Chapter In

Ensminger, Jean and Joseph Henrich
2014 *Experimenting with Social Norms: Fairness and Punishment in Cross-Cultural Perspective*. Jean Ensminger and Joseph Henrich, eds. New York: Russell Sage Foundation.

Chapter 4

Major Empirical Results: Markets, Religion, Community Size, and the Evolution of Fairness and Punishment

Joseph Henrich, Jean Ensminger, Abigail Barr, and Richard McElreath

Building on the theoretical framework laid out in chapter 2 and the background and methods described in chapter 3, this chapter presents our major empirical findings looking across our populations. In chapter 2, we proposed that a particular set of social norms has coevolved with the emergence of markets and the institutions of complex societies in order to facilitate exchange among individuals not involved in durable long-term relationships such as those associated with kinship, reciprocity, and status. Our experiments, with their salient contextual cues of cash and anonymity, are well suited to tap these particular norms. Thus, our framework predicts not only that these context-specific equity norms will vary in strength, but also that the strongest equity norms (measured by offers) will tend to be found in the most complex, market-integrated societies. Further, since theoretical work suggests that such norms can be maintained in larger groups only through costly diffuse punishment, while smaller groups can rely on punishment or other reputation-based mechanisms to sustain such norms, we also expect larger groups to engage in more costly punishment than smaller groups. Our major empirical findings can be summarized in four points:

1. *Fairness and punishment show both reliable patterns and substantial variability across diverse populations.* On the fairness side, as measured by offers, mean offers across our populations ranged from about 20 percent to roughly 50 percent of the stake, spanning one-quarter of the spectrum of possible offers. On the punishment side, the probability of second- or third-party punishment in the ultimatum game (UG) and the third-party punishment game (TPG) always declined as offers increased from zero to half of the stake, in all populations. However, the willingness of individuals to punish varied across populations, with the fraction of each population willing to engage in costly punishment of the lowest possible offer varying from 3 to 100 percent in the UG and from 26 to 100 percent in the TPG.

2. *Fairness increases with a population's degree of market integration.* Offers in all three experiments increased with market integration (measured as the percentage of calories purchased in the market), even after controlling for a range of economic and demographic variables. The effect of market integration replicates our prior finding involving market integration and UG offers—with eleven new populations added—and extends these findings to two other bargaining games (Henrich, . . . and Gintis 2004; Henrich et al. 2005a).[1]

3. *Fairness increases with an individual's participation in a world religion.* Compared to those who practice local or traditional religions, participants in Islam or Christianity made higher offers in the dictator game (DG) and the ultimatum game, though the effect is crowded out or otherwise reduced in the third-party punishment game. Overall, as we move from an entirely subsistence-based society with a traditional religion to a fully market-integrated society with a world religion, our measures of market integration and world religion predict an increase in offers of between roughly 20 and 23 percent of the stake in the DG and UG (using OLS regressions). This spans most of the range of variation we observe across mean offers in different populations. In the TPG, the predicted increase is 11 percent.
4. *Willingness to engage in punishment increases with community size.* Our measures of costly punishment in both the UG and TPG show that greater willingness to punish is associated with larger communities, controlling for sociodemographic and economic variables, including market integration and world religion. The estimated effect is dramatic: in the smallest communities (fifty people), the most common preference is not to engage in any costly punishment, while in the largest communities (nearly five thousand people), the most common preference is to punish even small deviations from an equal split (offers of 40 percent).

In this chapter, we lay out the analyses supporting these findings and provide an overall discussion that compares our theoretical interpretations with various alternatives; we then consider some standard criticisms. Our first section discusses the universal patterns and variation observed across our populations for each experiment and analyzes how they diverge from behavioral predictions rooted in pure self-interest. Next, we examine the relationships of market integration and world religions to our offer measures for all three experiments, using a baseline set of seven other economic and demographic predictor variables. We then examine the relationship between community size and our two measures of costly punishment. In each section, we discuss each experimental game in turn and then summarize our findings.

Throughout this chapter, we use terms like "fairness" and "punishment" as a shorthand to describe the motivations and behaviors of our participants. We do not mean to imply, however, that these are context-general or dispositional traits or characteristics of individuals or populations. It remains an empirical question as to how broadly these behavioral patterns apply, though from our theoretical perspective, they may apply only to contexts involving monetary transactions and lacking long-term, relationship-specific demands (for example, status, kinship) or reciprocity motivations. Our experiments probably do not, for example, generally cue and measure the social norms associated with complex kinship relationships, food-sharing, or cooperative fishing. In some societies, however, many interactions may fall into a kind of default category that is most applicable when other norms or motivations do not apply.

UNIVERSAL PATTERNS AND VARIATION IN PROSOCIAL BEHAVIOR ACROSS POPULATIONS

To examine both the universal patterns and the variation observed across our samples, we first present results from the DG, then the UG, and finally the TPG. In presenting these results, we emphasize two patterns that are robust across our samples. First, in all three experiments mean and modal offers for our populations span only a limited range, from 0 percent to 51 percent, with few offers above 50 percent. We did not, for example, find societies in which most people give more than half, or in which most people give zero. This result replicates the findings of the first round of experiments (Henrich, . . . and Gintis 2004; Henrich et al. 2005a, 2005b) using only the UG.[2]

Second, there is a decline in the likelihood of punishment in the UG and TPG as offers increase from 0 to 50 percent, with 50 percent always showing the lowest likelihood of punishment (which was sometimes tied with other offer amounts as well). Although this declining pattern may seem intuitive for many, our experimental approach could have found a vast set of alternative patterns. We might have found, for example, that the likelihood of punishment in some populations increases as offers approach 50 percent. Or we might have found that the likelihood of punishment increases up to 20 percent, then decreases to 50 percent. Thus, what we actually observed represents only a tiny subset of what was possible. At the other end of the offer spectrum in the UG, we found only two patterns of punishment as offers increase from 50 percent to 100 percent: either there is no punishment of these high offers (or only a sparse scattering) or the probability of punishment *increases* as offers approach 100 percent. Such an increase occurred in six populations out of fourteen. These broad patterns are important because they dramatically reduce the state space of possible explanatory theories.

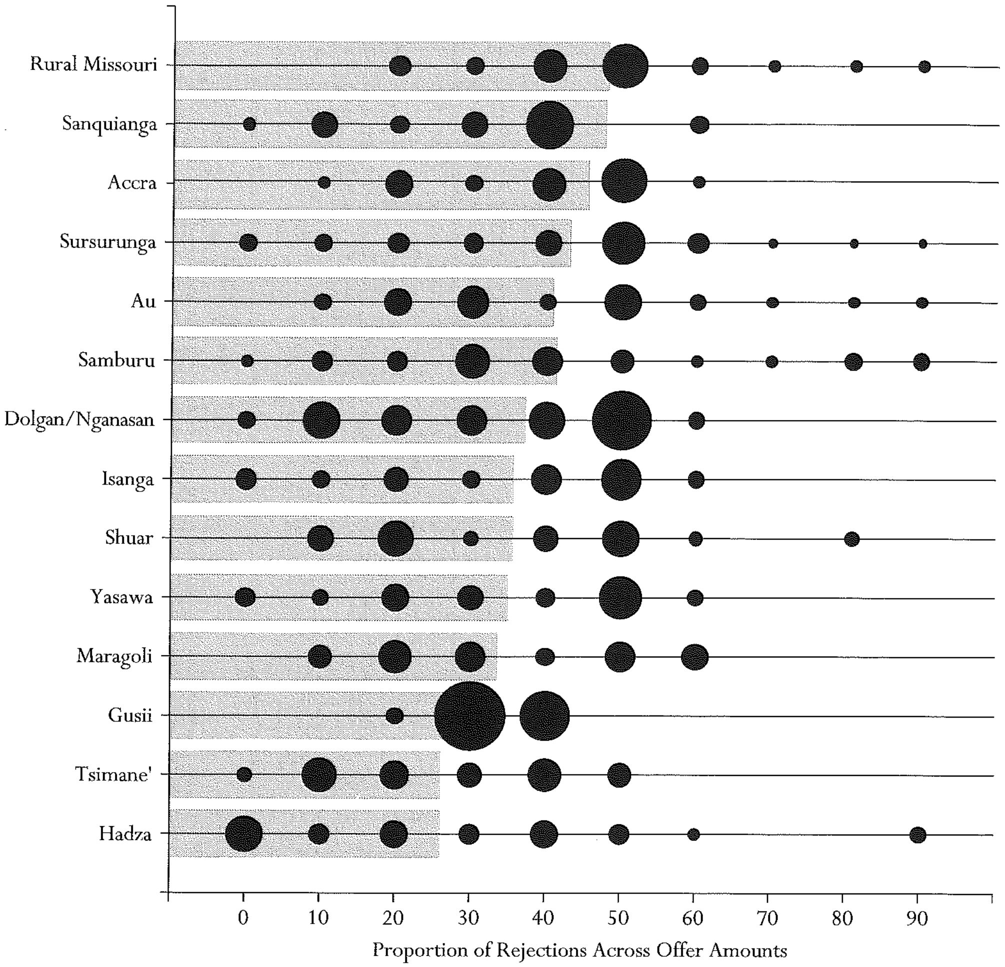
Within these robust patterns, there is substantial variation across populations. All five of our experimental measures of fairness and punishment show more variation among populations than is typically observed among subjects from industrialized societies. Mean offers vary from 26 percent to 47 percent in the DG, from 25 percent to 51 percent in the UG, and from 20 percent to 43 percent in the TPG. The fraction of each population willing to engage in costly punishment for each population ranges from 3 percent to 100 percent in the UG and from 26 percent to 100 percent in the TPG. At the other end of the punishment spectrum, the willingness to punish UG offers of 100 percent of the stake ranges from 0 percent in many populations to nearly 50 percent of the samples from two populations. We argue that much of this variation reflects the presence and strength of an equity norm (fifty-fifty division) that is applicable to contexts involving ephemeral interactions and money, combined with the locally appropriate sanctioning mechanisms used to enforce such norms (that is, costly punishment or reputation).

Finally, our findings—taken together and in light of our previous project—show that predictions assuming purely self-regarding preferences fail in all populations studied in all three experiments: either players engage in costly punishment, despite the one-shot nature of the experiments, or they offer too much, given the likelihood of punishment. In several places many people engage in both violations—punishing at a personal cost and offering too much. It seems that the assumption of pure self-interest fails in different ways, and to varying degrees, in different societies.

The Dictator Game

Figure 4.1 shows the distributions of offers in the dictator game. The *x*-axis gives the possible offers as a percentage of the total stake, with the size of the bubbles at each offer displaying the proportion of that sample that made that offer. Overall, of our 427 DG offers, 5.2 percent (22) are zero, 37.7 percent (131) are fifty-fifty splits, 85.5 percent occur between 10 percent and 50 percent (inclusive), and 9.4 percent are greater than half the stake (40 offers, 21 of which were at 60 percent). Our populations differ in modes, means, and standard deviations. Mean offers range from about 26 percent among the Tsimane' and Hadza to about 47 percent in the United States and Sanquianga. Modal offers are zero among the Hadza, 10 percent for the Tsimane', 20 percent for the Maragoli, 30 percent for the Gusii, and 50 percent in the rest of the populations, except for the Shuar, who show modes at both 20 percent and 50 percent. The standard deviations in offers vary across societies, from 5.4 among the Gusii farmers in the highlands of Kenya to 25.0 among Hadza foragers. Although these data do indicate substantial variability across populations, we emphasize that it is not the case that "anything goes," as few offers

FIGURE 4.1 *The Dictator Game: Distribution of Offers*



Source: Project data.
Notes: Reading horizontally for each of the fifteen populations listed along the left vertical axis, the area of each bubble represents the fraction of our sample that made that offer. Each horizontal set of bubbles thus provides the distribution of offers for each population. The gray bar reaches to the mean offer for each population and is the measure by which the table is sorted. Three offers of 100 percent, two from the Dolgan/Nganasan and one from Accra, are not shown.

above half the stake were observed, and all the population-level variation (mean and modes) is confined to only a fraction of the space of potential variation.

The prediction of models based on pure self-interest in the DG—offers of zero (Camerer and Fehr 2004)—is not well supported overall. Only about 5 percent of all offers were zero, and 41 percent of those occurred among the Hadza. The modal offer among the Hadza is zero, although 71 percent of Hadza offered more than zero. We note that the assumption of fully self-regarding preferences is more strongly supported than the assumption of fully other-regarding preferences, as only three individuals out of our 427 offered 100 percent of the stake. Fortunately, current models do not require us to pick between these two extremes (Camerer and Fehr 2006), but instead allow us to theorize and measure the mix of motivations in decisionmaking.

Those familiar with DG results from university student samples will observe that our nonstudent adults, especially our Americans, are substantially more prosocial than typical students. Several lines of converging evidence indicate that this difference probably arises from the fact that students' prosocial preferences are still developing and have not yet hit their adult plateau. First, several dictator games done by different researchers in nonstudent (older) adult pools in the United States and Switzerland show the same pattern we found (Carpenter, Burks, and Verhoogen 2005; Henrich and Henrich 2007)—mostly fifty-fifty offers—in contrast to the typical student findings. Second, we performed our DG protocol among both American students (see chapter 9) and nonstudent adults in rural Missouri (see chapter 18). Our Missouri findings match other nonstudent U.S. adult samples, and our student results match typical student findings—that is, it is not our specific protocol that is causing the prosociality. Third, as noted in chapter 2, research explicitly examining behavioral experiments across the life span show that other-regarding preferences in Westerners develop slowly (Harbaugh and Krause 2000; Harbaugh, Krause, and Liday 2002), change over the university years (Carter and Irons 1991), and do not plateau until at least the midtwenties (Henrich 2008; Sutter and Kocher 2007).

