Uncovering the Heterogeneity behind Cross-Cultural Variation in Antisocial Punishment

Adrian Bruhin* Kelly Janizzi§ Christian Thöni¶

University of Lausanne

November 21, 2019

Abstract

Antisocial punishment in public good games, i.e., punishment of individuals who contributed the same or more than their punisher, varies substantially across cultures. We exploit the data of Herrmann et al. (2008) and estimate a finite mixture model to uncover the heterogeneity behind this variation in a parsimonious way. The finite mixture model reveals that, overall, the population consists of two cleanly segregated punisher types: 35.3% Type AF subjects who engage in antisocial punishment as well as free rider punishment and 64.7% Type F subjects who engage exclusively in free rider punishment. Moreover, we find that in cultures with high levels of antisocial punishment, Type AF subjects are more frequent. Despite its parsimony, this classification of subjects into types predicts mean earnings per group and enhances our understanding of the large variation in the effectiveness of peer punishment across cultures.

Keywords: Antisocial Punishment, Public Good Games, Finite Mixture Models, Heterogeneity, Cultural Variation

JEL Classification: C92, C72, H41

*University of Lausanne, Faculty of Business and Economics (HEC Lausanne), Quartier UNIL-Chamberonne, Bâtiment Internef, CH-1015 Lausanne

§University of Lausanne, Faculty of Business and Economics (HEC Lausanne), Quartier UNIL-Chamberonne, Bâtiment Internef, CH-1015 Lausanne

¶University of Lausanne, FDCA, Quartier UNIL-Chamberonne, Bâtiment Internef, CH-1015 Lausanne; phone: +41 21 692 2843; mail: christian.thoeni@unil.ch; web: http://sites.google.com/site/christianthoeni/; ORCID: 0000-0003-1190-8471

1 Introduction

Many social interactions can be described as a public goods game, where individuals repeatedly contribute at a personal cost to a joint project benefiting everyone in the group. Examples comprise teamwork in firms, civic engagement in school boards or political parties, and neighborhood associations. Repeated public goods games represent a social dilemma in which individual rationality would call for free riding among selfish agents, while the joint payoffs are maximized when all agents contribute fully. The vast literature on experimental public goods games reliably shows that many subjects are willing to contribute initially, but over time contributions decay to low levels (Ledyard, 1995). The reason for this decay is that a majority of subjects can be characterized as conditional cooperators who are willing to contribute if and only if others contribute as well (Fischbacher and Gächter, 2010; Thöni and Volk, 2018).

Peer punishment is a prominent mechanism to counter the decay of cooperation (Yamagishi, 1986; Ostrom et al., 1992; Fehr and Gächter, 2000). In public goods games with punishment subjects have access to a punishment technology, enabling them to punish their peers at a cost after observing their contributions. The early studies in this literature suggested that peer punishment is often deterrent, i.e., sufficiently strong to enforce high contributions from all subjects in the group (Chaudhuri, 2011).

However, subsequent research demonstrated that the beneficial effects of peer punishment should not be taken for granted. One of the problems of the peer punishment mechanism is that subjects might use it not only to punish low contributors. Gächter et al. (2005) in a four country comparison and later Herrmann et al. (2008) in a larger sample document substantial differences in the punishment strategies across subject pools. In particular, they show large differences in the degree to which subjects engage in *antisocial punishment*, the punishment of high contributors.[1]

Unsurprisingly, antisocial punishment is detrimental to the willingness to cooperate and renders peer punishment largely ineffective. Gächter et al. (2010) analyze differences in punishment behavior across cultures. While the level of free rider punishment is fairly similar

[1]Herrmann et al. (2008) define antisocial as punishment of subjects with a weakly higher contribution than the punisher. Related studies use the term *perverse punishment* and define it as punishing above-average contributors (Bochet et al., 2006; Cinyabuguma et al., 2006) or punishing the highest contributor (Casari and Luini, 2009). See Fu and Putterman (2018) for a comparison of the different definitions.

across subject pools, there is substantial cross-cultural variation in the level of antisocial punishment, leading to substantial differences in the effectiveness of peer punishment.

Despite the central role of antisocial punishment for the success of peer punishment mechanisms, its cause and determinants are not well understood. Some argue that it is driven by a competition for status (Sylwester et al., 2013), or a dislike of morally superior acts (Monin, 2007). Experimental evidence suggests that descriptive norms may cause antisocial punishment (Parks and Stone, 2010; Irwin and Horne, 2013). Theoretically, antisocial punishment could stem from inequality averse punishers, who want to avoid earning less than non-punishing subjects (Thöni, 2014).

While the previous literature focused on the motives of antisocial punishment, our main goal is to uncover the *latent heterogeneity* behind the cross-cultural variation in antisocial punishment. In particular, we exploit the data of Herrmann et al. (2008) and estimate a finite mixture model. The data comprises the behavior of 1,120 subjects across 16 subject pools who played repeated four-person public good games with costly punishment. The subject pools are highly diverse and originate form the English Speaking, Protestant European, Southern European, Orthodox/Ex-Communist, Arabic Speaking, and Confucian cultural areas.[2]

Estimating a finite mixture model allows us to take latent heterogeneity parsimoniously into account. The finite mixture model assumes the population to consist of a finite number of types that differ in their punishment behavior. It identifies the prevalent types in the population and characterizes each of them by its relative size and type-specific parameter estimates. Moreover, after identifying and characterizing the prevalent types, the finite mixture model allows us to compute individual probabilities of type-membership and classify each subject into the type that best fits her behavior. Overall, this yields a parsimonious and easy to interpret characterization of the heterogeneity in punishment behavior.

A major advantage of our implementation of the finite mixture model is that we do not need to predefine the distinct types ex-ante. They arise endogenously from the data and capture the different punishment behaviors in a statistically efficient way. To determine the optimal number of types, we estimate versions of the model with different numbers of types and select the one that yields the cleanest classification of subjects into these types relative

[2]Our classification of subject pools into cultures is based on Gächter et al. (2010) and uses measures of cultural proximity by Inglehart and Baker (2000) and more recent updates of the World Cultural Map (http://www.worldvaluessurvey.org/).

to its goodness of fit.

The results reveal that we can characterize the heterogeneity in punishment behavior by two distinct types: Type AF and Type F subjects. Type AF subjects engage in both antisocial and free rider punishment and make up 35.3% of the population. In contrast, Type F subjects make up the remaining 64.7% of the population and exclusively engage in free rider punishment.

The individual classification of subjects into the two types explains the two most important features of punishment behavior across the 16 subject pools. First, it explains the substantial cross-cultural differences in antisocial punishment, because Type AF subjects are substantially more frequent in subject pools with high levels of antisocial punishment. In particular, they make up 44.3–59.0% in the Orthodox/Ex-Communist and Arabic speaking subject pools which exhibit high levels of antisocial punishment. In contrast, Type AF subjects only make up 17.7–22.2% of the Confucian, Protestant European, and English Speaking subject pools where antisocial punishment is much rarer. Second, the classification of subjects into types also explains why the level of free rider punishment remains fairly stable across all subject pools. Since both Type AF and Type F subjects punish free riders to a similar extent, free rider punishment remains stable even if the shares of the two types varies across cultures.

Finally, the individual classification of subjects into types is also a powerful predictor for the effectiveness of peer punishment. For instance, groups consisting exclusively of Type AF subjects earn on average 41.2% less than groups consisting exclusively of Type F subjects. Moreover, regressions reveal that the number of Type AF subjects per group explains 11.9% of the total variation in earnings across groups—considerably more than both the current period of the public good game and the same groups' average contributions in a public good game without punishment.

