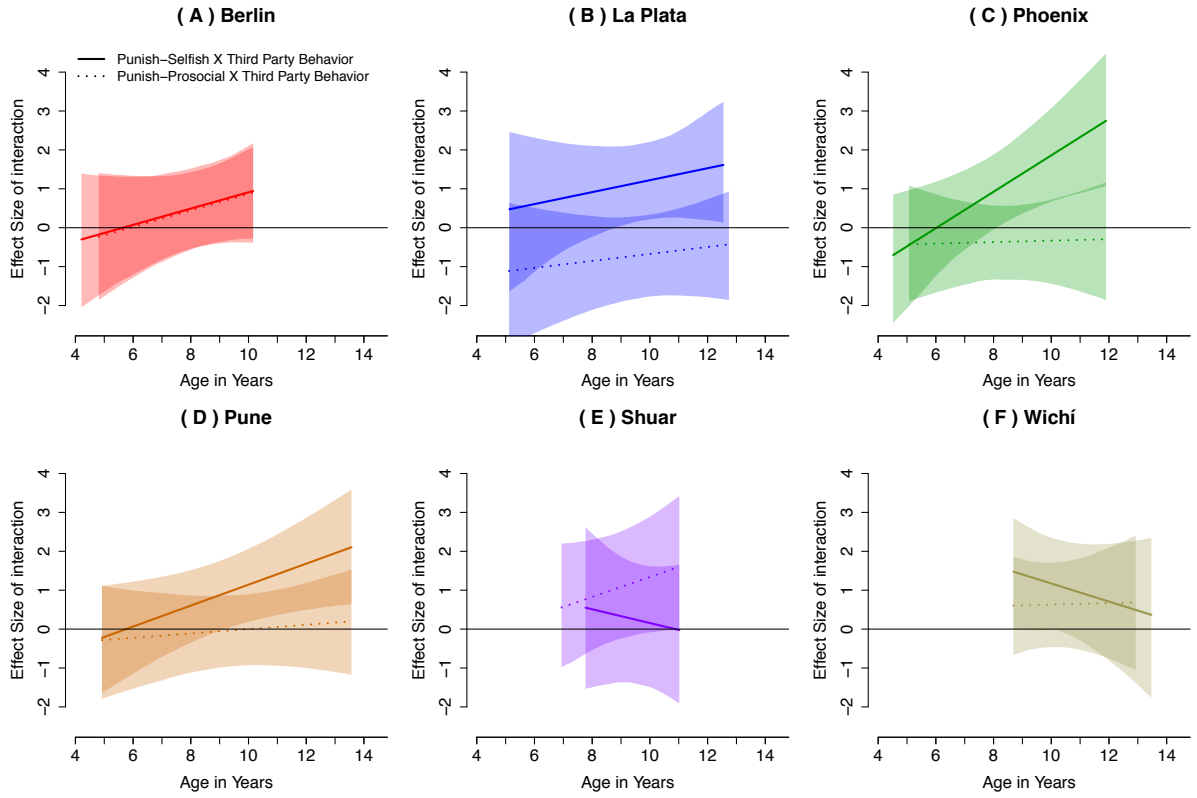

Figure 5: Comparing the influence of the norm primes on children's probability of punishing Selfish third parties (5A,C,E,G,I,K; identical to Figure 3A,C,E,G,I,K), to the influence of nearly identical norm primes on children's probability of making a prosocial choice (5B,D,F,H,J,L), results drawn from a previously published study using similar methods with the same sample of children (House et al., 2019). In each society, children's prosociality was previously responsive to norm primes, even in societies where children's punishment is not responsive to norm primes in the current study.