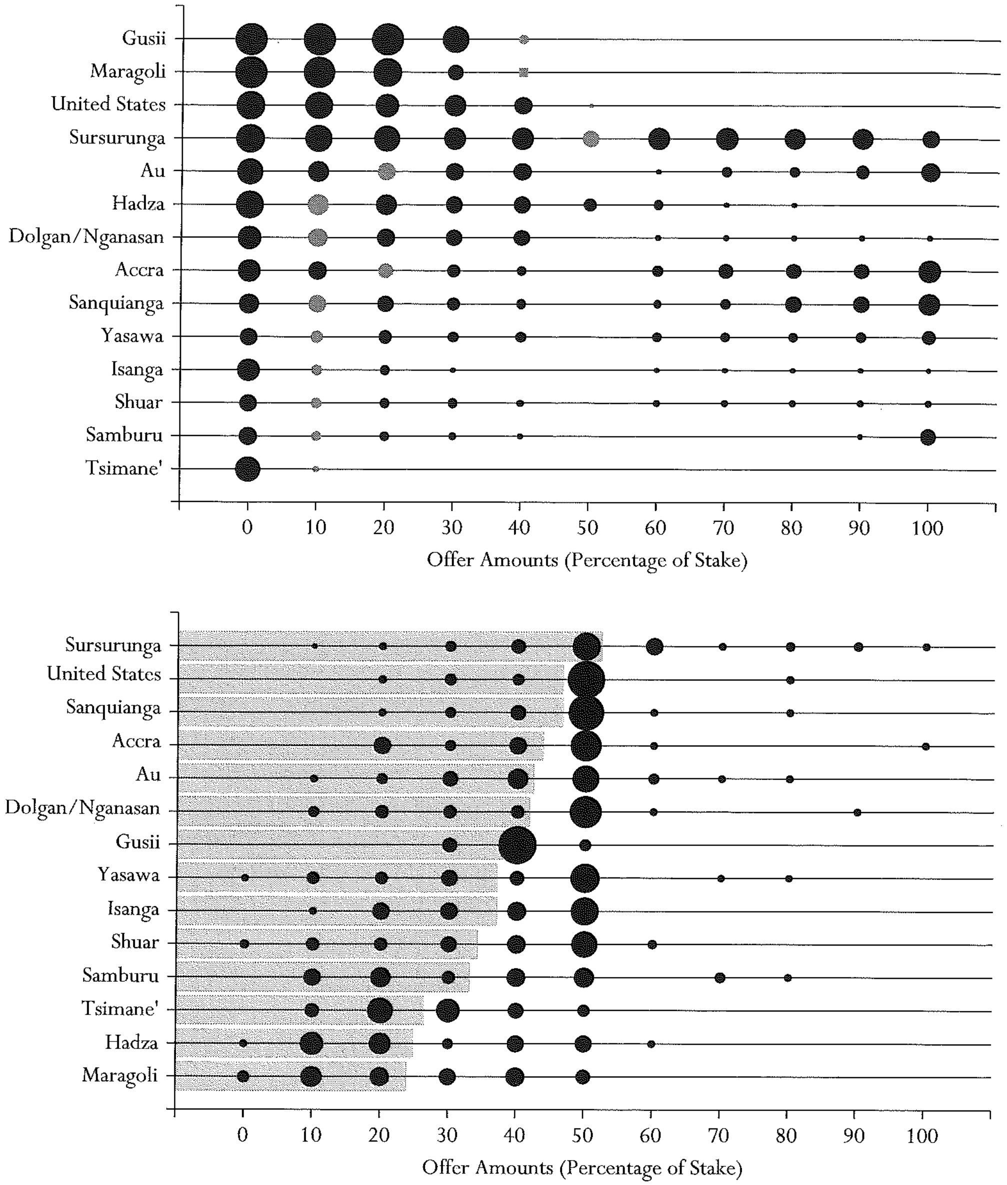
The Ultimatum Game

We begin with punishment in the ultimatum game as measured by people's willingness to reject offers. The top panel of figure 4.2 displays the distribution of rejections (punishments) across potential offers, and the bottom panel shows the distribution of UG offers. Reading the top panel horizontally for each population, the size of the bubble at each possible offer (listed along the *x*-axis) represents the proportion of that sampled population who indicated that they would reject that offer amount. The figure shows that the likelihood of rejection is highest for offers of zero and declines as offers approach 50 percent. For offers above 50 percent, there are two patterns. Either there are no, or only a few, scattered rejections or the probability of rejection increases as offers increase from 60 percent to 100 percent of the stake. We term this willingness to reject offers greater than 50 percent "hyper-fair punishment" and discuss it at greater length later in the chapter.

The variation among populations in their willingness to reject lower offers can be captured in at least three different ways. First, consider the variation in the proportion of each population who were willing to reject offers of 10 percent, an offer representing the cheapest punishment that still costs the responder some money. (Rejecting an offer of zero does not cost the responder anything.) Overall, 57 percent of players rejected 10 percent offers. However, in four populations—the Tsimane', Shuar, Isanga Village, and Samburu—fewer than 10 percent of people rejected offers of 10 percent. In contrast, 65 percent or more of the samples from four populations rejected such offers. In the top panel of figure 4.2, the populations are listed from top to bottom in order of decreasing willingness to reject offers of 10 percent.

A second way to compare punishment across populations is to use the pattern of rejection responses to calculate the income-maximizing offer (IMO) for each population. This is the offer that a proposer would make if he or she wanted only to maximize his or her own income and knew the probability of rejection for each possible offer (and could do all the necessary mental calculations). The IMO for each population is marked in gray in the top panel. IMOs range from 50 percent in the United States and among the Sursurunga to 10 percent among five other populations. That is, in five societies the threat of punishment is insufficient to drive an income-maximizing proposer above the theoretical prediction for offers (10 percent), based on pure self-interest.

FIGURE 4.2 *The Ultimatum Game: Distribution of Rejections (top) and Offers (bottom)*



Source: Project data.

Notes: The top panel shows the frequency of rejections across offers in the UG for each population. The bubbles' areas represent the portion of the sampled population (listed along the *y*-axis) who rejected offers at the amounts marked along the *x*-axis. The largest bubbles indicate that 100 percent of the sample rejected. Gray bubbles mark the income-maximizing offer (IMO). The square marks the IMO for the Maragoli, who made no rejections at that offer amount. The populations are ordered from bottom to top according to the frequency of rejections for offers of 10 percent of the stakes. No rejection data were collected for offers above 50 percent in the U.S. sample (see Chapter 3). The bottom panel shows the histogram of offers and mean offers for each population. The bubbles' areas represent the relative frequencies of offers at each of the amounts listed along the *y*-axis. The horizontal gray bars reach to the mean offer for each population. The populations are ordered by mean offer amount.

Our third approach compares the accept/reject decisions for each individual across offers to assign a minimum acceptable offer (MinAO) to each player. An individual's MinAO is the lowest offer that he or she is willing to accept. If a player accepts all offers between 0 percent and 50 percent, his or her MinAO is zero. If the player rejects 0, 10, and 20 percent but accepts 30, 40, and 50 percent, his or her MinAO is 30.[3] While the overall mean MinAO is 16, mean MinAOs for each population range from 6.1 among the cattle-herding Samburu to 27.9 in the United States.

For offers above 50 percent, we also observe variation across populations in people's willingness to engage in punishment. Of our fourteen populations, six displayed an increasing willingness to reject increasingly inequitable UG offers as they rose from 50 percent to 100 percent. Of those who showed this tendency, the fraction of the sample willing to reject offers of 100 percent ranged from 20 percent among Fijians to 50 percent in Accra, Ghana. We explore this phenomenon in more detail later (see also Henrich et al. 2006).

With regard to the universal patterns and variation in UG offers, the bottom panel of figure 4.2 provides histograms (in the form of bubbles) for the distribution of offers in each population. The gray horizontal bars show that mean offers range from around 25 percent of the stake among the Maragoli, Hadza, and Tsimane' to 51 percent among the Sursurunga.[4] Modal offers range from 10 percent among the Maragoli and Hadza to 50 percent in several societies. Eighty-four percent of all offers occur between 20 percent and 50 percent of the stake (inclusive), while only 6.8 percent of all offers are over 50 percent, and most of these came from New Guinea (Sursurunga and Au). Of these hyper-fair offers, half are offers of 60 percent of the stake.

Are Hyper-Fair Rejections Just Confusion?

Many researchers find the existence of hyper-fair rejections non-intuitive. To address this we explored the possibility that, despite our one-on-one testing procedures, those who rejected high offers might have somehow misunderstood the game. For every player 2 in the UG we counted the number of rejections for offers greater than 50 percent and ran two regressions. First, we used a negative binomial regression with robust standard errors to regress this count variable on education (measured in years of formal schooling). If those who rejected hyper-fair offers did so because of some misunderstanding regarding the game, we might expect more-educated people to have a better understanding and thus have fewer hyper-fair rejections. The coefficient, standard error, and *p*-value for education are −0.020, 0.029, and 0.49, respectively. Adding population dummies to remove any between-group variation yields a coefficient, standard error, and *p*-value of −0.0042, 0.045, and 0.92, respectively. Similar results obtain if one uses standardized value for education, as we do later in the chapter, to deal with regional differences in educational quality. Finally, if we include only those six populations showing an increasing tendency to reject as offers approach 100 percent, we obtain similar results to those found in the first regression. In short, we find no evidence that more-educated individuals are less likely to make hyper-fair rejections.

Our second test of the confusion hypothesis was to regress our hyper-fair rejections variable on the number of examples and test questions used, which provides a potential proxy for how much effort was required in explaining the game, as it was conveyed through repeated examples and test questions. Using a negative binomial regression with robust standard errors, the coefficient, standard error, and *p*-value for this predictor are −0.16, 0.072, and 0.03, respectively. Here the coefficient is negative, indicating that those who required more examples to learn the game (that is, had a tougher time understanding it) made fewer hyper-fair rejections. This is opposite to the prediction of the "confusion explanation."

Third, postgame interviews of players who punished high offers in the UG reveal that people understood the game and made sensible responses as to why they rejected high offers, such as, "It was too much, I cannot accept that much."

Finally, alongside these findings are a few other empirical patterns that contradict the notion that hyper-fair offers are a product of confusion or misunderstanding. To begin, in the TPG, which was generally more difficult to explain and took longer for players to comprehend (more examples), people did not punish hyper-fair offers. A look at the TPG explains why. If player 1 offers the full amount (100 percent) to player 2, player 3 cannot punish player 1 because we did not allow negative payoffs (and player 3 is not permitted to take money away from player 2). Player 3 could pay 10 percent of his or her stake, but this would not take any money away from player 1. If player 1 gives 90 percent to player 2, player 3 could pay 10 percent to take 10 percent away from player 1, but this is very inefficient punishment. It is not until player 1 gives 70 percent to player 2—that is, when matters are much less inequitable—that player 3 can administer the full brunt of his or her punishment to player 1. Consequently, punishment was not expected for high offers in TPG, and very little was seen. However, if the punishment of hyper-fair offers in the UG was the result of confusion, it is not obvious why similar confusions did not manifest themselves in hyper-fair rejections in the TPG. In fact, since the TPG was more difficult to understand than the UG, we would have expected more punishment of hyper-fair offers, if confusion was the reason.

The patterns of hyper-fair rejections observed in our data are consistent with those from UG experiments done in other non-Western societies by other researchers, as well as with nonstudent adults in the West and with Western undergraduates when more sensitive experimental tools are used. In Tatarstan and Sakha-Yakutia (Russia), Donna Bahry and Rick Wilson (2006) used our protocol and found the same patterns of hyper-fair rejections. Their analyses parallel ours in showing that confusion is unlikely to explain the presence of the phenomenon. Similarly, in China, Heike Hennig-Schmidt, Zhu-Yu Li, and Chaoliang Yang (2008) found hyper-fair rejections in UGs. Research among representative adult samples (nonstudents) in three countries in Europe using the UG has also revealed this tendency for hyper-fair rejections among nonstudent adults, though it is substantially weaker than in many of the non-Western populations discussed here (Bellemare, Kröger, and van Soest 2008; Güth, Schmidt, and Sutter 2003; Wallace et al. 2007).[5] Among Western undergraduates, milder versions of this phenomenon have been detected using bargaining instruments that permit the expression of weaker preferences for hyper-fair punishment in the responder (Andreoni, Castillo, and Petrie 2003; Huck 1999).

When applying the standard UG (not involving the strategy method) in phase 1 of the project, we observed hyper-fair rejections only among the Au and the Gnau of Papua New Guinea. We were able to observe this previously only among the Au and Gnau because these two groups, unlike the other populations studied in phase 1, showed some substantial willingness to make actual hyper-fair offers; hyper-fair rejections can only be observed in the standard form of the UG (see chapter 3) if actual hyper-fair offers are made. Our findings with the Au reported herein (chapter 7) replicate our previous efforts (Tracer 2003, 2004), while the Sursurunga findings, another New Guinea population that was added precisely to give more insight into this phenomenon, extend our observations and suggest some degree of regional generality (chapter 11). Like the Au and Gnau, the Sursurunga both make and reject hyper-fair offers. In fact, they make hyper-fair offers with sufficient frequency that they are the only population in phase 2 with a mean offer greater than 50 percent.

Self-interest and Risk Aversion in the Ultimatum Game

Despite the behavioral variation just described, we found that none of our populations conformed to the predictions of the oft-discussed model based on purely self-regarding preferences (Camerer 2003). This approach predicts that responders will accept any positive offer and thus proposers will make the smallest positive offer. Across our populations, responders either rejected positive offers or proposers offered too much, given the probability of rejection across offers. Four populations both rejected positive offers and gave too generously. Focusing on the rejection of offers of 10 percent (the cheapest opportunity for costly punishment), we calculated exact 95 percent confidence intervals (CIs) for each population and found that all populations except the Tsimane' could be distinguished from a zero probability of rejection at this offer amount. This remains true even if we calculate exact 99 percent CIs. Thus, strictly on the basis of responder behavior, our data indicate that all of the societies studied, except the Tsimane' responders, violate the narrow economic self-interest assumption in the UG. This is important because unlike the proposer, for whom we assume the possession of accurate beliefs about the likelihood of rejection across offers, the responder's decision to forgo free money is not contingent on anticipating another's behavior. There are many potential reasons why the self-regarding model might be failing here, including that the responders have inaccurate beliefs about the anonymity in the games, or that many people are motivated to punish low offers in this context.

On the proposer side, of our fourteen societies, four had either mean or modal offers near their income-maximizing offer. Of the remaining ten populations, nine had modal and mean offers above their IMO. For these nine, it may be that risk aversion—a standard modification of the self-regarding model—explains why mean and modal offers are higher than the IMO. For example, suppose a subject estimates that an offer of 40 percent of the pie will be accepted for sure (leaving 60 percent for the proposer) and that there is a two-thirds chance that an offer of 10 percent will be accepted. If this subject is risk-averse, he or she might value the certainty of keeping 60 percent of the pie more than the two-thirds chance of keeping 90 percent (and a one-third chance of getting nothing). In this case, the expected monetary gain is the same for the two offers (namely, 60 percent of the pie), but the expected utility of the certain outcome is greater. Thus, a risk-averse subject might make a high offer even if the probability of rejection of a low offer is small.

To examine this we assume that the utility (U) that individuals derive from money (I) is concave, such that $U = I^r$, where r provides a standard measure of risk aversion. If $r = 1$, people are risk-neutral; if $r < 1$, people are risk-averse; and if $r > 1$, people are risk-seeking. For each of the populations, we recalculated a utility-maximizing offer (UMO) by calculating the value of r closest to 1 that minimized the difference between the utility-maximizing offer and the (a) mean and (b) modal offers for each population. We assumed that proposers knew the actual (empirically observed) probabilities of rejection across offers for their group.

Using this approach, we found that five of our remaining ten populations required implausibly low values of r, two were somewhere in the middle, two obtained plausible values of r, and one required an r value that was implausibly high. For the first five populations, including the Tsimane', we found that r (for both mean and mode) was less than 0.3, an implausible amount of risk aversion. The five populations with implausibly low values of r are Isanga Village ($r = 0.22$), the Samburu ($r = 0.18$), the Shuar ($r = 0.18$), the Tsimane' ($r = 0.26$), and the Yasawans ($r = 0.27$).[6] To see how implausible such values are, a person with $r = 0.27$ would prefer a certainty of $1 over an even chance at $12 (yielding $6 on average). If this person faced five of these choices each week, he or she would earn $5 a week compared to the mean of $30 earned by someone with $r = 0.4$. Putting these five populations aside, two others require fairly low values of *r*, although these do not seem completely implausible (for the Accra $r = 0.47$, and for the Sanquianga $r = 0.54$). The Au and Dolgan/Nganasan require plausible values of *r*, 0.76 and 0.72, respectively. With regard to the self-regarding model and the potency of risk aversion in explaining our findings, this analysis parallels our previous work using the data from phase 1 (McElreath and Camerer 2004).