Our results shed new light on the discussion about self governance of cooperation. While there is evidence that behavior in public goods games without punishment does not vary substantially across cultures (Brandts et al., 2004), there are large and systematic differences in the experiment with punishment (Herrmann et al., 2008). The previous literature attributed the differences to the overall prevalence of antisocial punishment in a subject pool, but did not examine heterogeneity. Our results enhance this strand of the literature. They show that the substantial cross-cultural variation in antisocial punishment can be explained by a classification of subjects into two distinct types that is clean, parsimonious, and yet,

powerful at predicting variations in the effectiveness of peer punishment.

Due to its parsimony and predictive power, our classification of subjects into types could be used not only to inform the development of theoretical models that take behavioral heterogeneity into account but also to predict the effectiveness of policies within firms and societies that rely on peer punishment. Moreover, if the distribution of the two types across cultures turns out to be stable over time, it may even explain some differences in economic development. Societies with a high share of Type AF subjects may have to rely more on formal punishment mechanisms, which require explicit rules that probably are more costly and less flexible to implement.

More broadly, the paper also contributes to the literature that applies finite mixture models to uncover distinct preference types. Early papers in this literature focused on uncovering distinct behavioral strategies in the context of learning (El-Gamal and Grether, 1995) and complex dynamic decision making (Houser et al., 2004). Other studies used finite mixture models to classify subjects into rational and behavioral types in the context of decision making under risk (Bruhin et al., 2010; Fehr-Duda et al., 2010; Conte et al., 2011; Santos-Pinto et al., 2015; Bruhin et al., 2019b). Finally, more closely related studies applied finite mixture models to uncover distinct social preference types, mostly in dictator and ultimatum games (Iriberri and Rey-Biel, 2011, 2013; Breitmoser, 2013; Bruhin et al., 2019a) as well as in fairness games (Conte and Moffatt, 2014). So far, only one study used finite mixture models to analyze heterogeneous contributions in public good games (Conte and Levati, 2014). They analyze behavior in a series of one-shot public good games without punishment to classify subjects into three predefined types: egoists, conditional cooperators, and unconditional cooperators. Our study adds to this literature by classifying subjects into two distinct punishment types that arise endogenously in repeated public good games and predict differences in the effectiveness of peer punishment.

The paper has the following structure. Section 2 discusses the data and presents some descriptive statistics about the subjects and their punishment behavior. Section 3 presents the finite mixture model for uncovering the latent heterogeneity in antisocial punishment and classifying subjects into types. Section 4 presents the main results, the predictions regarding the effectiveness of peer punishment, and some robustness checks. Finally, Section 5 concludes.

Figure 1: Geographical Location of the 16 Subject Pools

2 Data and Descriptive Statistics

This section first summarizes the repeated public good games from Herrmann et al. (2008) and describes the 16 different subject pools. Subsequently, it presents descriptive statistics on the subjects and their punishment behavior.

2.1 Data

Herrmann et al. (2008) report data from repeated public good games with and without punishment. Both games are played over ten periods in groups of four and a partner matching. Subjects are endowed with 20 tokens and the marginal per capita return is 0.4, i.e., each token contributed to the public good yields a return of 0.4 tokens for every member of the group.

Subjects usually first played the game without punishment, followed by the game with punishment. The games with punishment feature a 1:3 punishment technology. After observing all group members' contributions in a given round, each subject decides how many tokens to spend on punishing one or more of her peers. Every token spent on punishment

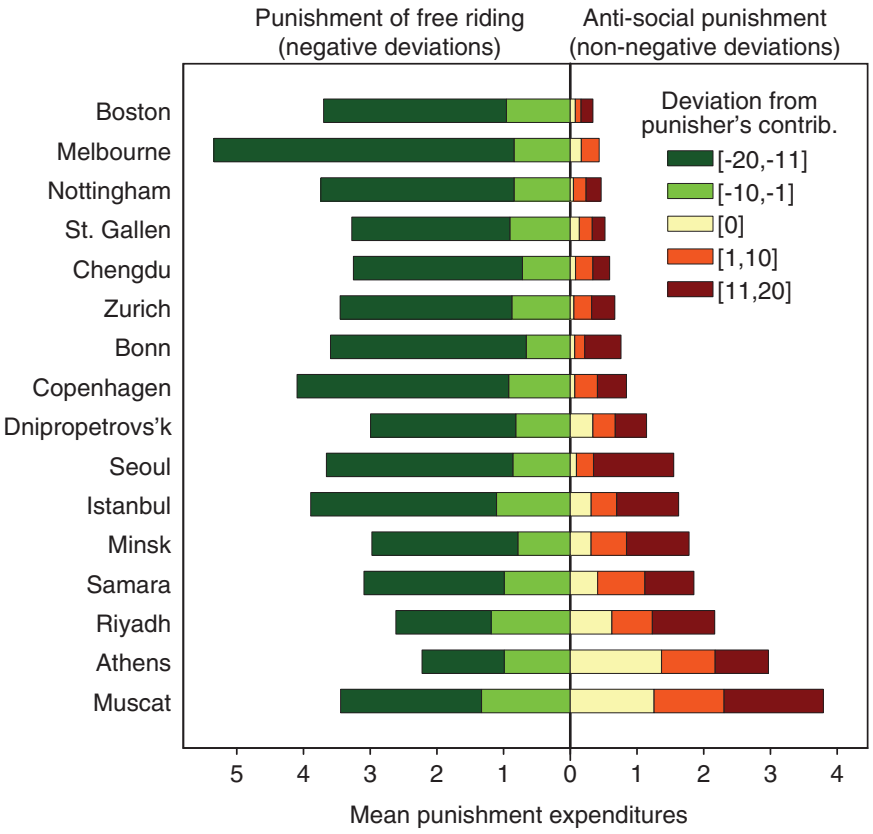
Table 1: Categorization of subject pools into cultural areas and number of subjects

Subject pool	Country	Culture	#subjects
Boston	USA	English Speaking	56
Nottingham	UK		56
Melbourne	Australia		40
Copenhagen	Denmark	Protestant Europe	68
Bonn	Germany		60
Zurich	Switzerland		92
St. Gallen	Switzerland		96
Minsk	Belarus	Orthodox/Ex-communist	68
Dnipropetrovs'k	Ukraine		44
Samara	Russia		152
Athens	Greece	Southern Europe	44
Istanbul	Turkey		64
Riyadh	Saudi Arabia	Arabic Speaking	48
Muscat	Oman		52
Seoul	South Korea	Confucian	84
Chengdu	China		96

reduces the peer's payoff by 3 tokens, while costing one unit to the punisher. We focus our analysis primarily on the punishment behavior in the public good game with punishment. However, in Section 4.4, we also use contributions in the public good game without punishment to predict the groups' earnings in the public good game with punishment.

Herrmann et al. (2008) feature data from 1,120 subjects from 16 culturally diverse subject pools. Figure 1 shows the location of these subject pools on a world map. In all locations subject pools are convenience samples (university students). We follow Gächter et al. (2010) and use measures of cultural proximity to categorize the subject pools into six cultural areas: English Speaking, Protestant Europe, Orthodox/Ex-Communist, Southern Europe, Arabic Speaking, and Confucian. Table 1 shows the categorization along with the number of subjects per subject pool.

Figure 2: Summary of Punishment Behavior across Subject Pools



Source: Herrmann et al. (2008)

2.2 Descriptive Statistics

Figure 2 stems from Herrmann et al. (2008) and summarizes the subjects' punishment behavior in each subject pool. It shows how the mean punishment expenditures depend on the difference between the punished subject's and her punisher's contribution. The right panel exhibits the mean expenditures on antisocial punishment, i.e., punishment of subjects who contributed at least as much as their punisher. The left panel exhibits the mean expenditures on free rider punishment, i.e., punishment of subjects who contributed less than their punisher.