It is worth noting that work using experiments designed to measure risk preferences in small-scale societies, including African agro-pastoralists and South American subsistence farmers, have not revealed levels of risk aversion anywhere near these values (Henrich and McElreath 2002). Moreover, efforts to establish a link between measures of risk preferences derived from risk experiments and offers above the local IMO have failed (Henrich et al. 2005a).

Lastly, in contrast to all other populations, the Maragoli had an IMO greater than their mean and modal offers in the UG. Following the same logic as earlier, we estimated the lowest value of *r* that would bring the utility-maximizing offer into correspondence with the actual mean and modal offers. The amount of risk-seeking required to accomplish this is extremely implausible. The value of *r* estimated was 13.7, meaning that the Maragoli would pass up $10 for certain in favor of a 50 percent chance at $10.50 (or zero). In chapter 12, Gwako provides an extended discussion of the Maragoli research and the unique situation of this population.

Overall, these findings replicate our team's previous research using the standard UG (Henrich, . . . and Gintis 2004; Henrich et al. 2005a). Despite employing a uniform methodology across sites that differs from the methodology used in phase 1 and using the strategy method (eliciting responses for all possible offers), our new findings still parallel those of phase 1 in four important ways. First, the broad patterns of variation across our societies are the same. The ranges of mean and modal offers are similar. Second, the earlier findings from the Hadza, Tsimane', and Au, which were each aberrant in different ways, have largely been replicated in these new experiments, with the same patterns reemerging. The Hadza again made relatively low offers, but punished sufficiently that their mean and modal offers were close to their IMOs. The Tsimane' again did not reject, and made low offers, as did some Tsimane' villages in Gurven's previous work. The Au were again willing to reject low offers and make high offers (including offers even greater than 50 percent), and they were also willing to reject offers greater than 50 percent with increasing probability. Our second New Guinea population also revealed these same unusual patterns, even more strongly than the Au. Finally, as detailed later, we replicated the relationship between market integration and UG offers. These parallels between the results from phases 1 and 2 suggest that our methodological decision to play the UG after the DG in phase 2 probably had no important impact on the overall pattern of results, except perhaps in the case of the Maragoli (see chapter 12).

The Third-Party Punishment Game

The third-party punishment game reveals patterns similar to those already seen in the DG and UG. With regard to punishment, all our populations showed at least some willingness to punish low offers, with the likelihood of punishment declining as offers increased (see top panel of Figure 4.3). There was substantial variation, however, in individuals' willingness to punish across populations. This can be illustrated, first, by considering the frequency of punishment for the lowest offers, and second, by using minimum acceptable offers (paralleling the treatment for the UG). For offers of zero, two-thirds of all player 3s were willing to pay to punish. Across populations, this fraction varied from around 26 percent among the Tsimane' to over 90 percent among the Samburu and Maragoli and 100 percent for the Gusii; see the top panel of figure 4.3 at zero on the *x*-axis. (Note that we do not have TPG results for U.S. nonstudent adults, but see chapter 9 for U.S. students.) Next, using each player's vector of punish or do-not-punish decisions across offers, we were able to calculate a minimum acceptable offer for 90 percent of our sample. The MinAO in the TPG represents the lowest offer for which a player will not punish. Mean MinAOs range from about 4 percent among the Tsimane' to 41 percent among the Gusii, giving a mean across populations of 21 (the mean of sample means).

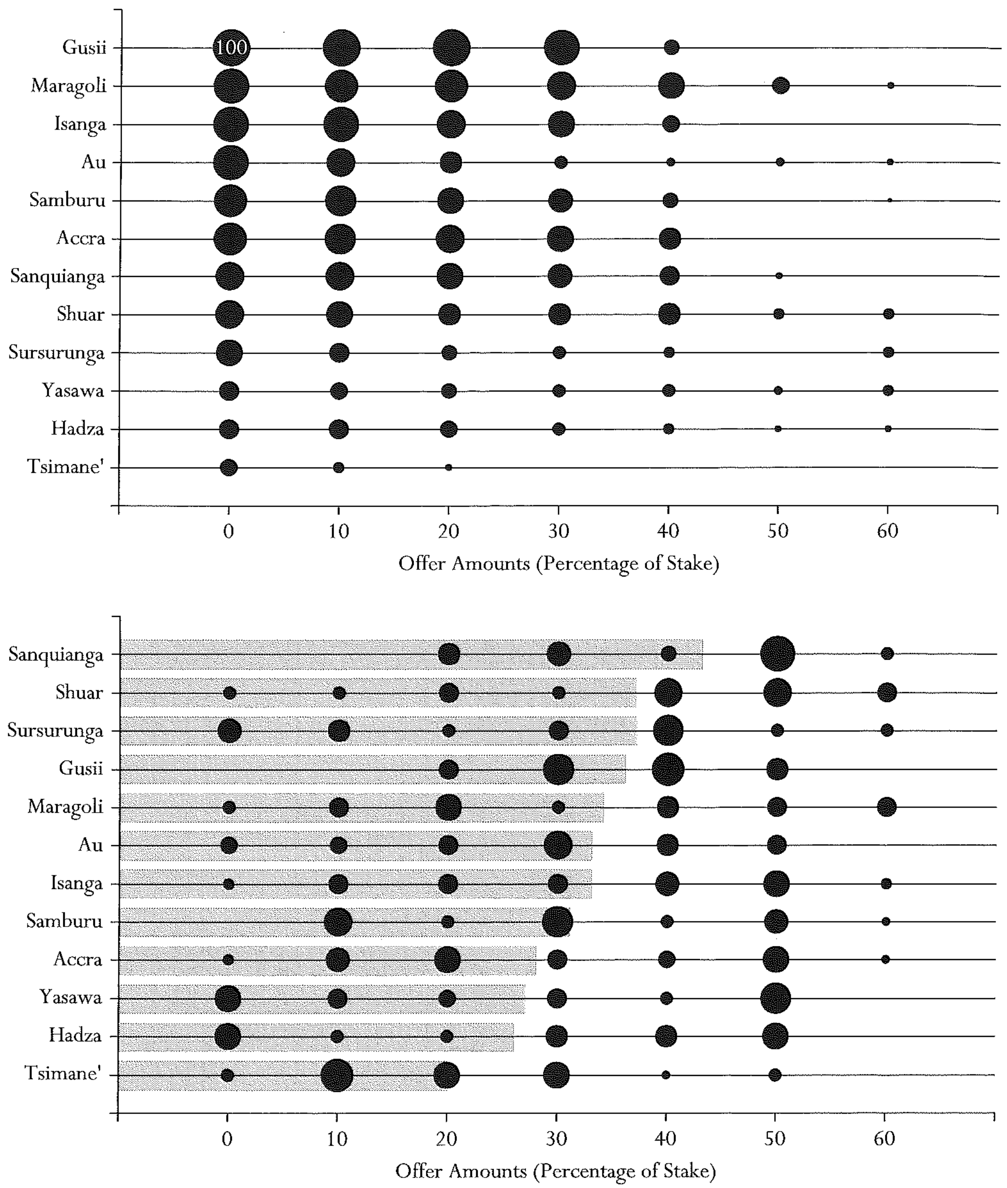
As noted earlier, the design of this experiment precludes us from observing the kind of hyper-fair punishment that was observed in the UG. In our rules for this game, player 3 could not reduce the final payoff of player 1 below zero because—for practical reasons—we could not deliver negative payoffs to our participants. This means that for high offers, punishing was neither an effective means to reduce player 1's payoff—that is, to punish him—nor a way of rectifying inequity among players. Assuming that the motive behind punishing has something to do with either reducing inequity or hurting player 1 for non-prosocial behavior, we would expect few to punish at high offers. Few did. Out of our 338 player 3s, four people punished offers of 100 percent, and the same four punished offers of 90 percent. These four were spread among three populations: Fijians, the Hadza, and the Shuar.

The distribution of offers in the TPG parallels the results described for the UG and DG, although offers were generally a bit lower (see Figure 4.3 bottom). Ninety-four percent of offers were between 0 percent and 50 percent of the stake (inclusive). Twenty-three percent of people offered 50 percent, which was the overall modal offer. Of the twenty-two people who offered more than 50 percent, thirteen made offers of 60 percent. Mean offers in the TPG ranged from 20 percent among the Tsimane' to 43 percent in Sanquianga; modal offers ranged from 10 percent among the Hadza to 50 percent in several other societies.

As with the UG and DG, the predictions derived from an assumption of narrow economic self-interest do not fare well in the TPG. A standard prediction assuming pure self-interest is that player 3 will not reduce his or her income in this one-shot interaction by paying to punish. Player 1, recognizing the situation of player 3, should offer zero to player 2. With regard to punishment, 95 percent exact confidence intervals show that the probability of punishment in every population studied can be distinguished from zero. The populations with the lowest likelihoods of punishment for an offer of zero are the Tsimane', with 26 percent of players punishing (at 95 percent CI, 10 to 48 percent), and the Hadza, with 27 percent punishing (at 95 percent CI, 12 to 48 percent). Since the IMOs for all populations were zero—reflecting the relatively limited abilities of punishers to inflict penalties on low offers[7]—all the offers were too high from the point of view of the typical self-regarding predictions. The lowest mean offer was 20 percent, and the lowest modal offer was 10 percent.

As in the UG, we again consider how a standard modification of the self-regarding model—the consideration of differing risk preferences as captured by the concavity of utility with increasing income—might allow us to better predict offers. As before, we assumed a standard relationship between utility and income, $U = I^r$, and estimated *r* based on the observed probabilities of punishment and the mean and modal offers for each population.

FIGURE 4.3 *The Third-Party Punishment Game: Distribution of Punishments (top) and Offers (bottom)*



Source: Project data.

Notes: The top panel displays the distributions of decisions to punish across the possible TPG offers. For each population labeled along the *y*-axis, the areas of the bubbles display the fraction of the sampled population who were willing to punish at that offer amount (along the *x*-axis). Inside the zero offers for the Gusii we placed a "100" to indicate the size of a bubble if everyone punished. The populations are ordered according to their willingness to punish offers of zero. The bottom panel provides the histogram for offers made in each population. The area of each bubble represents the fraction of the sampled population who made the offer. Both plots stop at offers of 60 percent along the *x*-axis because very little punishment and few offers occurred above this amount. Populations are ordered according to their mean offers.

It turns out that there are no positive values of *r* that will budge the utility-maximizing offer from zero. Adding risk aversion, at least in this specification, adds no explanatory power.

In table 4.1, we summarize the mean statistics on offers and rejections by game and society.

EXPLAINING THE VARIATION: MARKET INTEGRATION, RELIGION, AND COMMUNITY SIZE

This section analyzes the variation in our five experimental measures to test the three hypotheses presented in chapter 2, which were derived from considering the coevolution of social norms, institutions, and intrinsic motivations in an economic-evolutionary framework. Our framework proposes that global environmental shifts to a more stable climate regime at the beginning of the Holocene period (twelve thousand years ago) created possibilities for the emergence of larger-scale sedentary human societies, conditions that at most had existed only ephemerally during the preceding hundred thousand years. We argue that a crucial impediment to the emergence of these more complex and evolutionarily novel, larger-scale societies was the development of the social norms and institutions (informal and formal) that were capable of domesticating our evolved social psychology and had adapted over tens of millennia for life in small groups (families, bands, and tribes). We hypothesize that increasingly complex societies prospered to the degree that norms and institutions sustained more intense social interactions in larger and larger social and economic spheres, well beyond the familiarity of the local sphere of durable relationships. It is these norms and institutions—and their gradual internalization over the life course as intrinsic motivations—that recalibrate and harness our evolved social psychology, thereby allowing individuals to successfully interact in larger-scale contexts and outside tightly knit social networks. By incrementally facilitating trust, fairness, and cooperation in an increasingly diverse array of interactions beyond the local group, these emerging norms permitted social groups to make more productive use of diverse skills, knowledge, and resources, as well as gradually increasing cooperation in exchange, defense, public works, and internal policing (such as reducing crime).

One important element in the evolution of societal complexity is the expansion of both the breadth and intensity of market exchange. At its most efficient, market exchange requires trust, fairness, and cooperation among individuals engaged in infrequent, ephemeral, or anonymous interactions. The greater the degree to which expectation-motivation sets related to trust, fairness, and cooperation are shared, the lower the transaction costs and the higher the expected rewards. However, studies of both nonhuman primates and small-scale societies suggest that during most of human history transactions beyond the local group, and certainly those beyond the ethnolinguistic unit, were often fraught with danger, mistrust, and exploitation.[8] Reliable transactions among strangers are commonplace for many people today, but they probably have not always been part of human evolutionary history. Thus, in refining our theoretical proposal, we suggest that market norms may have evolved as part of this overall process to facilitate and extend mutually beneficial interactions in contexts where established and ongoing social relationships (based on kinship or reciprocity, for example) could not be fully relied upon.

Our behavioral experiments are well suited to tap exactly these "market norms," as they involved both money and anonymity. Money is most frequently used in market transactions and at least in some circumstances signals a desire to avoid a longer-term nonmarket relationship.[9] Owing to the anonymity in our games, players lack the cues necessary to apply the expectations and motivations associated with other kinds of relationships (Fiske 1992), such as those

TABLE 4.1 *Mean Summary Statistics on Offers and Rejections, by Society*

Society	Dictator Game Offer	Standard Deviation (N)	Ultimatum Game Offer	Standard Deviation (N)	Third-Party Punishment Game Offer	Standard Deviation (N)	Ultimatum Game Phase 2 Minimum Acceptable Offer	Standard Deviation (N)	Third-Party Punishment Game Phase 3 Minimum Offer Not Fined	Standard Deviation (N)
Accra	42	16.9 (30)	44	15.9 (30)	28	16.8 (39)	13	17.3 (30)	28	17.7 (36)
Au	41	19.6 (30)	44	14.5 (30)	33	23.5 (30)	20	21.0 (30)	31	20.0 (30)
Dolgan/ Nganasan	37	20.8 (30)	43	16.2 (30)	n.a.	n.a.	17	20.2 (26)	n.a.	n.a. n.a.
Gusii	33	5.4 (25)	40	4.5 (25)	36	9.4 (30)	38	5.8 (25)	41	5.5 (30)
Hadza	26	25.3 (31)	26	16.6 (31)	26	19.4 (27)	17	17.4 (26)	8	15.0 (24)
Isanga	36	18.3 (30)	38	12.6 (30)	33	17.1 (20)	7	10.1 (30)	33	14.5 (19)
Maragoli	35	17.1 (25)	25	15.6 (25)	34	20.8 (30)	30	7.6 (25)	33	16.6 (23)
Orma	42	15.0 (26)	n.a.	n.a. n.a.	n.a.	n.a. n.a.	n.a.	n.a. n.a.	n.a.	n.a. n.a.
Samburu	40	23.2 (31)	35	19.1 (31)	31	18.0 (30)	6	12.3 (31)	19	10.9 (26)
Sanquianga	47	15.6 (30)	48	10.1 (30)	43	16.0 (32)	12	18.1 (30)	24	21.6 (31)
Shuar	35	19.1 (21)	37	16.5 (21)	37	17.9 (15)	7	13.9 (20)	19	22.2 (15)
Sursurunga	41	18.6 (30)	51	16.3 (30)	37	18.9 (32)	25	20.6 (21)	10	14.6 (25)
Tsimane'	26	15.5 (38)	27	11.1 (36)	20	13.3 (27)	7	5.4 (33)	4	7.8 (23)
U.S./rural Missouri	47	10.3 (15)	48	10.3 (26)	n.a.	n.a. n.a.	28	19.5 (28)	n.a.	n.a. .
Yasawa	35	17.9 (35)	40	17.5 (34)	27	20.4 (30)	7	13.8 (32)	4	7.8 (23)
Total	37	18.9 (427)	39	16.5 (409)	32	18.6 (342)	16	18.0 (387)	22	19.3 (305)

Source: Project data.

Note: This table provides summary statistics for our three experiments.

based on kinship, reciprocity, or status. However, we emphasize that the norms tapped by our experiments may apply more broadly than to just "market exchanges": they may deal with any interactional circumstance not governed by some other form of longer-term social relationship. Such norms may act as the default expectation-motivation sets in some societies whenever relationship-specific information is lacking. As noted earlier, understanding the contextually circumscribed nature of these norms is important because we do interpret our findings as capturing, not anything dispositional about individuals' or societies' general tendencies, but only something about their motivations and expectations in this context (exchanges involving money and anonymity). Suggesting the existence of a default set of rules is different from making a dispositional attribution.

If this line of reasoning is correct, we should expect market integration and fairness norms (offers closer to fifty-fifty) to coevolve such that they are positively correlated across societies. Some might wonder why fifty-fifty is the predicted market norm. We theorize that market norms evolved to allow people to successfully interact without either party having any background information on the other. In the absence of any differentiating information, either about the situation or the other individual, neither party has a claim to more than half.

Since this hypothesis derives from a cultural-institutional evolutionary process, there is no unidirectional causality. Societies with stable market norms readily expand such that those social groups in contact with market societies will be inclined to adopt, and eventually internalize, these prosocial norms. At the same time, those social groups that already possess suitably appropriate norms for market engagement will be better able to readily engagė in successful market interactions (and thus be more market-integrated). In short, we expect that greater market integration will be associated with higher offers in all three experiments.[10] Replicating our earlier UG findings, our new findings presented here support this prediction for offers in all three experiments.

Second, as detailed in chapter 2, we also explore the hypothesis that religious beliefs, rituals, and institutions coevolved with the norms and nonreligious institutions that support larger-scale complex societies (Atran and Henrich 2010; Shariff, Norenzayan, and Henrich 2010). The idea is that cultural evolution increasingly favors potent, moralizing high gods, who, along with the institutional and ritual machinery for instilling and maintaining such beliefs (Henrich 2009), incentivize prosocial behavior toward coreligionists with a range of rewards and punishments, including afterlife incentives. Empirically, anthropological work shows that the presence of high moralizing gods increases with greater societal size and complexity (Roes 1995; Roes and Raymond 2003). Small-scale societies often possess only local, relatively weak, highly anthropomorphic gods who lack moral righteousness (they do both good and bad things from the local's perspective), are unreliable and unpredictable, and cannot—for example—grant eternal life in paradise. Thus, in contrast to the religions that are likely to have dominated most of human history, the Abrahamic religions of Christianity and Islam that have spread globally in the last few thousand years provide a powerful moralizing god who is believed to be dominant over all peoples, omniscient, and equipped with ample powers to reward and punish, including decision authority to provide individuals with eternal bliss or everlasting suffering (Wright 2009). Such religions may have emerged—through a variety of potential processes—to buttress the emerging social norms and institutions that support cooperation in increasingly large-scale societies. In both dictator and ultimatum games, players who report practicing either Christianity or Islam offered more than those professing a traditional local religion.

Finally, in chapter 2 we also discussed theoretical work that uses tools from evolutionary game theory to show how different kinds of mechanisms can stabilize prosocial norms, such as those creating fairness and trust outside of durable longer-term relationships. Roughly speaking, the work shows at least two different classes of norm-stabilization mechanisms: one involving reputational effects in which norm-violators are sanctioned in other interactions through, for example, the withdrawal of help in dyadic exchanges, and a second involving the use of costly punishment (which can work with or without reputational systems). Since the effectiveness of reputational systems in sustaining norms degrades rapidly as communities expand (roughly with the natural logarithm of community size), this research predicts that in large communities norms must be maintained by costly punishment. Thus, in large communities at least some people will have internalized a greater taste for costly punishment, while smaller communities will tend to rely on either costly punishment or indirect sanctioning mechanisms that operate via reputations. The prediction is that costly punishment as measured in our two experiments, increases with community size. And since some theoretical work suggests that reputational breakdown is roughly proportional to the natural logarithm of the group size (Cancho, Solé, and Köhler 2004), we use both community size and the natural logarithm of community size as the key theoretical predictor of variables. This effect emerges in both of our measures of second- and third-party punishment and is robust to the inclusion of demographic and economic control variables, including market integration.

We first study a series of linear regression models that examine the relationship between offers in each game and our variables measuring market integration and participation in a world religion. Then we explore the question of why the predictive effects of participation in world religion disappear in the TPG, while at the same time substantial effects for some economic variables, specifically income, wealth, and household size, emerge as potent predictors of TPG offers. In the next subsection, we examine the relationship between willingness to punish and community size by looking at results from both the UG and TPG. For theoretical reasons, we examine both the effects of community size and the natural logarithm of community size.

Explaining the Variation in Offers

In exploring the variation in offers, we first analyze all the offers together from all three games and then analyze the offers from each game separately. Offers are measured as a percentage of the stake. Our baseline model regresses offers on nine predictor variables: market integration (MI), world religion (WR), age, sex, education, income, wealth, household size (HS), and community size (CS). MI measures market integration as the percentage of the diet in calories purchased in the market, as opposed to homegrown, hunted, fished, or gathered calories. As explained in chapter 3, we use the average MI for each individual's community (village, camp, and so on). WR is an individual-level binary variable, with "1" indicating participation in Islam or Christianity and "0" indicating the practice of a local or traditional religion, or a report of no religion. Our income and wealth measures are derived from detailed protocols eliciting data disaggregated by source (see chapter 3) and have been converted to U.S. dollars and scaled to units of $1,000. Income is measured at the individual level and wealth at the household level. Age is measured in years at the time of the experiment. For education, we created standardized values (with mean zero and standard deviation one) within each population based on self-reports of the number of years a player had spent in formal schooling. We did this because one year of formal schooling is unlikely to be even roughly equivalent across these diverse societies. This approach allows us to get the most from the substantial within-population variation in formal education in our samples. Community size (CS) is the number of individuals (in units of one hundred people) in the local social group, usually a village (though camps were used for the Hadza and the town in Missouri).[11] In most of our populations we sampled from two or more local villages or camps.

In addition to analyses involving these variables, we ran a variety of supplemental analyses. To address the comparability of our income and wealth variables across such diverse populations we performed two sets of supplemental regression analyses that were run in parallel to those shown here, in which our absolute income and wealth measures were replaced with alternative measures. First, income and wealth were replaced with the same variables scaled to the local means and standard deviations for each population, giving us locally relative measures of income and wealth. This allowed us to detect effects based on relative differences in income and wealth. Second, by converting our income and wealth variables into U.S. dollars based on the international exchange rates at the time of the experiments, we might have introduced distortions from their real values based on local purchasing power. Such distortions could have resulted from several factors, including the distance of many of our sites from large market centers (where exchange rates are set) or ephemeral fluctuations in world exchange rates that were unconnected to the material conditions on the ground at our sites. To address this each researcher compiled a list of the local prices at the time of the experiments for twenty-seven commonly used items, including several staples. From this list we found five items that were present in all field sites and were purchased, at least occasionally: sugar, salt, rice, D-cell batteries, and cooking oil. Using the local prices of these items for each site, we converted our measures of income and wealth from the local currency into quantities of each of these items, giving us five new income and wealth variables now measured in quantities of these local consumables. For example, if a subject's yearly income was 1,000 shillings and sugar was locally priced at 5 shillings per kilogram, we converted that person's income measure to 200 kilograms of sugar. That is, he or she could purchase 200 kilograms of sugar locally with his or her yearly income. This approach avoids the use of international exchange rates and grounds people's income and wealth in the kinds of products that are commonly purchased in these locations. We do not present these supplementary regression analyses here. In general they robustly support the conclusions drawn from the analyses presented here.[12]

Analyses of the Variation Across All Offers

Here we combine all offers from each of our three games. For our baseline model we estimated the coefficients in the following equation for each dependent variable (units are in parentheses, if applicable):

$$\begin{aligned}\text{Offer}\,(\%\text{ of stake}) = \text{Constant} + \beta_{MI} * \text{Market Integration}\,(\%) + \beta_{WR} * \text{World Religion}\\ + \beta_I * \text{Income}\,(\$1{,}000) + \beta_W * \text{Wealth}\,(\$1{,}000) + \beta_H * \text{Household Size}\,(\#\text{ of people})\\ + \beta_A * \text{Age (years)} + \beta_S * \text{Sex} + \beta_E * \text{Education} + \beta_{CS} * \text{CS}\,(100\text{ people})\end{aligned}$$

To this equation we add dummy variables for cases in which the offers occur in the TPG or in the UG. The DG provides the reference game for the coefficients on these dummy variables.

Model 1 of table 4.2 is the baseline model for all offers. Market integration and world religion are the only large and significant predictors of all offers. In the models in table 4.2, we used clustered robust standard errors because the analyses compile all observations across our three experiments, which involved some repeated observations from the same individuals. We clustered on individuals to address the problem of the non-independence of repeated observations from the same person. The standard errors are below the coefficients.

TABLE 4.2 *Dictator Game, Ultimatum Game, and Third-Party Punishment Game: Linear Regressions for All Offers*

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Market integration	0.131***	0.182***	0.142***	0.0839***	0.0822***	0.0695***	0.0691***
	(0.0311)	(0.0295)	(0.0307)	(0.0197)	(0.0187)	(0.0182)	(0.0182)
World religion	5.795***	−0.373	5.676***	5.573***	5.506***	6.158***	6.141***
	(2.051)	(2.216)	(1.893)	(1.864)	(1.812)	(1.819)	(1.817)
Income	0.0935	0.0952	0.0655	0.182***	0.200***	0.218***	0.227***
(U.S.$1,000)	(0.0897)	(0.112)	(0.0771)	(0.0644)	(0.0630)	(0.0610)	(0.0606)
Sex	−1.561	−1.389	−1.419	−1.118			
(Female = 1)	(1.413)	(1.384)	(1.368)	(1.279)			
Age (years)	0.0541	0.0667	0.0654	0.0659	0.0610		
	(0.0491)	(0.0466)	(0.0472)	(0.0441)	(0.0423)		
Household size	−0.259	−0.294	−0.366*	−0.183			
	(0.212)	(0.216)	(0.209)	(0.190)			
Education	0.557	0.564	0.574				
(standardized by population)	(0.663)	(0.641)	(0.637)				
Community size	−0.0647	0.0688	−0.0856				
(100 people)	(0.0897)	(0.0684)	(0.0920)				
Wealth	−0.00134	0.000670					
(U.S.$1,000)	(0.00592)	(0.00583)					
Ultimatum game	0.886	0.502	0.987	0.925	1.318	1.353	
	(1.262)	(1.260)	(1.207)	(1.105)	(1.030)	(1.009)	
Third-party punishment game	−2.828*	−4.396**	−3.117*	−4.903***	−4.098***	−4.301***	−4.953***
	(1.712)	(1.718)	(1.679)	(1.535)	(1.398)	(1.387)	(1.254)
Africa		0.194					
		(3.630)					
South America		5.740					
		(3.679)					
Oceania		11.43***					
		(4.188)					
Constant	27.22***	24.37***	27.64***	27.69***	26.03***	28.79***	29.47***
	(3.361)	(4.615)	(3.235)	(3.057)	(2.741)	(2.021)	(1.902)
Observations	840	840	887	987	1,071	1,120	1,120
Number of clusters	541	541	565	634	691	719	719
R-squared	0.086	0.119	0.092	0.080	0.076	0.068	0.067

Source: Project data.

Notes: Clustered robust standard (clustering on individuals) errors are in parentheses below the coefficient. Education has been standardized to a mean of zero and standard deviation of one within each population.

$*p < 0.1$; $**p < 0.05$; $***p < 0.01$

Some might worry that, despite the global-level cultural and linguistic diversity captured in our sample of societies, the relationships observed arise from some shared history among our societies, which would create independence problems for statistical inferences. To address this we included continental controls in model 2 using Eurasia (in which we include the United States, since it is predominantly of European cultural descent) as the reference population. Comparing models 1 and 2, the coefficient on MI increases from 0.13 to 0.18 and remains highly significant. MI is robust to continental-level controls. The coefficient on WR in model 2, however, drops dramatically to become indistinguishable from zero. This occurs because individuals with WR = 0 are scattered quite unevenly across the continents, residing mostly in Africa, and because the impact of world religion does not extend to the TPG. We explore both of these issues later.

Model 2 shows a large positive coefficient on the continental dummy variable for Oceania. This reflects our two populations in New Guinea, which tended to make high offers even after controlling for all demographic and economic differences.

To explore the robustness of our findings for market integration and world religion, we examined alternative specifications by dropping the terms with the least significant coefficients. Doing so allowed us to bring in more data, as we lacked some variables for some populations and sometimes for particular individuals. In table 4.2, models 3 to 7 show the coefficients from five different specifications. Coefficients on both MI and WR remain large and significant across models. The coefficients on MI range from 0.069 to 0.14. The coefficient on WR ranges from 5.5 to 6.16. Taken together, these coefficients indicate that moving from a fully subsistence-based society with a local religion to a fully market-integrated society with a world religion predicts an increase in offers of between thirteen and eighteen percentage points.