The figure reveals a key pattern in the subjects' punishment behavior that motivates this paper: on the one hand, there is substantial variation across subject pools in antisocial punishment, on the other hand, free rider punishment is remarkably similar. For instance, subjects who contribute more than their punishers receive almost no punishments in Boston, Melbourne, and Nottingham but get heavily punished in Riyadh, Athens, and Muscat. In contrast, (relative) free riders are heavily punished across all subject pools. Our goal is to

analyze whether this pattern is the result of latent heterogeneity in the subjects' punishment behavior and whether uncovering this latent heterogeneity can predict differences in the effectiveness of peer punishment.

3 Finite Mixture Model

This section presents the specification of the finite mixture model which allows us to uncover latent heterogeneity in the subjects' punishment behavior in a parsimonious way. It also explains how we classify subjects into types and how we determine the optimal number of types.

3.1 Specification

The finite mixture model assumes that the population consists of K types of subjects who distinctly differ in their inclination to punish their peers. It uses a Tobit-specification where the dependent variable is subject i's non-negative expenditure for punishing her peer j in period t: $P_{ijt} = \max(P^*_{ijt}, 0)$.[3] The unit of observation is a single punishment decision. As there are three peers and ten periods we observe 30 punishment decisions per subject. The latent type-specific inclination to punish is

$$P^*_{ijt} = \beta_0 + \beta_{1k}(c_{jt} - c_{it}) + \beta_2(c_{jt} - c_{it})I(c_{jt} - c_{it} < 0) + \gamma X_{it} + \epsilon_{it}\,, \tag{1}$$

where, β_{1k} represents subject i's inclination to engage in antisocial punishment when she belongs to type k, i.e., her inclination to punish the peer j in case the peer's contribution, c_{jt}, is at least as high as her own contribution, c_{it}. Similarly, $-(\beta_{1k} + \beta_2)$ is i's type-specific inclination to engage in free rider punishment, i.e., her inclination to punish the peer j in case the peer's contribution falls short of her own contribution. γ measures the effects of a vector X_{it} comprising the following control variables: i's contribution in period t, the average contribution of the other two peers in the group in t, the punishment she received in the previous period $t-1$, as well as the current period t and an indicator for the final period. ϵ_{it} is a normally distributed error term with variance σ^2. To guarantee that the finite mixture model discriminates the types according to their inclination to engage in antisocial

[3] We ignore that P_{ijt} is also censored from above, since only in 1% of all punishment decisions subjects choose to punish maximally.

punishment, and not according to some other dimension, only the parameter β_{1k} is type-specific. All other parameters are common across the K types.

This specification implies that subject i's type-specific contribution to the finite mixture model's density over all 30 punishment decisions, D_i, is

$$f(D_i;\psi_k,\sigma)=\prod_{t=1}^{10}\prod_{j=1}^{3}\left[\frac{1}{\sigma}\phi\left(\frac{P_{ijt}-\hat{P}^*_{ijt}(\psi_k)}{\sigma}\right)\right]^{I(P_{ijt}\geq 0)}\left[1-\Phi\left(\frac{\hat{P}^*_{ijt}(\psi_k)}{\sigma}\right)\right]^{1-I(P_{ijt}\geq 0)},$$

where $\psi_k=(\beta_0,\beta_{1k},\beta_2,\gamma)$ is the type-specific parameter vector which determines the predicted inclination to punish, $\hat{P}^*_{ijt}(\psi_k)$. ϕ represents the PDF and Φ the CDF of the standard normal distribution.

We do not directly observe to which of the K types subject i belongs. Thus, we have to weight her type-specific density contributions by the corresponding ex-ante probabilities of type-membership, π_k, to obtain her likelihood contribution to the finite mixture model,

$$\ell(\Psi;D_i)=\sum_{k=1}^{K}\pi_k\, f(D_i;\psi_k,\sigma)\,,$$

where the vector $\Psi=(\psi_1,\ldots,\psi_K,\sigma,\pi_1,\ldots,\pi_{K-1})$ comprises all parameters that need to be estimated, and $\pi_K=1-\sum_{k=1}^{K-1}\pi_k$. Note that the ex-ante probabilities of type-membership are the same across all subjects and correspond to the shares of the types among the population.[4]

3.2 Classification of Subjects into Types

Once we estimated the parameters of the finite mixture model, we can classify each subject into the type she most likely belongs to, given her punishment behavior, D_i, and the estimated parameters, $\hat{\Psi}$. To do so, we apply Bayes' rule and obtain subject i's individual ex-post probabilities of type-membership,

$$\tau_{ik}=\frac{\hat{\pi}_k\, f(D_i;\hat{\psi}_k,\hat{\sigma})}{\sum_{m=1}^{K}\hat{\pi}_m\, f(D_i;\hat{\psi}_m,\hat{\sigma})}\,. \tag{2}$$

[4]Since i's likelihood contribution is highly non-linear, we apply the expectation maximization (EM) algorithm to obtain the model's maximum likelihood estimates (Dempster et al., 1977). The EM algorithm proceeds iteratively in two steps: In the E-step, it computes the individual ex-post probabilities of type-membership given the actual fit of the model (see equation 2). In the subsequent M-step, it updates the fit of the model by using the previously computed ex-post probabilities to maximize each types' log likelihood contribution separately.

Subsequently, we classify the subject into the type with the highest ex-post probability of type-membership.

3.3 Determining Optimal Number of Types

The individual ex-post probabilities of type-membership also help us to find the optimal number of types, K^*, that represent the best compromise between the model's flexibility to capture behavioral heterogeneity and its parsimony.

On the one hand, if K is too small, the model lacks the flexibility to cope with the heterogeneity in the data and may disregard minority types. If K is too large, on the other hand, the model is overspecified and tries to capture types that do not exist. Such an overspecified model results in considerable overlap between the estimated types and an ambiguous classification of subjects. Thus, we select the optimal number of types, K^*, such that it provides the model with a good fit to the data and, at the same time, also yields an unambiguous classification of subjects.

We apply the normalized entropy criterion (NEC) to summarize the ambiguity in the individual classification of subjects into preference types (Celeux and Soromenho, 1996; Biernacki et al., 1999). The NEC allows us to select the finite mixture model with $K > 1$ types that yields the cleanest possible classification of subjects into types relative to its fit. The NEC for K preference types,

$$NEC(K) = \frac{E(K)}{L(K) - L(1)}\,,$$

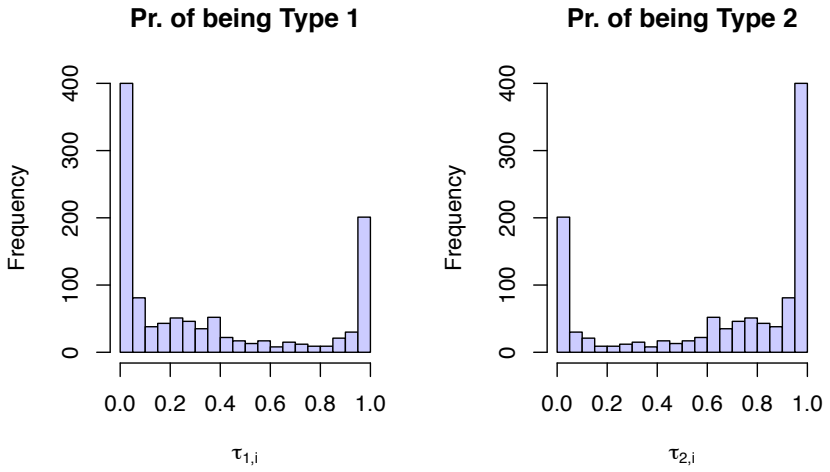
uses the entropy,

$$E(K) = -\sum_{k=1}^{K}\sum_{i=1}^{N} \tau_{ik}\, \ln \tau_{ik} \geq 0\,,$$

normalized by the difference in the log likelihood between the finite mixture model with K types, $L(K)$, and a model with just one type, $L(1)$. The entropy, $E(K)$, quantifies the ambiguity in the ex-post probabilities of type-membership, τ_{ik}. If all τ_{ik} are either close to one or close to zero, implying that each subject is classified unambiguously into exactly one type, $E(K)$ is close to zero. However, if many τ_{ik} are close to $1/K$, implying that many subjects cannot be cleanly assigned to one type, $E(K)$ is large.

One disadvantage of the NEC is that it is not defined in case of $K = 1$. Hence, we cannot apply the NEC to decide whether estimating a finite mixture model is meaningful in the first place or whether we should instead estimate a representative agent model with just one type.

Figure 3: Clean Classification of Subjects into $K^* = 2$ Types



The figure shows the distribution of the individual ex-post probabilities of type-membership τ_{ik} (see equation (2)) for the finite mixture model with $K^* = 2$ types. Most subjects' τ_{ik} are either close to 1 or 0, indicating that most subjects are cleanly classified into one of the two types. Thus, the two types are distinct and exhibit only minor overlap.

However, to make that decision, we analyze whether the classification we derive from the finite mixture model with the optimal number of types successfully predicts differences in the effectiveness of peer punishment in Section 4.4.

4 Results

We start by discussing the finite mixture model's optimal number of types and estimated parameters. Subsequently, we analyze the distribution of types across subject pools and cultures and assess how well the classification of subjects predicts differences in the effectiveness of peer punishment. Finally, we present three robustness checks.

4.1 Optimal Number of Types

To determine the optimal number of types, K^*, we estimate the finite mixture model with $K = 2$ and $K = 3$ types and analyze how cleanly subjects are classified into these types.

The model with $K = 2$ types yields a clean classification of subjects into types. Figure 3 shows the corresponding distribution of the individual ex-post probabilities of type-membership, τ_{ik}. The figure reveals that the finite mixture model classifies most subjects

unambiguously into one of the two types with the corresponding τ_{ik} lying in the vicinity of one. Thus, the two types types are distinct and exhibit only minor overlap. The normalized entropy criterion confirms that the classification of subjects into types is clean ($NEC = 0.363$).

In contrast, the model with $K = 3$ types yields a substantially more ambiguous classification of subjects into types. Figure A1 in Appendix A.1 shows that the individual ex-post probabilities of type-membership are ambiguous for a large number of subjects, particularly in the third type. This ambiguity indicates that there is substantial overlap between the three types and the model may overfit the data. The $NEC = 0.536$ confirms this higher degree of ambiguity. Thus, we conclude that the model with $K^* = 2$ types represents the best compromise between parsimony and flexibility. Moreover, we will show in Section 4.4, that the parsimonious classification of subjects into types predicts differences in the effectiveness of peer punishment and, thus, also outperforms a representative agent model with just one type.

4.2 Estimated Parameters

Table 2 exhibits the estimated parameters of the finite mixture model with $K^* = 2$ types. The finite mixture model uses the Tobit-specification discussed in equation (1) where the dependent variable is subject i's expenditure to punish her peer j in period t.[5]

Type AF subjects make up 35.3% of the population and engage in antisocial punishment as well as in free rider punishment. Their estimated inclination to engage in antisocial punishment, $\hat{\beta}_{1k}$, is positive. That is, when a peer at least matches their own contribution, Type AF subjects punish the peer for an attempt to contribute even more. At the same time their estimated inclination to engage in free rider punishment, $-(\hat{\beta}_{1k} + \hat{\beta}_2)$, is also positive. That is, when the peer's contribution falls short of their own, Type AF subjects punish the peer for an attempt to contribute even less.

Type F subjects, on the other hand, make up 64.7% of the population and engage exclusively in free rider punishment. As their estimated inclination to engage in antisocial punishment is negative while the one to engage in free rider punishment is positive, Type F subjects tend to increase their punishment whenever a peer attempts to reduce his contri-

[5]Note that the signs and the significance levels of the parameter estimates have a direct interpretation. However, due to the Tobit-specification, the size of their marginal effects on P_{ijt} is non-linear.

Table 2: Estimates of the Finite Mixture Model

Dependent variable[a]: i's expenditure to punish j in period t	Type AF	Type F
Type-specific estimates		
Share among the population (π_k)[b]	0.353	0.647
	(0.041)	(0.041)
Inclination to antisocial punishment (β_{1k})	0.116***	-0.279***
	(0.023)	(0.026)
Inclination to free rider punishment $-(\beta_{1k} + \beta_2)$[c]	0.201***	0.597***
	(0.035)	(0.028)
Common estimates		
i's contribution (γ_1)	-0.201***	
	(0.016)	
Mean contribution other two group members (γ_2)	0.062***	
	(0.015)	
Received punishment in $t-1$ (γ_3)	0.214***	
	(0.022)	
Period (γ_4)	-0.111***	
	(0.018)	
Final period (γ_5)	0.601***	
	(0.168)	
Constant (β_0)	-1.958***	
	(0.240)	
Standard deviation of error term (σ)	3.661	
	(0.141)***	
Number of subjects	1,120	
Number of observations	33,600	
Log likelihood	-28,108.47	
AIC	56,238.94	
BIC	56,331.59	

Individual-specific cluster robust standard errors in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

[a] Since the dependent variable is non-negative, we use the Tobit specification shown in equation (1).

[b] There are no significance stars as relative sizes lie within the interval $[0, 1]$.

[c] Estimates and standard errors are based on the Delta-method.

bution.

The parameter estimates common to both types suggest that subjects tend to punish less when they already spent a lot on their own contribution and more when the mean contribution of the other group members is higher. There is also some evidence for revenge, as subjects who got punished in the previous period tend to punish in the current period. Finally, expenditures on punishment decline over time but shoot up again in the final period.

4.3 Distribution of Types across Subject Pools and Cultures

In this section, we analyze how the two types of subjects are distributed across subject pools and cultures. To do so, we use the individual ex-post probabilities of type-membership, τ_{ik}, to classify each subject into the type best describing her behavior. Subsequently, we analyze how the share of Type AF subjects among the population differs across subject pools and cultures.

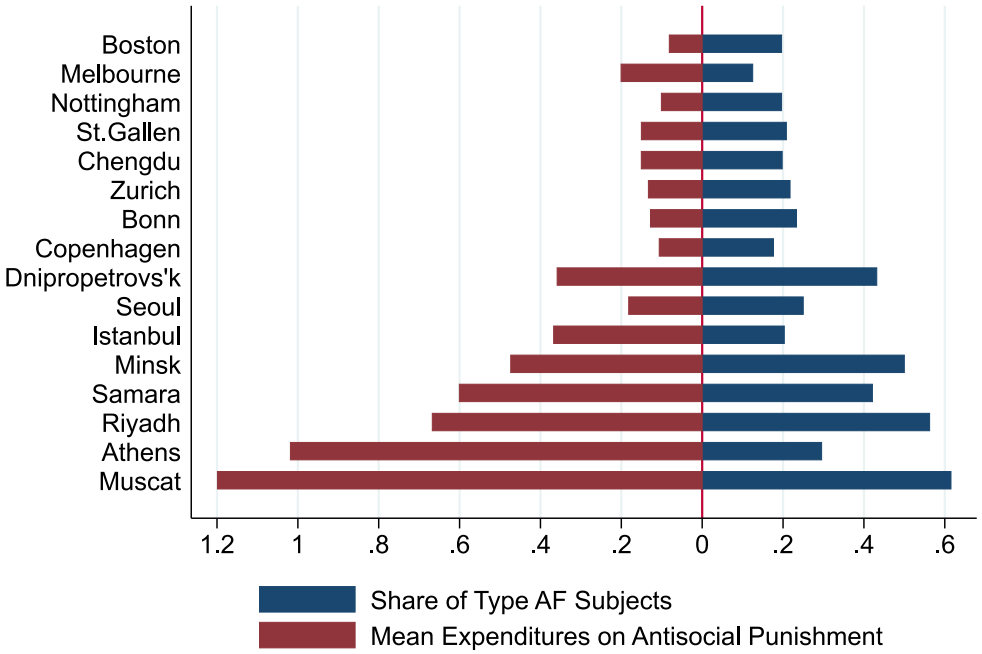
Panel (a) of Figure 4 shows the share of type AF subjects across the 16 subject pools along with the mean expenditure on antisocial punishment. The figure reveals that Type AF subjects make up a substantially higher share of the population in subject pools that exhibit a high mean expenditure on antisocial punishment. For instance, they make up more than 50% of the population in Muscat and Riyadh but less than 20% in Nottingham and Copenhagen.