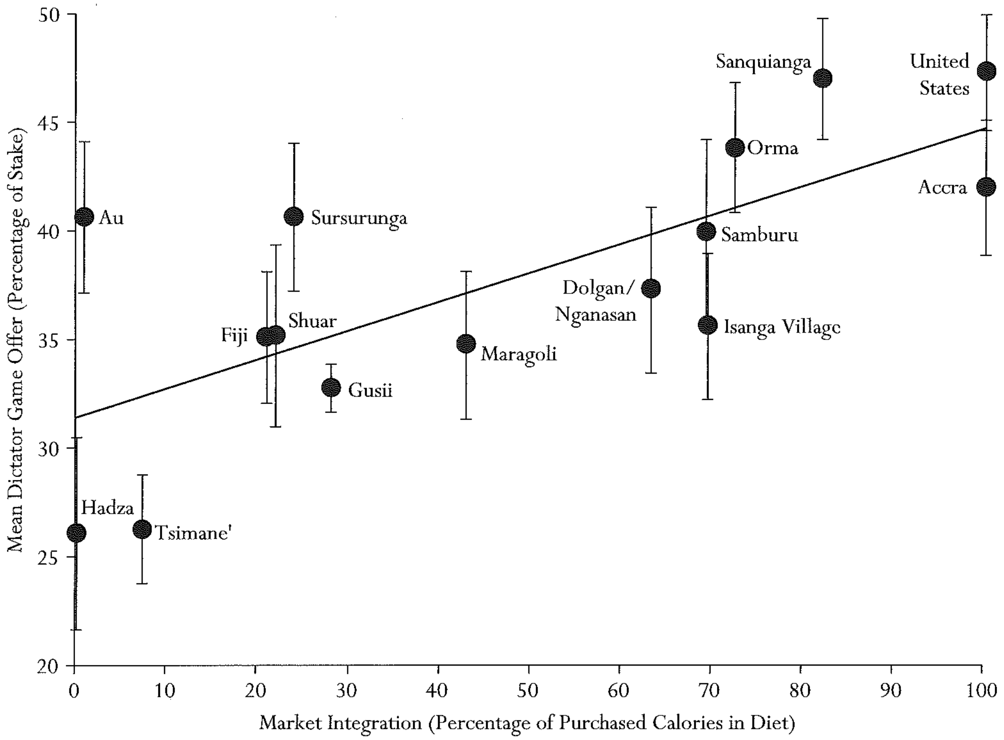
Although the coefficient on income is not significant at conventional levels in the baseline model, it is significant in models 4 to 7. Independent of MI and WR, an increase of $1,000 in income per year increases the percentage offered by between 0.18 and 0.23. As noted in chapter 3, income is mostly wage labor participation, which in itself may capture another dimension of market integration.

The TPG dummy in table 4.2 is either significant or marginally so across all seven specifications. Playing the TPG, as opposed to the DG, predicts a decrease in the percentage offered of between 2.8 and 5 percentage points. Later, we discuss this consistent drop.

While our measures of wealth, income, age, market integration, and household size are all intercorrelated, the correlations are sufficiently small that they do not create collinearity problems for our regressions. Focusing on market integration, the highest correlations in absolute values with our other predictor variables is 0.36 with community size and −0.34 for household size. World religion is the most highly correlated with MI and CS at just under 0.20. Income and wealth show by far the highest correlation at 0.55, which is one of the reasons why we reran everything with wealth dropped.

The findings related to world religion and market integration are robust to modifications in our income and wealth measures. First, in supplemental analyses that recalibrated our income and wealth measure based on local prices, the significant coefficient on MI varies little around a value of 0.13. WR's coefficient varies from about 5.8 to 6.0 and is significant across the board. No other variables in these models are even marginally significant. These findings hold when wealth, household size, and community size are dropped from the models to bring in the Gusii, Dolgan/Nganasan, and Accra samples. Similarly, in our supplementary analyses using relative wealth and income (standardized), using the same procedure of dropping the least significant predictors, both MI and WR remain large and highly significant, while no other coefficients are consistently significant.

FIGURE 4.4 *Mean Dictator Game Offers for Each Population, Plotted Against Mean Value of Market Integration*



Source: Project data.
Note: Error bars are bootstrapped standard errors (bca corrected) on the population mean.

Explaining the Variation in Dictator Game Offers

Now we proceed to analyze the offers for each of the three games separately. For the DG, we begin with a simple bivariate plot in figure 4.4 of mean MI value for each population against mean DG offers. Population mean MI values account for 52 percent of the variation in mean DG offers ($\rho = 0.72$, 0.4–1.0, 95 percent bootstrap CI, $P < 0.01$, N = 15). In designing this second round of our project, we sought out an additional population in New Guinea, where societies are known for this kind of gifting behavior. In our first round, the Au of New Guinea revealed highly unusual behavior, including relatively high offers with little market integration. The Au pattern replicated in our second New Guinea population, the Sursurunga. However, because we targeted a second population that skews our world sample unrepresentatively toward New Guinea, we also examined this relationship with either the Au or the Sursurunga dropped. Dropping the Sursurunga, mean MI accounts for 58 percent of the variation ($\rho = 0.76$, 0.44–0.95, 95 percent bootstrap CI, $P < 0.001$). Dropping the Au instead, but retaining the Sursurunga, mean MI captures 71 percent of the variation ($\rho = 0.84$, 0.59–0.96, 95 percent bootstrap CI, $P < 0.001$).

Proceeding to our multivariate analyses, table 4.3 compares six regression models using the same set of predictor variables deployed in the earlier overall analysis. The baseline model,

TABLE 4.3 *Linear Regressions for Dictator Game Offers*

Variables	(1)	(2)	(3)	(4)	(5)	(6)
Market integration	0.166***	0.205***	0.176***	0.161***	0.123***	0.121***
	(0.0401)	(0.0359)	(0.0364)	(0.0336)	(0.0274)	(0.0263)
World religion	6.429*	2.888	6.933**	7.671**	7.240**	7.823***
	(3.644)	(3.745)	(3.166)	(3.123)	(3.060)	(2.908)
Sex	−2.580	−2.691	−2.595	−2.829	−2.972	−3.367*
(Female = 1)	(2.203)	(3.302)	(2.047)	(1.951)	(1.841)	(1.724)
Household size	−0.114	−0.0812	−0.345	−0.396	−0.239	
	(0.312)	(0.372)	(0.293)	(0.283)	(0.269)	
Education	1.149	1.148	0.937	1.010	0.781	
(standardized by population)	(1.160)	(0.919)	(1.053)	(1.026)	(0.991)	
Income	−0.00349	−0.0534				
(U.S.$1,000)	(0.151)	(0.168)				
Age	−0.0201	−0.0133	−0.0254			
(years)	(0.0803)	(0.0690)	(0.0735)			
Community size	−0.0616	−0.0111	−0.0882	−0.0816		
(100 people)	(0.0909)	(0.0753)	(0.0924)	(0.0901)		
Wealth	0.000699	0.00127				
(U.S.$1,000)	(0.00856)	(0.00727)				
Africa		−0.899				
		(3.754)				
South America		1.548				
		(2.645)				
Occania		6.069				
		(4.442)				
Constant	27.81***	26.89***	29.30***	29.00***	28.76***	26.81***
	(5.238)	(5.594)	(4.820)	(3.675)	(3.609)	(3.110)
Observations	311	311	336	354	384	416
R-squared	0.102	0.113	0.104	0.098	0.087	0.083

Source: Project data.

Notes: Model 2 uses clustered robust standard errors (clustering on population). Other models use robust standard errors; standard errors are in parentheses below the coefficient. Education has been standardized to a mean of zero and standard deviation of one within each population.

$*p < 0.1$; $**p < 0.05$; $***p < 0.01$

model 1, indicates that only MI and WR are large and significant or marginally significant predictors of DG offers. A comparison of models 1 and 2 in table 4.3 reveals that when both continental controls and clustered robust errors (clustering on site[13]) are included in model 2, the coefficient on MI increases from nearly 0.17 to 0.21 and remains highly significant. However, again owing to the uneven distribution across continents of WR = 0 individuals (who are mostly in Africa), the coefficient on WR drops to about 3.0 and becomes insignificant. To explore this we reran our baseline regression for only those populations in Africa. The coefficient on WR in this regression increases to above its value in our baseline regression, going from 6.4 to 6.6, though it is not well estimated owing to the smaller sample size ($p = 0.29$, N = 137, R-squared = 0.13, using robust standard errors). If we drop household size from the regression to increase the sample size, because this coefficient is both small and nonsignificant, the coefficient on WR increases to 7.9 ($p = 0.17$, N = 162, R-squared = 0.11).

Next, we check the robustness of our findings to alternative model specifications. Models 3 to 6 in table 4.3 allow us to examine what happens if we sequentially drop the least significant predictors. While both MI and WR have large coefficients in the baseline model, WR is only marginally significant at conventional levels. However, as the least significant variables are dropped and the sample size expands from 311 to 416 participants, WR becomes significant at conventional levels, with coefficients ranging from 6.9 to 7.8. MI's coefficient remains large and highly significant across all models.

Together, moving from a fully subsistence-oriented society with a traditional religion to a market-integrated economy with a world religion predicts an increase of twenty to twenty-four percentage points in DG offers, which captures the full range of variation across populations in mean DG offers.[14]

Since the effect of participating in a world religion arises from a small portion of our sample (11 percent) that is not widely distributed across our populations, we reran our baseline model using the four populations that have nontrivial frequencies of individuals with WR = 0 and for which we have data for wealth and community size. The coefficient of WR jumps from 6.4 (baseline model) to 14.8 ($p = 0.002$, N = 107, R-squared = 0.23, using robust standard errors). Then, to add our Accra and Siberian samples (giving us six populations), which lack wealth or community size data but do have nontrivial frequencies of those with WR = 0, we reran the same regression, dropping these variables in order to bring in data from these two populations. Now the coefficient on WR is 11.3 ($p = 0.003$, N = 157, R-squared = 0.18). This suggests that the WR effects are not driven by the difference between a few societies lacking a prevalent world religion.

The findings for the coefficients on MI and WR in table 4.3 are also robust to modifications in our income and wealth measures. In supplementary analyses that recalibrated income and wealth to deal with differences in local purchasing power, we found that the coefficient on MI varies little with these recalibrations, ranging from 0.16 to 0.17 (all significant across the board). WR's coefficients range from 6.0 to 6.4 and are marginally significant across the board. When we drop wealth (recalibrated) and CS out of the regression to bring additional samples, all coefficients for WR across the models with the different valuations for income are significant at conventional levels, ranging from 6.0 to 6.6. MI's coefficients are highly significant, ranging from 0.11 to 0.13. We then drop household size from the model, to bring in the Gusii sample. Here MI's coefficients range from 0.12 to 0.13 (all highly significant). WR's coefficients again range from 6.0 to 6.6 and are all significant.

We also sought to address concerns about the comparability of income and wealth across populations by estimating a series of models using standardized versions of income and wealth to obtain a purely relative measure of these variables for each population. We then repeated the procedure used in table 4.3, starting with the baseline model and dropping the least significant variables. These supplementary analyses support the earlier finding for MI and WR, with the coefficients being as large or larger as those in table 4.3. Relative wealth here, however, also has a large and highly significant negative effect on offers, with relatively wealthier people within a given society offering less.

Since it is plausible that income and wealth may have different effects within versus between populations, we also broke both income and wealth down into population mean measures and separated within-population deviations from the population mean. The population averages allow us to assess, for example, whether individuals from populations with higher absolute incomes offer more. Within a population, measures of deviations from a local average allow us to assess whether relative income and wealth predict lower offers. For example, it is theoretically possible that individuals from populations with high mean incomes offer more, while relatively richer people (locally speaking) offer less. Our efforts found no evidence of this in the DG, however, or elsewhere. None of the coefficients on any of these wealth or income measures approached conventional levels of significance. Meanwhile, the coefficients on MI and WR do not vary much from those presented earlier. Note that because the variables *mean* wealth and *mean* income are correlated at 0.99, we never entered both into the model at the same time.

Explaining the Variation in Ultimatum Game Offers

For the UG, table 4.4 repeats the procedure used above for DG offers. Model 1, the baseline model, shows that MI, WR, and age are large, positive, and significant predictors of UG offers.

Model 2 adds continental controls to our baseline model and uses clustered robust standard errors. Here, MI increases from 0.14 in model 1 to 0.19 in model 2. WR drops dramatically, however, owing to the uneven distribution of individuals who do not participate in world religions. As a check, we reran our regression just for our African populations. Now the coefficient on WR jumps back up to 5.6 ($p = 0.25$, N = 112, R-squared = 0.21, using robust standard errors). We examine this more later.

Models 3 to 8 show that both MI and WR are robust, positive predictors of higher UG offers and are highly significant at conventional levels across the board. MI's coefficients range from 0.10 to 0.17, while WR's range from 7.9 to 9.8. For MI, this implies that a 20 percent increase in calories purchased in the market predicts an average increase of between 2.0 and 3.4 percentage points in the UG.

Along with MI and WR, age is also a robust positive predictor of UG offers. Each additional decade of adulthood predicts an increase of between 1.2 and 1.6 percentage points in UG offers. Since previous work has suggested a nonlinear fit for age (Bahry and Wilson 2006), we explored adding a squared term for age. We do not have evidence for this nonlinear age relationship. Unlike MI and WR, age also does not emerge as a significant predictor for offers in the DG and TPG.

Our results suggest that going from a fully subsistence economy with a traditional religion to a fully market-dependent economy with a world religion means an increase in UG offers of eighteen to twenty-six percentage points, covering most of the range of variation we observe across societies.

To further examine our findings for world religion, we reran model 1 (our baseline) using only the four populations in which individuals had nontrivial frequencies of individuals with WR = 0. The coefficient of WR is 9.1 ($p = 0.032$, N = 107, R-squared = 0.29, using robust standard errors). Then, to bring in Accra and the Siberian sites (giving us six populations), we reran the same regression dropping wealth and community size. Now the coefficient on WR is 7.0 ($p = 0.041$, N = 157, R-squared = 0.24, using robust standard errors). This indicates that the effect of world religion is not driven by differences between the few societies lacking world religions and the rest.

We also examined the robustness of our analyses using a modification of our income and wealth variables. First, paralleling our earlier efforts to address differences in local purchasing power in the DG, we used our income and wealth variables measured according to the prices of local consumables to verify that our analyses are not strongly influenced by the use of international currency exchange rates. Coefficients on MI range between 0.14 and 0.15, while WR's coefficients range from 9.96 to 10.1. All are significant, except when income and wealth are revalued using local cooking oil prices; then the p-value for MI slips to 0.056. The coefficient on age ranges from 0.14 to 0.17. In order to include the data from Accra, our Siberian site, and the Gusii, we also ran the same analyses shown earlier, first with wealth and community size dropped, and then with household size dropped. The coefficients on market integration, world religion, and age all remain large and significant across all revaluations.

TABLE 4.4 *Linear Regressions for Ultimatum Game Offers*

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Market integration	0.135**	0.193***	0.132**	0.161**	0.174***	0.141***	0.142***	0.102***
	(0.0651)	(0.0520)	(0.0645)	(0.0628)	(0.0597)	(0.0315)	(0.0312)	(0.0223)
World religion	9.956***	−0.466	9.841***	8.366***	8.447***	8.693***	8.833***	7.940***
	(2.719)	(5.138)	(2.689)	(2.449)	(2.417)	(2.337)	(2.303)	(2.244)
Age (years)	0.138**	0.164*	0.143**	0.147**	0.171***	0.183***	0.163***	0.150***
	(0.0674)	(0.0896)	(0.0658)	(0.0628)	(0.0635)	(0.0617)	(0.0606)	(0.0567)
Income	0.138	0.168**	0.136	0.0798				
(U.S.$1,000)	(0.108)	(0.0579)	(0.112)	(0.0957)				
Community size	−0.256	0.0246	−0.261	−0.319	−0.327	−0.216*	−0.214*	
(100 people)	(0.222)	(0.129)	(0.221)	(0.230)	(0.233)	(0.116)	(0.115)	
Education	0.959	1.014	1.067	0.987	1.062	1.182		
(standardized by population)	(0.875)	(0.806)	(0.860)	(0.818)	(0.803)	(0.744)		
Sex	−1.362	−1.605						
(Female = 1)	(1.960)	(2.790)						
Wealth	−0.00713	−0.00402	−0.00679					
(U.S.$1,000)	(0.00824)	(0.00417)	(0.00856)					
Household size	−0.253	−0.283	−0.251	−0.299	−0.272			
	(0.271)	(0.475)	(0.269)	(0.273)	(0.276)			
Africa		−0.742						
		(3.446)						
South America		7.656*						
		(3.491)						
Oceania		19.04***						
		(5.290)						
Constant	22.90***	18.84***	22.27***	23.56***	22.21***	20.39***	20.98***	21.39***
	(3.824)	(3.818)	(3.677)	(3.657)	(3.700)	(3.193)	(3.171)	(3.119)
Observations	294	294	294	316	319	346	347	377
R-squared	0.148	0.263	0.146	0.146	0.147	0.132	0.127	0.111

Source: Project data.

Notes: Model 2 uses clustered robust standard errors (clustering on site); other models use robust standard errors; standard errors are in parentheses below the coefficient. Education has been standardized to a mean of zero and standard deviation of one within each population.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Then, to further address concerns about the comparability of income and wealth across populations, we estimated a series of models using standardized versions of wealth and income. We repeated the procedure of dropping nonsignificant variables, as in table 4.4. These supplemental analyses show that MI and WR have large coefficients and remain significant (usually highly significant) across all specifications. The coefficient on age also remains significant at conventional levels across specifications.

To address the possibility that income and wealth have different effects within versus between populations, we broke both income and wealth down into population mean measures and separated within-population deviations from the population mean. Analyses using these versions yield no evidence of income or wealth effects in the UG. None of the coefficients on any of these wealth or income measures approached conventional levels of significance. Meanwhile, the coefficients on MI and WR remain significant and fairly stable. This remains true if we drop wealth, CS, and HS to bring in all of our samples.

Explaining the Variation in Third-Party Punishment Offers

For offers in the third-party punishment game, our analyses once again show the predictive importance of market integration, but in contrast to our analyses of DG and UG offers, they do not show consistent associations between offers and world religion. These analyses also find that lower incomes, greater wealth, and smaller households are all associated with higher offers (model 1 in table 4.5). We explore the disappearance of the effects for world religions and the emergence of economic variables as predictors in the TPG in the next section.

One way to address concerns about both shared culture histories and the non-independence of individuals from the same groups is to compare model 1 (baseline) with model 2 (with continental controls and clustered robust standard errors) in table 4.5. Since we lack data from the Siberian and U.S. sites for the TPG, the regression uses Africa as the reference for the continental dummies. These modifications have no quantitative effect on our main results in the TPG. The coefficients on market integration, income, wealth, and household size remain large and significant with the addition of continental controls and use of clustered robust errors. The coefficient on community size increases and becomes highly significant in model 2.

Models 3 to 6 allow comparisons of the coefficients on MI as the least significant predictors are sequentially dropped from the model. MI remains significant and meaningful. An increase of 20 percent in calories purchased in the market is associated with an increase of roughly two percentage points. For income, a $1,000 increase is associated with a *decline* of roughly 2.2 percentage points. For wealth, each $1,000 increment is associated with a roughly 1.3-percentage-point increase. For the size of players' households, an additional member is associated with approximately a one-percentage-point decline in offers.

As earlier, we also examined the robustness of our findings to modifications in our income and wealth variables. First, when we deploy income and wealth revalued using local prices of consumables, supplemental analyses show that the coefficients on MI remain large, varying from 0.08 to 0.11, and significant or marginally significant across all models, except when salt is used to recalibrate (there $p = 0.12$ for MI). Wealth and household size remain significant across all six models, as does household size. Income is significant in all models except where the price of cooking oil is used to recalibrate our measures; there it is marginally significant.

Second, we also estimated a series of models using standardized versions of wealth and income. Market integration remains marginally significant across most specifications, although if only MI, income (standardized), and household size are used as predictors, MI's effects disappear. In these models, the significant effects of income (standardized), wealth (standardized), and household size also vanish. Importantly, compared to the models using either U.S. dollars or local consumable prices for wealth and income, these models explain much less of the variation in offers, with the percentage of variance explained dropping from about 10 percent to 4 percent or less. Capturing the effect of MI in the TPG would seem to depend on controlling for the actual values of income and wealth in a manner not observed in the DG and UG.

TABLE 4.5 *Linear Regressions for Third-Party Punishment Game Offers*

Variables	(1)	(2)	(3)	(4)	(5)	(6)
Market integration	0.101**	0.100***	0.102**	0.103**	0.0987**	0.121***
	(0.0503)	(0.0171)	(0.0503)	(0.0501)	(0.0490)	(0.0427)
Wealth	1.279***	1.028***	1.286***	1.288***	1.292***	1.174***
(U.S.$1,000)	(0.262)	(0.129)	(0.258)	(0.256)	(0.253)	(0.248)
Income	−2.200**	−3.167***	−2.201**	−2.184**	−2.061**	−1.841*
(U.S.$1,000)	(0.972)	(0.666)	(0.966)	(0.950)	(0.911)	(0.947)
Household size	−1.097**	−1.030**	−1.112**	−1.115***	−1.128***	−1.006**
	(0.436)	(0.353)	(0.430)	(0.429)	(0.428)	(0.421)
Community size	0.126	0.306***	0.126	0.127	0.127	
(100 people)	(0.100)	(0.0384)	(0.101)	(0.0999)	(0.0991)	
Age	0.0529	0.0281	0.0619	0.0630	0.0631	
(years)	(0.0877)	(0.0784)	(0.0836)	(0.0834)	(0.0836)	
Sex	−1.035	−0.719	−0.913	−0.881		
(Female = 1)	(2.620)	(2.346)	(2.617)	(2.590)		
World religion	0.836	−3.672**	0.584			
	(3.027)	(1.381)	(2.954)			
Education	−0.495	−0.985				
(standardized by						
population)	(1.449)	(2.004)				
South America		10.48***				
		(1.633)				
Oceania		9.404***				
		(1.537)				
Constant	30.62***	28.19***	30.43***	30.78***	30.49***	33.45***
	(5.297)	(4.590)	(5.264)	(5.022)	(4.875)	(3.035)
Observations	235	235	235	235	235	242
R-squared	0.102	0.135	0.101	0.101	0.101	0.082

Source: Project data.

Notes: In the TPG, and only the TPG, we lack data on income and wealth among the Tsimane'. Model 2 uses clustered robust standard errors (clustering on site); other models use robust standard errors; standard errors are in parentheses below the coefficient. Education has been standardized to a mean of zero and standard deviation of one within each population.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

Offer Regressions Using the Minimum Acceptable Offer as a Predictor

Offers in the UG and TPG measure some combination of internalized motivations (regarding fairness, equality, relative payoffs, and so on) and beliefs about the likelihood of punishment or rejection. In the DG, there is no punishment (unless participants do not believe that the game is anonymous, a concern we discuss later), so we assume that the DG measures intrinsic motivation. To assess the degree to which the UG and TPG capture intrinsic motivations versus fear of punishment, we reran our baseline regressions with each population's mean MinAO for the UG

TABLE 4.6 *Linear Regressions for Offers in the Ultimatum Game and the Third-Party Punishment Game With and Without the Mean Minimum Acceptable Offer as a Predictor*

Variables	Ultimatum Game (1)	Ultimatum Game (2)	Third-Party Punishment Game (3)	Third-Party Punishment Game (4)
Market integration	0.135**	0.146*	0.101**	0.0946*
	(0.0651)	(0.0772)	(0.0503)	(0.0529)
World religion	9.956***	9.723***	0.836	0.403
	(2.719)	(2.729)	(3.027)	(3.155)
Mean MinAO UG/TPG		0.156		0.0863
		(0.159)		(0.174)
Age (years)	0.138**	0.132**	0.0529	0.0508
	(0.0674)	(0.0667)	(0.0877)	(0.0883)
Income	0.138	0.0971	−2.200**	−2.036**
(U.S.$1,000)	(0.108)	(0.124)	(0.972)	(0.954)
Community size	−0.256	−0.296	0.126	0.0960
(100 people)	(0.222)	(0.269)	(0.100)	(0.121)
Education	0.959	1.089	−0.495	−0.535
(standardized by population)	(0.875)	(0.880)	(1.449)	(1.449)
Sex	−1.362	−1.530	−1.035	−0.644
(Female = 1)	(1.960)	(1.986)	(2.620)	(2.688)
Wealth	−0.00713	−0.00917	1.279***	1.295***
(U.S.$1,000)	(0.00824)	(0.00904)	(0.262)	(0.261)
Household size	−0.253	−0.159	−1.097**	−1.097**
	(0.271)	(0.315)	(0.436)	(0.439)
Constant	22.90***	20.65***	30.62***	29.65***
	(3.824)	(4.589)	(5.297)	(5.670)
Observations	294	294	235	235
R-squared	0.148	0.153	0.102	0.103

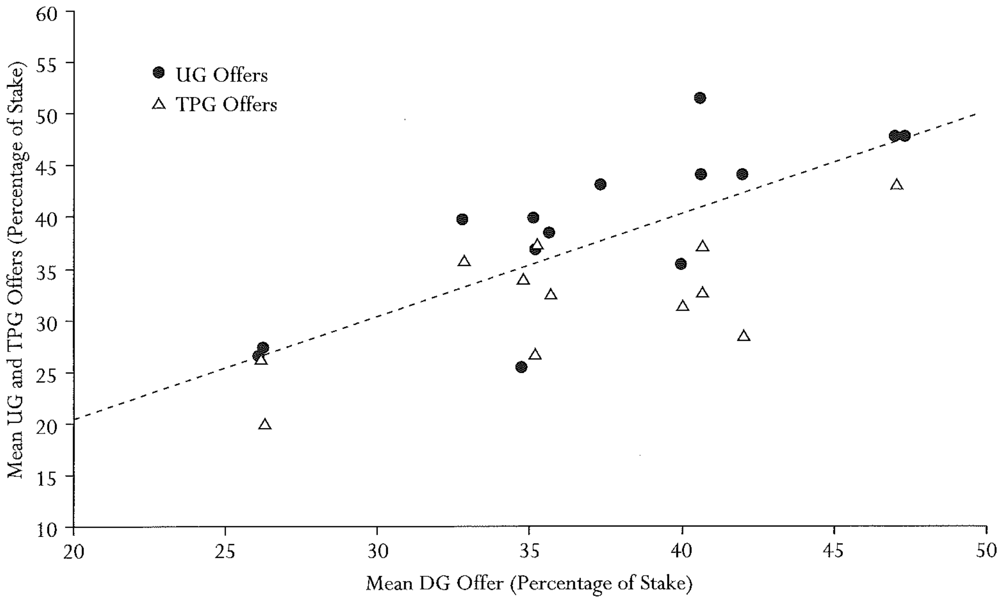
Source: Project data.
Note: Robust standard errors are in parentheses below the coefficient.
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

or TPG, as appropriate, included as an additional explanatory variable (table 4.6). The idea is that if individuals are responding to the local chances of rejection or punishment in making their offers, the mean MinAO of an individual's population should capture something of this response. A comparison of models 1 and 2 for the UG and models 3 and 4 for the TPG shows little impact on our key theoretical variables. The coefficient on MI in the UG increases from 0.14 to 0.15 with the addition of mean MinAO (in the UG) as a predictor, while in the TPG the coefficient on MI changes little. In both the UG-offer and the TPG-offer models, the coefficients on MinAO mean are positive, as expected, with more punishment predicting greater offers, though neither is significant at conventional levels. Little else changes in the analysis.

Why Does the Effect of World Religion Disappear in the Third-Party Punishment Game?

Offers in the TPG differ from offers in the DG and UG in two interesting respects. First, TPG offers are generally lower than offers in the other two games. Overall mean offers across these three games are 32 (TPG), 37 (DG), and 39 (UG) percent of the stake. Of our twelve populations that played the TPG, only three had TPG mean offers greater than their DG mean offers (greater by only 1.7 percentage points, on average), while nine had mean offers in the TPG less than in the DG (less by 6.3 percentage points on average). Both parametric and nonparametric comparisons of means show the TPG means are lower than both the DG and UG means (all $p < 0.02$). The DG and UG mean offers, however, cannot be distinguished at conventional levels of confidence.[15] Figure 4.5 plots mean DG offers for each population along the *x*-axis and mean UG and TPG offers on the *y*-axis. The dotted line is the unity line, such that dots above indicate the UG or TPG mean is greater than the DG mean offer. This plot shows that while TPG offers tend to be lower than DG offers, UG mean offers tend to be a bit higher. Of our thirteen populations in the UG, all but two have mean UG offers greater than mean DG offers.

FIGURE 4.5 *Mean Dictator Game Offers for Each Population Plotted Against Both Their Mean Offers in the Third-Party Punishment Game and in the Ultimatum Game*



Source: Project data.

The second respect in which TPG offers differ from UG and DG offers is that while world religion is a potent predictor for the latter, it is not for the former. Moreover, while WR is not associated with TPG offers, our three indicators of material well-being—income, wealth, and household size—all are.

Here we explore one hypothesis that may help account for both of these phenomena: that is, the general drop in TPG offers (versus DG and UG offers) and the reversal in the predictive importance of WR versus the economic variables (income, wealth, and household size). We acknowledge at the outset that this is post hoc theorizing. Although our hypothesis emerges from a long-standing and well-documented empirical phenomenon termed "crowding out," which suggests that external incentives (for example, third-party punishments and rewards) can—under some circumstances—reduce other-regarding intrinsic motivations (Bowles 2008;

Frey and Jegen 2004). We suspect that adding a third party who can impose fines—as in our TPG—may drive out some of people's intrinsic motivation toward equal divisions, whether those are based on altruism, inequity aversion, norm adherence, or some other motivation. This phenomenon may have manifested relatively strongly in our version of the TPG because our setup prevented third parties from punishing sufficiently to discourage purely self-interested players from offering zero.

To explore this hypothesis we reasoned that any differences between DG or UG offers and TPG offers owing to the crowding-out effect should be manifest in the relative predictive effects of world religion versus income and/or wealth (or wealth per household member). Participation in a world religion may imbue individuals with ethical principles or prosocial motivations (toward those beyond close durable relationships) that they seek to demonstrate to themselves, God, or others. World religions, through ritualized reminders, may make us judge ourselves in ways more critical, explicit, and harsh than other systems of belief. The threat of a fine might either destroy the signaling content of an individuals' generosity or diminish the intrinsic pleasure or satisfaction derived from it (or both). For example, one might take pleasure in personal altruism toward one's fellow humans or experience the pride and approval of the All Mighty by taking the time to donate blood; however, if one gets paid $10 for the blood, that good feeling and sense of divine approval may disappear, while the cash might not provide sufficient compensation for the cost in time and discomfort. With regard to our experiments, this reduction in nonpecuniary intrinsic motivations may be moderated by a player's wealth or income, which would figure into the impact of a game decision on one's material self-interest. For example, when richer individuals trade off their—now reduced—fairness motivations against their material self-interest, fairness motivations are relatively more important (compared to those of poorer individuals) because the same amount of money has a smaller impact on the material self-interest of richer individuals (concave utility functions). Alternatively, less-well-off people may be looking for a rationalization to be less fair, and they find it once the responsibility for enforcing fairness is deflected to a third party. From this, the prediction is straightforward: populations with greater participation in world religions will show a larger difference between their DG and TPG offers once economic variables are controlled for. Greater income or wealth ought to favor smaller differences, since the decision has a smaller effect on material self-interest when values are higher.