Similarly, Panel (b) of Figure 4 shows the distribution of type AF subjects across the six cultures. Again, Type AF subjects make up a much higher share in cultures with a high mean expenditure on antisocial punishment. In particular, they constitute 44.3–59.0% of the population in the Orthodox/Ex-Communist and Arabic Speaking subject pools which exhibit high mean expenditures on antisocial punishment. In contrast, Type AF subjects only make up 17.7–22.2% of the population in the Confucian, Protestant European, and English Speaking subject pools where mean expenditures on antisocial punishment are substantially lower. The only exception is the Southern European subject pool, where Type AF subjects only make up 24.1% of the population but the mean expenditure on antisocial punishment is the second highest across all cultures.

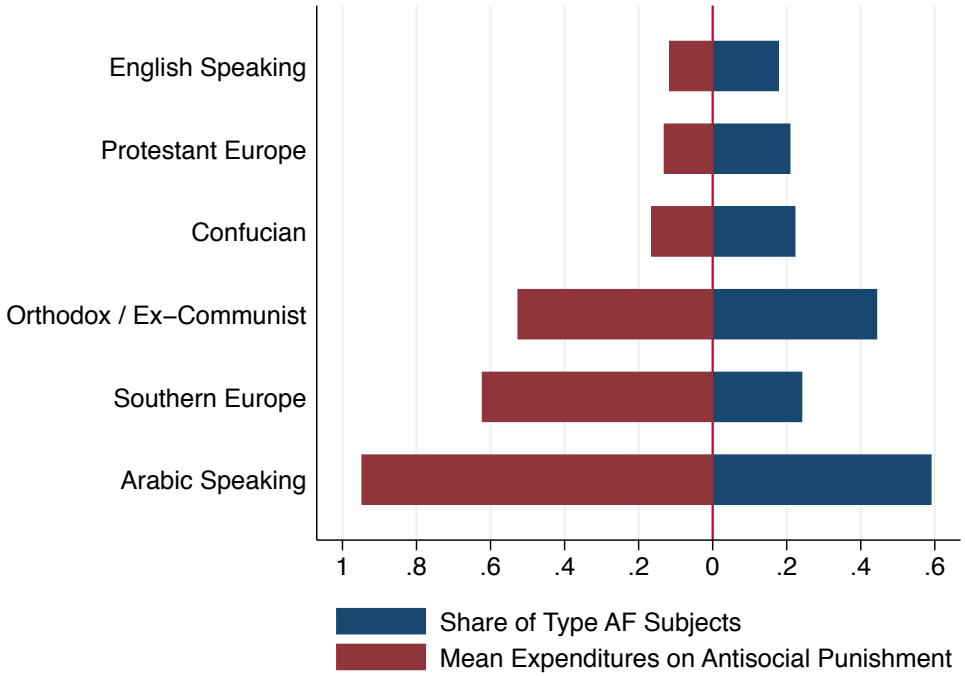
Overall, the distribution of types across subject pools and cultures explains the key pattern in the subjects' punishment behavior (see Figure 2). On the one hand, there is substantial variation in antisocial punishment across subject pools, because the share of Type

Figure 4: Share of Type AF Subjects and Mean Expenditure on Antisocial Punishment

(a) By Subject Pool



(b) By Culture



AF subjects tends to be higher in subject pools with a high mean expenditure on antisocial punishment. On the other hand, free rider punishment is remarkably stable, because both Type AF and Type F subjects punish free riders and, hence, differences in their share among the population matter much less for the mean expenditure on free rider punishment.

4.4 Predicting the Effectiveness of Peer Punishment

In a next step we assess the power of the subjects' classification into types to predict the effectiveness of peer punishment. Ideally, the parsimonious classification we obtain from the finite mixture model not only describes the key pattern in punishment behavior but also predicts the implications of this behavior for the effectiveness of peer punishment.

If the classification of subjects into types predicts the effectiveness of peer punishment, we expect groups with a high number of Type AF subjects to earn less on average. Since the Type AF subjects engage in antisocial punishment, they (i) directly reduce their own and their peers' earnings and (ii) undermine their peers' willingness to make high contributions. These two effects should dampen average earnings in groups with many Type AF subjects.

Figure 5 reveals that, in fact, there is a negative relationship between a group's mean earnings and its number of Type AF subjects. In particular, average earnings per subject are 15.00 tokens in groups with four Type AF subjects vs. 25.49 tokens in groups with no Type AF subjects. Thus, groups consisting exclusively of Type AF subjects earn on average 41.2% less than groups consisting exclusively of Type F subjects.[6]

Next, we examine how well the classification of subjects into types predicts differences in earnings relative to other variables. To do so, we estimate a linear regression model in which we explain a groups earnings in each period by adding the following explanatory variables: (i) the current period and an indicator for the final period, (ii) the group's mean earnings in the public good game without punishment during the corresponding period, and (iii) the number of Type AF subjects in the group.

Table 3 shows the results. The first column reveals that the current period and the indicator for the final period explain a mere $R^2 = 3.1\%$ of the total variance in mean earnings. The second column shows that, when we add the group's mean earnings in the public good

[6]Figure A2 in Appendix A.2 shows how the number of Type AF subjects in the group is related to mean expenditures on punishment (Panel a) and to mean contributions (Panel b). It confirms that a higher number of Type AF subjects in the group reduces mean earnings not only directly via higher punishments but also indirectly via lower mean contributions.

Figure 5: Mean Earnings by Number of Type AF Subjects in Group

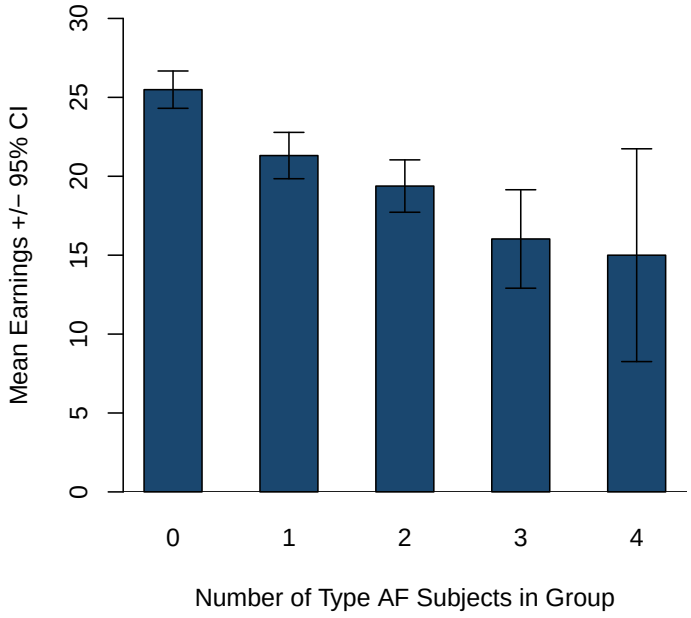


Table 3: Regressions to Assess Predictive Power

Dependent variable: mean earnings in group g in t	(1)	(2)	(3)
Period t	0.662***	0.845***	0.868***
	(0.064)	(0.073)	(0.073)
Indicator for final period	-4.411***	-3.718***	-3.632***
	(0.551)	(0.580)	(0.578)
Mean Earnings in t in public good game without punishment		0.551***	0.619***
		(0.126)	(0.111)
Number of Type AF subjects in group			-3.150***
			(0.389)
Constant	18.170***	3.180	5.112*
	(0.570)	(3.424)	(3.093)
R^2	0.031	0.063	0.181
Number of groups	269	269	269
Number of observations	2,690	2,690	2,690

Group-specific cluster robust standard errors in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

game without punishment during the current period as an additional explanatory variable, the R^2 raises slightly by 3.2 percentage points to 6.3%. However, the third column reveals that, when we add the number of Type AF subjects in the group as another explanatory variable, the R^2 almost triples to 18.1%. This jump in the fraction of the explained variance confirms that the finite mixture model's parsimonious classification of subjects into types exhibits substantial predictive power and explains 11.9% of the total variance in mean earnings—way more than any of the other explanatory variables.

4.5 Robustness Checks

After discussing the main results and showing that the finite mixture model's parsimonious classification of subjects into types predicts differences in the effectiveness of peer punishment, we now present robustness checks to address three important concerns.

4.5.1 Stability of Types

First, we assess whether the classification of subjects into types is stable over the course of the experiment. This is an important question for two reasons. First, only if the classification is stable, it can be used for making behavioral predictions in other contexts. Second, analyzing the stability of the classification also sheds light on whether or not subjects learn and adapt their punishment behavior over the course of the experiment.

To assess the stability of the classification of subjects into types, we re-estimate the finite mixture model with a modified unit of classification. Instead of forcing the model to classify all observations of a subject into the same type, we allow it to classify the observations of the first five periods into a different type than the observations of the final five periods. Thus, every subject features two ex-post probabilities of type-membership, one for the first five periods and one for the final five periods.

The transition matrix below reveals how many subjects remain classified into the same type over the course of the experiment and how many switch types.

		Type-membership in periods 6-10	
		Type F	Type AF
Type-membership in periods 1-5	Type F	701	105
	Type AF	161	153

Overall, the classification of subjects into types remains fairly stable: 854 subjects, making up 76.3% of the population, are on the main diagonal and, thus, remain in the same type over the course of the experiment. Moreover, a χ^2-test of independence rejects the null hypothesis that type-membership in the first five periods is independent of type-membership in the final five periods (p-value < 0.001). Hence, we conclude that the classification of subjects into types is mostly stable and that there is no evidence of subjects learning and adapting their behavior.[7]

4.5.2 Flexibility of Specification

Next, we address the concern that the specification of the finite mixture model may be too rigid. Since it only allows for type-specific heterogeneity in the inclination to engage in antisocial punishment, β_{1k}, potential heterogeneity in other dimensions could bias the results.

To investigate this concern, we split the sample into Type AF and Type F subjects and separately estimate standard Tobit models in each of the resulting two subsamples. Since the two estimations are independent of each other, the parameters can freely vary across the two subsamples.

Table 4 shows the results of the two separate Tobit models. The estimates for the two types' inclination to engage in antisocial and free rider punishment remain qualitatively unchanged compared to their type-specific counterparts from the finite mixture model. The Type AF subjects' estimates are somewhat smaller in absolute values than those from the finite mixture model but remain significant. The type F subjects' estimates remain almost unchanged and are highly significant. Not surprisingly, the other parameters, which are common to both types in the finite mixture model, now vary somewhat between the types. However, with the exception of the estimates for the influence of the final period and for the constant, the differences are small. Overall, the results remain robust when we estimate separate Tobit models in the subsamples of Type AF and Type F subjects. Thus, there is no evidence that the specification of the finite mixture model is too rigid.

[7] The parameter estimates of the corresponding finite mixture model can be found in Table A1 in Appendix A.3. They are virtually identical to the ones of the baseline model in Table 2.

Table 4: Tobit Models Separately Estimated for Type AF and Type F Subjects

Dependent variable[a]: i's expenditure to punish j in period t	Type AF	Type F
Inclination to antisocial punishment (β_{1k})	0.0615**	-0.242***
	(0.025)	(0.022)
Inclination to free rider punishment $-(\beta_{1k}+\beta_2)$[b]	0.146***	0.577***
	(0.024)	(0.018)
i's contribution (γ_1)	-0.222***	-0.172***
	(0.028)	(0.147)
Mean contribution of other group members (γ_2)	0.072***	0.053***
	(0.023)	(0.014)
Received punishment in $t-1$ (γ_3)	0.272***	0.150***
	(0.036)	(0.022)
Period (γ_4)	-0.176***	-0.071***
	(0.032)	(0.020)
Final period (γ_5)	0.395	0.664***
	(0.269)	(0.199)
Constant (β_0)	-1.210**	-2.202***
	(0.446)	(0.262)
Standard deviation of error term (σ)	4.168***	3.244***
	(0.226)	(0.160)
Number of subjects	335	785
Number of observations	10,050	23,550
Log likelihood	-10,819.04	-16,501.59

Individual-specific cluster robust standard errors in parentheses (* $p<0.1$, ** $p<0.05$, *** $p<0.01$).

[a] Since the dependent variable is non-negative, we use a Tobit specification.

[b] Estimates and standard errors are based on the Delta-method.

4.5.3 Type-Membership and Observable Characteristics

Finally, we address the concern that the type-specific heterogeneity may not be latent but rather a function of the subjects' observable characteristics. If that were the case, estimating a finite mixture model would not be necessary and the subjects' type-membership could be predicted based on their observable characteristics.

To study this concern, we analyze whether the subjects' observable characteristics predict their individual ex-post probabilities of type-membership, τ_{ik}. The data contains a number of individual characteristics, such as gender, age, and socioeconomic variables (see Table S2 in the Supporting Material of Herrmann et al., 2008). Table A2 in Appendix A.4 shows the results of the corresponding linear probability models. Most observable characteristics are not significantly correlated with the probability of being a Type AF subject. The only exceptions are the indicator whether the subject has an urban background, the subject's gender, and some indicators for the cultures. Subjects with an urban background are, depending on the specification of the linear probability model, 5.6–6.0 percentage points less likely to belong to Type AF. Women, on the other hand, are overall 4.4 percentage points more likely to belong to Type AF. The specification of the linear probability model which interacts the subjects' gender with their culture reveals that Protestant and Southern European women drive the overall gender effect. This result is in line with the previous finding that higher levels of economic development and gender equality favor the manifestation of gender differences in preferences across countries (Falk and Hermle, 2018). However, overall, the predictive power of the subjects' observable characteristics is fairly limited and barely exceeds 10% of the total variation in the individual ex-post probabilities of type-membership. Hence, individual type-membership seems to be largely driven by latent heterogeneity and estimating a finite mixture model is justified.

5 Conclusion

In this paper, we present a parsimonious characterization of the heterogeneity in punishment behavior by two distinct types of subjects. Overall, there are 35.3% Type AF subjects and 64.7% type F subjects.