To analyze this we had to use the population means because different players were used in the DG and TPG. We regressed the mean population differences between the DG and TPG on mean values of WR, income, and wealth (absolute and per household member). As noted previously, however, at the population aggregate level, income and wealth (and wealth per household member) are all highly intercorrelated (all > 0.9), so we paired each with WR in table 4.7 (models 1, 2, and 3). The effects of WR and each of our economic variables go in the direction predicted by our hypothesis. The coefficients on WR are large in all three models, though only significant or marginally significant when income or wealth per household member is used. The coefficients on income, wealth, and wealth per household member are all negative and large, though only the two wealth measures are significant at 0.05. To include all twelve of our data points we use income in models 4 and 5.

Model 4 adds the mean MinAO in the TPG to model 1, to control for the effects of the threat of punishment. The coefficient on WR increases, achieving conventional levels of significance. Model 5 adds MI to this, and the coefficient on WR remains significant. Societies with more participation in a world religion show larger drops in offers from the DG to the TPG. Model 5 captures 57 percent of the variation in the difference between DG and TPG mean offers.

Table 4.8 provides analyses parallel to those observed in table 4.7, now for the difference between mean UG and TPG offers. Model 1 shows that both WR and income have effects in the

TABLE 4.7 *Linear Regressions on the Difference Between Mean Dictator Game and Third-Party Punishment Game Offers*

Variables	(1)	(2)	(3)	(4)	(5)
World religion	6.350**	4.931	5.443*	7.466**	6.926**
	(2.696)	(2.992)	(2.471)	(3.064)	(2.826)
Income	−2.973			−2.462	−3.482
(U.S.$1,000)	(2.038)			(1.946)	(2.215)
Wealth		−1.135**			
(U.S.$1,000)		(0.344)			
Wealth per household member			−5.378**		
			(2.212)		
Mean MinAO TPG				−0.0987	−0.178
				(0.102)	(0.122)
Market integration					0.0927*
					(0.0450)
Constant	0.846	1.852	1.156	1.644	0.849
	(2.178)	(2.839)	(2.204)	(2.489)	(1.551)
Observations	12	11	10	12	12
R-squared	0.202	0.468	0.416	0.254	0.572

Source: Project data.
Notes: Robust standard errors are in parentheses below the coefficient. Wealth per household member is household wealth divided by household size.
*p < 0.1; **p < 0.05; ***p < 0.01

TABLE 4.8 *Linear Regressions for the Difference in Ultimatum Game and Third-Party Punishment Game Offers*

Variables	(1)	(2)	(3)
World religion	11.02***	11.88**	11.86**
	(3.237)	(3.669)	(3.668)
Income	−4.322	−4.295	−5.585
(U.S.$1,000)	(2.718)	(2.950)	(3.075)
Difference (UG MinAO − TPG MinAO)		0.100	0.255*
		(0.130)	(0.128)
Market integration			0.0864
			(0.0618)
Constant	−0.654	−0.854	−2.499
	(1.862)	(1.810)	(2.074)
Observations	12	12	12
R-squared	0.255	0.279	0.369

Source: Project data.
Note: Robust standard errors are in parentheses below the coefficient.
*p < 0.1; **p < 0.05; ***p < 0.01

predicted directions, with large coefficients, though only WR's coefficient is significant at conventional levels. Adding the difference in the mean MinAOs in the UG and TPG as a predictor in model 2 to control for differences in the expectation of punishment shows that the coefficients on both income and WR change little. Model 3 shows that adding MI to this estimation changes little, though now the coefficient on the difference in MinAO is larger and marginally significant.

Lest the reader be concerned that this drop in offers found in moving from the DG to the TPG occurs only in the unusual contexts of our diverse small-scale societies—where, for example, antisocial punishment may occur (Henrich et al. 2006; Herrmann, Thöni, and Gächter 2008)—our protocols have also been administered to university students in the United States (see chapter 9). The undergraduate mean offers are 32 percent (DG), 39 percent (UG), and 27 percent (TPG). That undergraduates show a 5 percent drop in moving from the DG to the TPG makes them about average in our sample of societies. This suggests that alternative explanations based on antisocial punishment are not likely (Fehr, Hoff, and Kshetramade 2008; Herrmann et al. 2008).

Note that we have not included these student samples in our analyses because, unlike all our other populations, these participants are not random samples of typical communities but instead represent a narrow age range of individuals of high socioeconomic status (SES) (Emory University), most of whom are not economically self-sufficient. It is well established that student behavior in experiments like ours has not reached its adult plateau (Carpenter et al. 2005; Harbaugh et al. 2002; Sutter and Kocher 2007). Student prosociality in experiments hugs the lower bound of prosociality observed in older adults (Bellemare and Kroger 2007; Bellemare et al. 2008), and student prosociality continues to increase over the university years (Carter and Irons 1991). Entering these samples into our database for this analysis would mix adult variation among populations with developmental variation among a sample that is also skewed socioeconomically relative to the U.S. population as a whole.

While necessarily tentative, the analysis presented in this section provides some very preliminary support for our religion-crowding-out hypothesis described earlier, as world religion and the economic variables seem to account for a nontrivial fraction of the differences between the TPG offers and offers in the other two experiments. Our small sample of twelve populations makes stronger conclusions impossible.

Why Does the Punishment Threat Fail to Reduce Offers in the Ultimatum Game?

When we performed analyses parallel to those in the previous section on the differences between DG and UG offers, nothing was ever significant at conventional levels, and the R-squared never went above 0.06. However, given our tentative results suggesting that crowding-out effects are important, we must ask why the threat of punishment present in the UG does not cue the same loss of intrinsic motivation that we observe for the TPG. We consider two possibilities and evaluate their consistency with the data.

Hypothesis 1 The strength of punishment is structurally stronger in the UG compared to the TPG. For low UG offers, punishment is cheap for player 2 and costly for player 1. For higher offers in the UG, nearing fifty-fifty, punishment becomes quite expensive for player 2, and the costs inflicted on player 1 are lower. In contrast, in our particular design of the TPG, punishing low and high offers costs exactly the same amount and the damage to player 1 is the same. Consequently, the threat of punishment in the TPG may be structurally too weak to fully compensate for the loss of intrinsic motivation created by the threat of punishment. An income-maximizer in our TPG will still give zero even when he believes that punishment is likely. (This is not the case in the UG.) This *compensatory hypothesis* suggests that the same motivational loss occurs for both the UG and the TPG, but that only the UG has sufficiently potent punishment to compensate.

We have examined the compensatory hypothesis, and it does not hold up. The above-mentioned analyses of the difference between mean DG and UG offers find only small and nonsignificant coefficients on MinAO in the UG. Populations with greater punishment in the UG do not show greater differences in mean offers. We also substituted other statistics for the mean, including the eightieth- and ninetieth-percentile MinAO. (This is the offer that gives an 80 percent or 90 percent chance of acceptance.) These measures of the threat of punishment also showed no relationship to the difference in mean DG and UG offers. In short, the difference between DG and UG offers does not seem related to any differences in the actual threats of punishment, which are substantial. This remains true even when the mean DG offer is controlled for, thereby addressing differences in intrinsic motivation. Note, however, that we are not suggesting that the threat of punishment does not have an important influence on the decision of player 1; it is just that the difference between the TPG and the UG cannot be accounted for by this. (Detailed modeling and analysis showing how the threat of punishment can be linked to players' decisions can be found in Barr et al. 2009; see also Henrich et al. 2006).

Hypothesis 2 There may be an important psychological difference between the threat of second-party punishment, which can be motivated by revenge, and the threat of third-party punishment. Perhaps the possibility of second-party punishment is perceived as endemic to any interaction, while third-party punishment is cognitively perceived and encoded as an external source of rewards and punishments. We do not have a direct test for this hypothesis. However, our findings are broadly consistent with it in that they show the same patterns for explaining the difference between mean DG offer and mean TPG offers, and between mean UG offers and mean TPG offers.

Finally, we want to emphasize that in this section we are making a post hoc attempt to grapple with something that unexpectedly emerged from our findings. Our project was not designed to address these questions, so these unexpected patterns may be merely an artifact of our design. The DG and UG were played in rapid succession, with the same people as player 1 in both games (without feedback in between).[16] The TPG was mostly played with different participants, or alternatively, at least three weeks later in the same community, or it was played in a different community within the same population (see chapter 3). Comparing the DG and TPG raises the fewest concerns, since the DG always came first and the TPG was played much later and/or with different people (for whom it was the first game). Nevertheless, comparisons of the difference between the DG and the UG, on the one hand, and the DG and the TPG, on the other, raise a variety of methodological concerns that cannot be adequately addressed given our design. Overall, we think the only conclusion this analysis strongly favors is the need for further investigation.

Punishment in the Third-Party Punishment and Ultimatum Games

Our analyses of rejections in the UG and fining in the TPG indicate that people from larger communities engage in more costly punishment. As laid out in chapter 2, theoretical work examining various mechanisms capable of sustaining costly norms, including those associated with fairness among strangers, suggests that smaller groups can sustain costly norms with reputational systems that, for example, allow individuals to withdraw help from norm-violators instead of punishing them at a personal cost (Panchanathan and Boyd 2004). Larger cooperative groups require costly punishment because reputational systems rapidly break down as group size increases (Panchanathan and Boyd 2003). At the same time as reputational systems are collapsing, the anonymity of larger groups mitigates the threat of counterpunishment—punishing someone back who punished you—thereby increasing the range of conditions in which costly punishment can sustain larger-scale cooperation and prosocial norms. That is, the possibility of counterpunishment threatens the effectiveness of punishment in maintaining cooperation in smaller groups. This problem declines in larger populations because anonymity increases, so "pushing back" is more costly or difficult. This line of theory predicts that larger coherent communities must have diffuse punishment—otherwise, they would break down and not remain large communities for long. Since some theoretical work suggests that reputational breakdown should be roughly proportional to the natural logarithm of the group size (Cancho et al. 2004), here we use both community size and the natural logarithm of community size (LNCS) as key theoretical predictors of variables. We also discuss the use of the square of community size.

As described earlier, we reduced our vectors of punishment decisions in both the TPG and the UG to a single number called a minimum acceptable offer (MinAO) for each person. This is the lowest offer below 50 for which an individual would not reject (in the UG) or pay to punish (in the TPG). To analyze the MinAOs in both the UG and the TPG, we used an ordered logistic regression (OLR) instead of an OLS regression because our dependent variables (MinAOs) are both discrete and bimodally distributed. The diagnostics for our initial linear regressions indicated that basic assumptions were being dramatically violated. The OLR assumes that the dependent variables are discrete and rank-ordered, but that the distance between discrete ranks is not meaningful.[17]

All coefficients shown and discussed here are reported as odds ratios, for ease of interpretation.

Minimum Acceptable Offers in the Third-Party Punishment Game

The third-party punishment game provides the most straightforward measure of an individual's willingness to engage in diffuse, norm-enforcing, costly punishment because, unlike the ultimatum game, the motivation behind this punishment cannot be simply revenge for a direct personal slight. Model 1 in both table 4.9 and table 4.10 shows that CS and LNCS (natural logarithm of community size) are large and significant predictors of MinAOs in the TPG. Comparing model 1 (baseline) and model 2 in table 4.9 shows that adding both continental controls and using clustered robust standard errors results in little change to our findings. Models 3 through 9 test the robustness of community size as a predictor, following the same procedure. The odds ratio (standing in for the coefficient) tells us how much an increase of one hundred people in community size influences an individual's chances of punishing in the next-higher MinAO category (for example, the increased chance of delivering an MinAO of 20 percent instead of 10 percent). The coefficient on CS remains large and highly significant across all specifications. Figure 4.6 graphically illustrates the effect size of CS. Figure 4.7 graphically presents the estimated association between LNCS and the TPG MinAO.

The reader should note that the coefficient on WR is positive and significant, or marginally so, across all these specifications. Participating in a world religion predicts more punishment in the TPG. We think this is worth noting, but do not make much of it, because if we use the LNCS (table 4.10), the WR effect evaporates.

Table 4.10 parallels the analysis in table 4.9, now using LNCS in place of CS. Comparing models 1 and 2 shows that the odds ratio for the coefficient of LNCS increases from 2.4 to 2.7

TABLE 4.9 *Ordered Logistic Regressions for the Minimum Acceptable Offer in the Third-Party Punishment Game, Using Community Size*

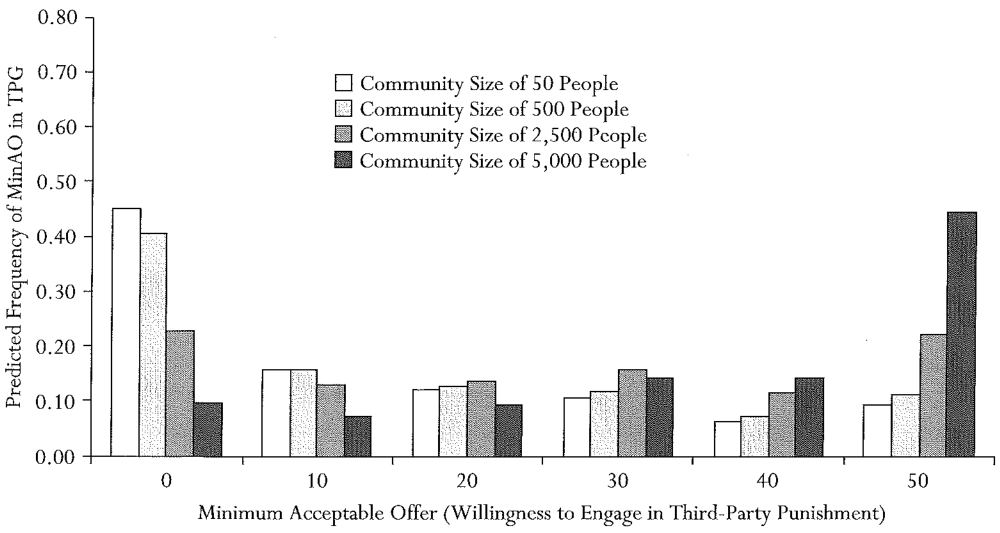
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Community size	1.042***	1.044***	1.041***	1.050***	1.050***	1.050***	1.055***	1.054***	1.056***
(100 people)	(0.0108)	(0.0143)	(0.0103)	(0.00802)	(0.00806)	(0.00806)	(0.00731)	(0.00731)	(0.00724)
World religion	2.055**	2.312**	2.267**	2.348**	2.346**	2.288**	1.772*	1.772*	
	(0.701)	(0.925)	(0.745)	(0.800)	(0.798)	(0.757)	(0.547)	(0.538)	
Sex	0.611*	0.603	0.647*	0.659*	0.663*	0.679*	0.761		
(Female = 1)	(0.170)	(0.224)	(0.166)	(0.156)	(0.159)	(0.159)	(0.167)		
Wealth	0.981	0.976	0.972	0.976	0.975	0.972			
(U.S.$1,000)	(0.0479)	(0.0468)	(0.0462)	(0.0446)	(0.0434)	(0.0411)			
Market integration	1.007	1.003	1.003	1.001					
	(0.00615)	(0.00754)	(0.00541)	(0.00476)					
Income	0.938	0.911	0.949	0.964	0.968				
(U.S.$1,000)	(0.0885)	(0.105)	(0.0893)	(0.0932)	(0.0933)				
Household size	0.995	0.999	0.995						
	(0.0400)	(0.0446)	(0.0364)						
Age	0.999	0.997							
(years)	(0.0113)	(0.0140)							
Education	0.959	0.938	0.885	0.894	0.892	0.888			
(standardized by population)	(0.139)	(0.129)	(0.116)	(0.108)	(0.107)	(0.105)			
South America		1.417							
		(0.599)							
Oceania		0.823							
		(0.579)							
Observations	197	197	211	242	242	243	268	268	269
Pseudo-R-squared	0.0562	0.0580	0.0479	0.0734	0.0734	0.0733	0.0776	0.0759	0.0725

Source: Project data.