Despite its parsimony, our classification of subjects into types gives a possible explanation for the observed cultural differences in the levels of antisocial punishment and cooperation

found by previous studies (Brandts et al., 2004; Herrmann et al., 2008; Gächter and Herrmann, 2009). In particular, it explains why the intensity of antisocial punishment varies substantially across cultures while the intensity of free rider punishment is mostly stable.

Moreover, our classification also predicts differences in the effectiveness of peer punishment and has direct policy implications. Managers of multinational firms and policy makers should take into account that in some cultures peer punishment might not be effective due to a large share of Type AF subjects who punish prosocial acts and undermine the willingness to contribute to the public good. Hence, depending on the share of Type AF subjects, in some cultures there may exist a need to rely predominantly on third-party punishment, whereas in other cultures relying on peer punishment could be more effective.

Relatedly, if the distribution of types across cultures is stable over prolonged time horizons, it may have developmental consequences. Cultures with a high share of Type AF subjects could benefit less from civic engagement and the spontaneous formation of social capital. Instead, they may need to rely predominantly on explicit rules and centralized punishment which are more rigid and arguably more costly to enforce.

References

Biernacki, C., G. Celeux, and G. Govaert (1999): "An improvement of the NEC criterion for assessing the number of clusters in a mixture model," *Pattern Recognition Letters*, 20, 267–272.

Bochet, O., T. Page, and L. Putterman (2006): "Communication and punishment in voluntary contribution experiments," *Journal of Economic Behavior & Organization*, 60, 11–26.

Brandts, J., T. Saijo, and A. Schram (2004): "How universal is behavior? A four country comparison of spite and cooperation in voluntary contribution mechanisms," *Public Choice*, 119, 381–424.

Breitmoser, Y. (2013): "Estimating social preferences in generalized dictator games," *Economics Letters*, 121, 192–197.

Bruhin, A., E. Fehr, and D. Schunk (2019a): "The Many Faces of Human Sociality - Uncovering the Distribution and Stability of Social Preferences," *Journal of the European Economic Association*, 17, 1025–1069.

Bruhin, A., H. Fehr-Duda, and T. Epper (2010): "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion," *Econometrica*, 78, 1375–1412.

Bruhin, A., M. Manai, and L. Santos-Pinto (2019b): "Risk and Rationality: The Relative Importance of Probability Weighting and Choice Set Dependence," Tech. rep., Working Paper.

Casari, M. and L. Luini (2009): "Cooperation under alternative punishment institutions: An experiment," *Journal of Economic Behavior & Organization*, 71, 273–282.

Celeux, G. and G. Soromenho (1996): "An entropy criterion for assessing the number of clusters in a mixture model," *Journal of Classification*, 195–212.

Chaudhuri, A. (2011): "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature," *Experimental Economics*, 14, 47–83.

Cinyabuguma, M., T. Page, and L. Putterman (2006): "Can second-order punishment deter perverse punishment?" *Experimental Economics*, 9, 265–279.

Conte, A., J. D. Hey, and P. G. Moffatt (2011): "Mixture Models of Choice Under Risk," *Journal of Econometrics*, 162, 79–88.

Conte, A. and V. Levati (2014): "Use of Data on Planned Contributions and Stated Beliefs in the Measurement of Social Preferences," *Theory and Decision*, 76, 201–223.

Conte, A. and P. G. Moffatt (2014): "The Econometric Modelling of Social Preferences," *Theory and Decision*, 76, 119–145.

Dempster, A., N. Liard, and D. Rubin (1977): "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–38.

El-Gamal, M. A. and D. M. Grether (1995): "Are People Bayesian? Uncovering Behavioral Strategies," *Journal of the American Statistical Association*, 90, 1137–1145.

Falk, A. and J. Hermle (2018): "Relationship of gender differences in preferences to economic development and gender equality," *Science*, 362, eaas9899.

Fehr, E. and S. Gächter (2000): "Cooperation and punishment in public goods experiments," *American Economic Review*, 90, 980–994.

Fehr-Duda, H., A. Bruhin, T. Epper, and R. Schubert (2010): "Rationality on the rise: Why relative risk aversion increases with stake size," *Journal of Risk and Uncertainty*, 40, 147–180.

Fischbacher, U. and S. Gächter (2010): "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments," *American Economic Review*, 100, 541–556.

Fu, T. and L. Putterman (2018): "When is punishment harmful to cooperation? A note on antisocial and perverse punishment," *Journal of the Economic Science Association*, 4, 151–164.

Gächter, S. and B. Herrmann (2009): "Reciprocity, culture, and human cooperation: Previous insights and a new cross-cultural experiment," *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364, 791–806.

Gächter, S., B. Herrmann, and C. Thöni (2005): "Cross-cultural differences in norm enforcement," *Behavioral and Brain Sciences*, 28, 822–823.

——— (2010): "Culture and cooperation," *Philosophical Transactions of the Royal Society B*, 365, 2651–2661.

Herrmann, B., C. Thöni, and S. Gächter (2008): "Antisocial punishment across societies," *Science*, 319, 1362–1367.

Houser, D., M. Keane, and K. McCabe (2004): "Behavior in a Dynamic Decision Problem: An Analysis of Experimental Evidence Using a Bayesian Type Classification Algorithm," *Econometrica*, 72, 781–822.

Inglehart, R. and W. E. Baker (2000): "Modernization, Cultural Change, and the Persistence of Traditional Values," *American Sociological Review*, 65, 19–51.

Iriberri, N. and P. Rey-Biel (2011): "The Role of Role Uncertainty in Modified Dictator Games," *Experimental Economics*, 14, 160–180.

——— (2013): "Elicited Beliefs and Social Information in Modified Dictator Games: What Do Dictators Believe Other Dictators Do?" *Quantitative Economics*, 4, 515–547.

Irwin, K. and C. Horne (2013): "A normative explanation of antisocial punishment," *Social Science Research*, 42, 562–570.

Ledyard, J. O. (1995): "Public Goods: A Survey of Experimental Research," in *The Handbook of Experimental Economics*, ed. by J. H. Kagel and A. E. Roth, Princeton University Press, 111–194.

Monin, B. (2007): "Holier than me? Threatening social comparison in the moral domain," *Revue Internationale de Psychologie Sociale*, 20, 53–68.

Ostrom, E., J. Walker, and R. Gardner (1992): "Covenants With and Without a Sword: Self-Governance Is Possible," *American Political Science Review*, 86, 404–17.

Parks, C. D. and A. B. Stone (2010): "The desire to expel unselfish members from the group," *Journal of Personality and Social Psychology*, 99, 303–310.

Santos-Pinto, L., A. Bruhin, J. Mata, and T. Astebro (2015): "Detecting heterogeneous risk attitudes with mixed gambles," *Theory and Decision*, 79, 573–600.

Sylwester, K., B. Herrmann, and J. J. Bryson (2013): "Homo Homini Lupus? Explaining Antisocial Punishment," *Journal of Neuroscience, Psychology, and Economics*, 6, 167–188.

Thöni, C. (2014): "Inequality aversion and antisocial punishment," *Theory and Decision*, 76, 529–545.

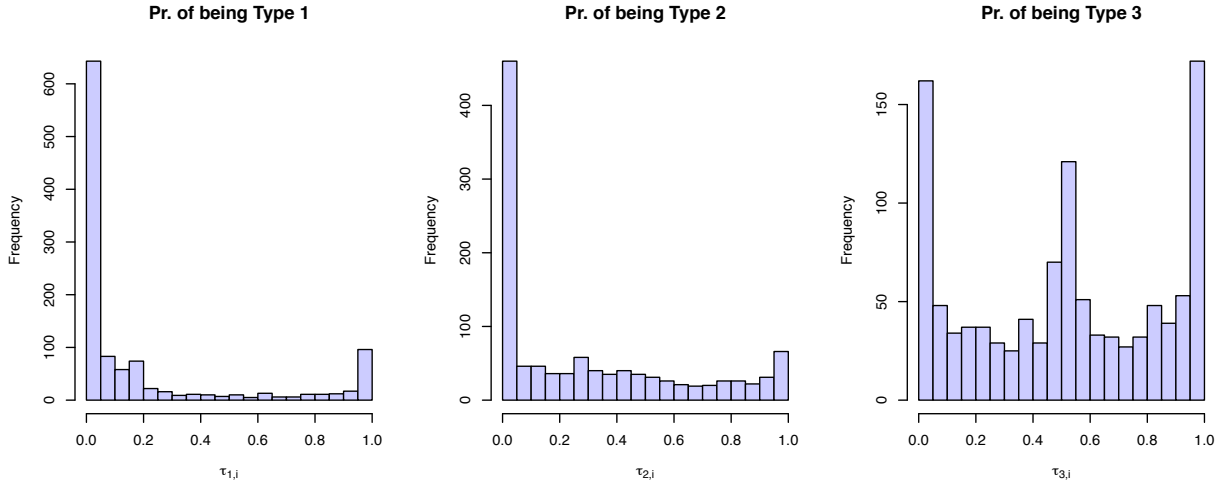
Thöni, C. and S. Volk (2018): "Conditional cooperation: Review and refinement," *Economics Letters*, 171, 37–40.

Yamagishi, T. (1986): "The Provision of a Sanctioning System as a Public Good," *Journal of Personality and Social Psychology*, 51, 110–116.

A Appendix

A.1 Model with $K = 3$ Types

Figure A1: Ambiguous Classification of Subjects into $K = 3$ Types

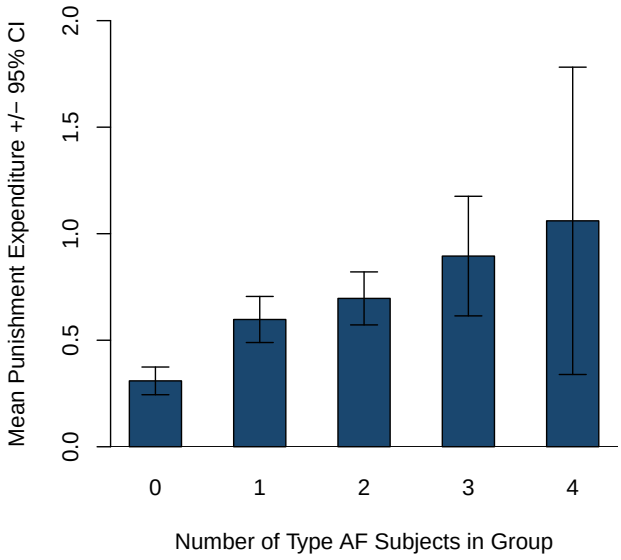


The figure shows the distribution of the individual ex-post probabilities of type-membership τ_{ik} (see equation (2)) for the finite mixture model with $K = 3$ types. Many subjects' τ_{ik} are in the middle of the probability interval, indicating an ambiguous classification. Thus, the the three types exhibit significant overlap which indicates that the model may overfit the data.

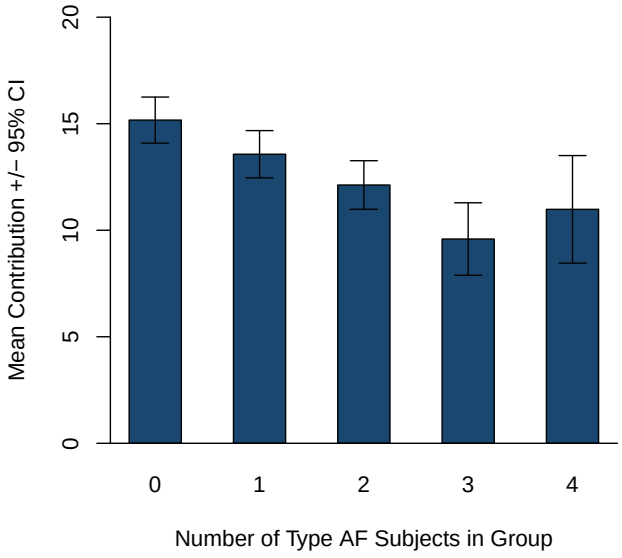
A.2 How the Number of Type AF Subjects in the Group relates to Punishment Expenditures and Contributions

Figure A2: Number of Type AF Subjects in the Group and

(a) Mean Expenditures on Punishment



(b) Mean Contributions



A.3 Model for Testing the Stability of Types

Table A1: Estimates of the Finite Mixture Model for Testing the Stability of Types

Dependent variable[a]: i's expenditure to punish j in period t	Type AF	Type F
Type-specific estimates		
Share among the population (π_k)[b]	0.354	0.646
	(0.027)	(0.027)
inclination to antisocial punishment (β_{1k})	0.136***	-0.299***
	(0.023)	(0.025)
inclination to free rider punishment $-(\beta_{1k} + \beta_2)$[c]	0.165***	0.601***
	(0.023)	(0.018)
Common estimates		
i's contribution (γ_1)	-0.196***	
	(0.012)	
Mean contribution of other group members (γ_2)	0.057***	
	(0.012)	
Received punishment in $t-1$ (γ_3)	0.219***	
	(0.018)	
Period (γ_4)	-0.100***	
	(0.024)	
Final period (γ_5)	0.572***	
	(0.170)	
Constant (β_0)	-1.960***	
	(0.217)	
Standard deviation of error term (σ)	3.594	
	(0.111)***	
Number of subjects	1,120	
Number of observations	33,600	
Log likelihood	-28,100.82	
AIC	56,223.63	
BIC	56,6316.28	

Individual-specific cluster robust standard errors in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

[a] Since the dependent variable is non-negative, we use a Tobit specification.

[b] There are no significance stars as relative sizes lie within the interval $[0, 1]$.

[c] Estimates and standard errors are based on the Delta-method.

A.4 Type-Membership and Observable Characteristics

Table A2: Linear Probability Model

Dependent variable: Probability of belonging to Type AF ($\tau_{i,AF}$)	(1)	(2)
Female	0.044*	
	(0.024)	
Female × English speaking		-0.002
		(0.054)
Female × Protestant European		0.107**
		(0.047)
Female × Orthodox/Ex-Communist		0.024
		(0.057)
Female × Southern European		0.179**
		(0.073)
Female × Arabic speaking		0.091
		(0.120)
Female × Confucian		-0.041
		(0.049)
Age/100	-0.354	-0.356
	(1.362)	(1.367)
$(\text{Age}/100)^2$	0.399	0.438
	(2.230)	(2.219)
No. known subjects	0.000	0.001
	(0.004)	(0.004)
Single Child	0.047	0.045
	(0.037)	(0.036)
Urban background	-0.060**	-0.056**
	(0.026)	(0.026)
Middle class	0.012	0.015
	(0.027)	(0.027)
Member in civic organization	-0.029	-0.026
	(0.032)	(0.032)
English speaking	-0.019	0.025
	(0.035)	(0.044)
Orthodox/Ex-Communist	0.174***	0.206***
	(0.039)	(0.046)
Southern European	0.047	0.024
	(0.042)	(0.049)
Arabic speaking	0.340***	0.357***
	(0.052)	(0.059)
Confucian	0.001	0.065
	(0.034)	(0.046)
Constant	0.361*	0.329
	(0.206)	(0.207)
R^2	0.098	0.106
Number of observations / subjects[a]	985	985

Heteroskedasticity robust standard errors in parentheses (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$).

Protestant European is the basis category.

[a] 135 observations dropped as individual characteristics are not available in all subject pools.