Notes: Model 2 uses clustered robust standard errors (clustering on site); other models use robust standard errors; standard errors are in parentheses below the coefficients. Education has been standardized to a mean of zero and standard deviation of one within each population. Coefficients are reported as odds ratios.

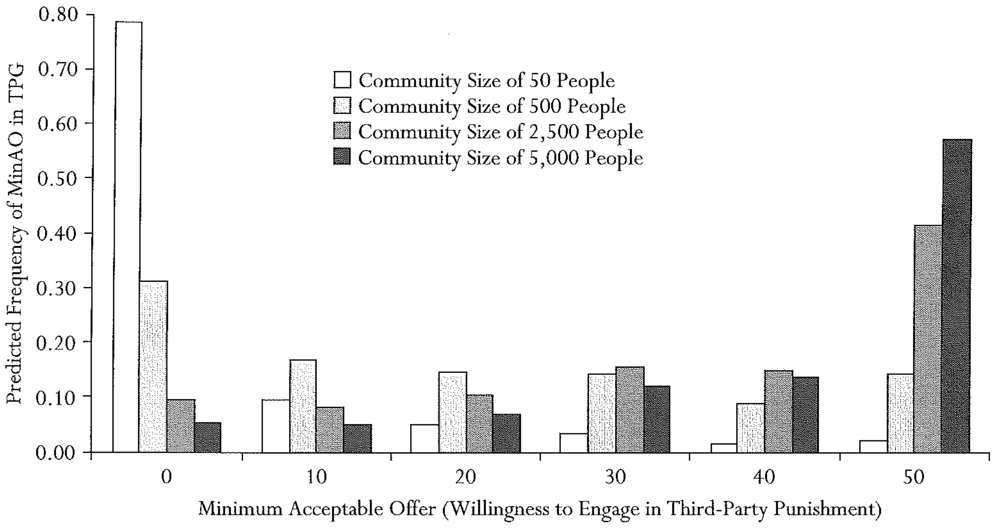
*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

FIGURE 4.6 *Community Size Predicting Third-Party Punishment Game Minimum Acceptable Offers*



Source: Project data.
Note: The coefficients used to create the plots are from an ordered logistic regression containing the eight other variables.

FIGURE 4.7 *LNCS Predicting Third-Party Punishment Game Minimum Acceptable Offers*



Source: Project data.
Note: The coefficients used to create the plots are from an ordered logistic regression containing the eight other variables.

TABLE 4.10 *Ordered Logistic Regressions for the Minimum Acceptable Offer in the Third-Party Punishment Game, Using LNCS*

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
LNCS	2.479***	2.746***	2.583***	2.249***	2.283***	2.270***	2.256***
	(0.416)	(0.743)	(0.425)	(0.219)	(0.217)	(0.213)	(0.208)
Market integration	0.981**	0.972*	0.978***	0.988**	0.988**	0.987**	0.989**
	(0.00883)	(0.0141)	(0.00810)	(0.00542)	(0.00522)	(0.00522)	(0.00514)
Household size	0.989	0.991	0.980				
	(0.0398)	(0.0430)	(0.0359)				
Sex	0.641	0.625	0.658	0.693	0.692		
(Female = 1)	(0.180)	(0.228)	(0.172)	(0.155)	(0.155)		
World religion	1.328	1.413	1.305	1.203			
	(0.472)	(0.547)	(0.442)	(0.429)			
Income	0.964	0.905	0.956				
(U.S.$1,000)	(0.0942)	(0.114)	(0.0873)				
Wealth	0.987	0.977					
(U.S.$1,000)	(0.0482)	(0.0345)					
Age	0.999	0.995					
(years)	(0.0114)	(0.0132)					
Education	0.960	0.929	0.916	0.806*	0.810*	0.829	
(standardized by population)	(0.135)	(0.136)	(0.120)	(0.0933)	(0.0940)	(0.0959)	
Africa		0.469*					
		(0.213)					
Oceania		0.422**					
		(0.151)					
Observations	197	197	212	267	267	267	269
Pseudo-R-squared	0.0868	0.0923	0.0840	0.0955	0.0952	0.0923	0.0900

Source: Project data.
Notes: Model 2 uses clustered robust standard errors (clustering on site); other models use robust standard errors; standard errors are in parentheses below the coefficients. Education has been standardized to a mean of zero and standard deviation of one within each population. Coefficients are reported as odds ratios.
$*p < 0.1$; $**p < 0.05$; $***p < 0.01$

and remains highly significant when phylogenetic controls for shared culture history are added and clustered robust standard errors are used. Little else changes in this specification. If anything, adding continental-level controls increases the effects of LNCS. In contrast to the results shown for CS, these models also indicate that MI has a negative effect on MinAO (that is, an odds ratio less than one). This indicates that more market-integrated societies punish less.[18]

We also tried using combinations of CS, CS-squared, and LNCS. The goodness-of-fit measures are somewhat better for the LNCS vis-à-vis CS and CS-squared (together), as we would expect from theory. When CS is entered along with CS-squared, both are significant. The effect of CS is large and positive (or odds ratio > 1) and the effect of CS-squared is smaller and negative, suggesting a declining effect of CS as populations get large. Goodness-of-fit measures (both pseudo-R-squared and AIC [Akaike information criterion]), however, still indicate that entering LNCS alone is superior to entering both CS and CS-squared.

As we did earlier for offers, we tested the robustness of these results to modifications in our income and wealth variables. First, we ran separate models for CS and LNCS with our wealth and income revalued by local prices in batteries, rice, sugar, cooking oil, and salt. We found no qualitative differences from tables 4.9 and 4.10. When CS is used, its coefficient hovers close to 1.04. When LNCS is used, its coefficient hovers close to an odds ratio of 2.48. Second, we also explored the effects of standardizing our income and wealth variables. When CS is used, its coefficient varies from an odds ratio of 1.04 to 1.05. When LNCS is used, its coefficient varies from an odds ratio of 2.47 to 2.56. All coefficients on CS and LNCS in these checks are highly significant at conventional levels across the board.

In our sample, local community sizes are highly correlated with overall ethnic group size. In a few cases, the boundaries of the ethnic group are somewhat unclear, so several different demarcations could be drawn, making it somewhat difficult to nail down a precise value for this relationship. However, using a sensible set of demarcations, we have estimated the correlation at 0.97 (Marlowe et al. 2008). This relationship makes good sense from our theoretical perspective: those populations with more third-party punishment can stabilize more fairness and cooperation in larger populations, leading to larger, more stable ethnic groups and success in competition with other ethnic groups (Kelly 1985; Sahlins 1961). Since we get similar findings whether we use ethnic group size or local community size, these analyses cannot tease out which is the best predictor. From our theoretical perspective, these are causally interconnected in any case. Such findings are consistent with recent work suggesting that cultural group selection, driven by differences in political complexity, can help explain the size and diversity of languages globally (Currie and Mace 2009).

Minimum Acceptable Offers in the Ultimatum Game

Willingness to engage in second-party punishment, as measured by rejecting in the ultimatum game, is also positively related to both CS and LNCS. Model 1 in tables 4.11 and 4.12 shows our baseline models. As with the MinAO in the TPG, the coefficients on both LNCS and CS are large and highly significant. Individuals from larger communities tend to punish more. The coefficient on MI is also a potent predictor in both tables 4.11 and 4.12, as it is in table 4.10. Figures 4.8 and 4.9 illustrate the magnitude of these coefficients. The coefficient on MI indicates that, ceteris paribus, individuals from more market-integrated communities punish less. In contrast to the analysis of the TPG MinAO here, income and household size are also significant at conventional levels in some models, with higher-income individuals generally punishing more and those from larger households punishing less.

TABLE 4.11 *Ordered Logistic Regressions for the Minimum Acceptable Offer in the Ultimatum Game, Using Community Size*

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Community size (100 people)	1.077***	1.116***	1.076***	1.075***	1.076***	1.061***	
	(0.0142)	(0.0149)	(0.0133)	(0.0135)	(0.0136)	(0.0131)	
Market integration	0.979***	0.963***	0.979***	0.979***	0.980***	0.985***	0.998
	(0.00566)	(0.00733)	(0.00544)	(0.00549)	(0.00529)	(0.00479)	(0.00295)
Household size	0.924**	0.981	0.924**	0.929**	0.925**	0.921***	0.956*
	(0.0294)	(0.0267)	(0.0294)	(0.0288)	(0.0286)	(0.0274)	(0.0251)
Income	1.054***	0.968***	1.054***	1.055***	1.056***	1.038***	1.031**
(U.S.$1,000)	(0.0181)	(0.00959)	(0.0177)	(0.0177)	(0.0175)	(0.0140)	(0.0125)
Wealth	0.998	0.998***	0.998	0.998	0.998		
(U.S.$1,000)	(0.00216)	(0.000395)	(0.00215)	(0.00216)	(0.00215)		
Age	1.006	1.004	1.007	1.007			
(years)	(0.00941)	(0.0130)	(0.00887)	(0.00882)			
World religion	1.143	0.843	1.138				
	(0.340)	(0.214)	(0.327)				
Sex	0.993	1.034					
(Female = 1)	(0.239)	(0.252)					
Education	0.972	1.031					
(standardized by population)	(0.130)	(0.155)					
Africa		0.0186***					
		(0.0117)					
South America		0.0134***					
		(4.13e-07)					
Oceania		0.0340***					
		(0.0401)					
Observations	272	272	272	273	293	324	354
Pseudo-R-squared	0.0315	0.0743	0.0314	0.0310	0.0289	0.0221	0.00744

Source: Project data.

Notes: Model 2 uses clustered robust standard errors (clustering on site); other models use robust standard errors; standard errors are in parentheses below the coefficient. Education has been standardized to a mean of zero and standard deviation of one within each population. Coefficients are reported as odds ratios.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$

TABLE 4.12 *Ordered Logistic Regressions for the Minimum Acceptable Offer in the Ultimatum Game, Using LNCS*

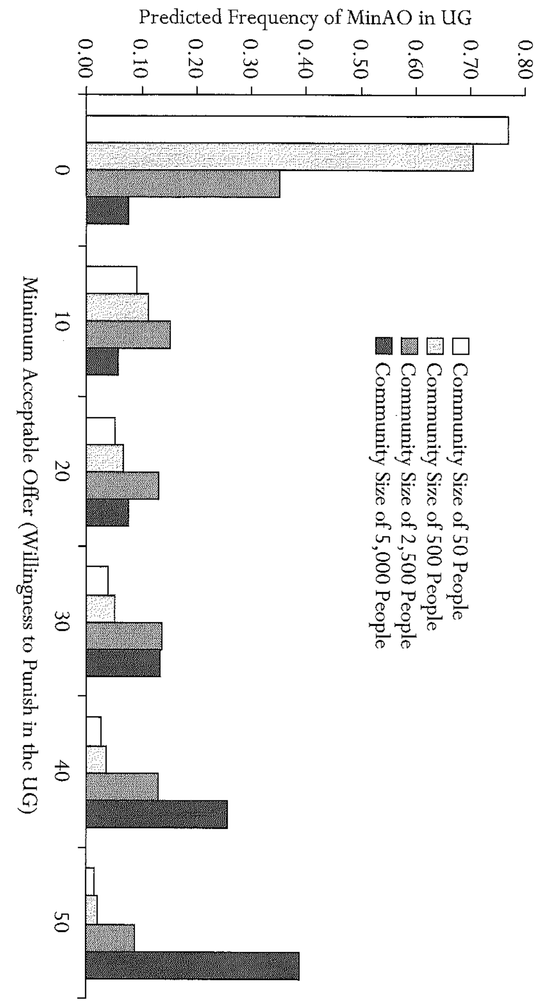
Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
LNCS	1.856***	3.116***	1.858***	1.759***	1.684***	1.599***	1.764***
	(0.294)	(0.946)	(0.288)	(0.267)	(0.241)	(0.207)	(0.161)
Market integration	0.974***	0.947***	0.974***	0.978***	0.979***	0.981***	0.976***
	(0.00781)	(0.0163)	(0.00758)	(0.00705)	(0.00686)	(0.00625)	(0.00504)
Household size	0.925**	0.986	0.924**	0.926**	0.923**	0.919***	
	(0.0297)	(0.0243)	(0.0297)	(0.0291)	(0.0289)	(0.0279)	
Income	1.058***	0.971***	1.058***	1.056***	1.054***	1.040***	1.055***
(U.S.$1,000)	(0.0194)	(0.00962)	(0.0190)	(0.0182)	(0.0175)	(0.0142)	(0.0158)
Wealth	0.998	0.998***	0.998	0.998	0.998		
(U.S.$1,000)	(0.00216)	(0.000326)	(0.00215)	(0.00212)	(0.00212)		
Age	1.007	1.003	1.007				
(years)	(0.00949)	(0.0127)	(0.00892)				
Education	0.998	1.066					
(standardized by population)	(0.132)	(0.162)					
World religion	0.741	0.355***	0.738	0.744			
	(0.244)	(0.0903)	(0.234)	(0.237)			
Sex	0.986	1.079					
(Female = 1)	(0.235)	(0.260)					
Africa		0.0116***					
		(0.0119)					
South America		0.00657***					
		(0.00995)					
Oceania		0.0265***					
		(0.0356)					
Observations	272	272	272	292	293	324	356
Pseudo-R-squared	0.0299	0.0832	0.0299	0.0258	0.0248	0.0207	0.0343

Source: Project data.

Notes: Coefficients are reported as odds ratios. Model 2 uses clustered robust standard errors (clustering on site); other models use robust standard errors; standard errors are in parentheses below the coefficient. Education has been standardized to a mean of zero and standard deviation of one within each population.

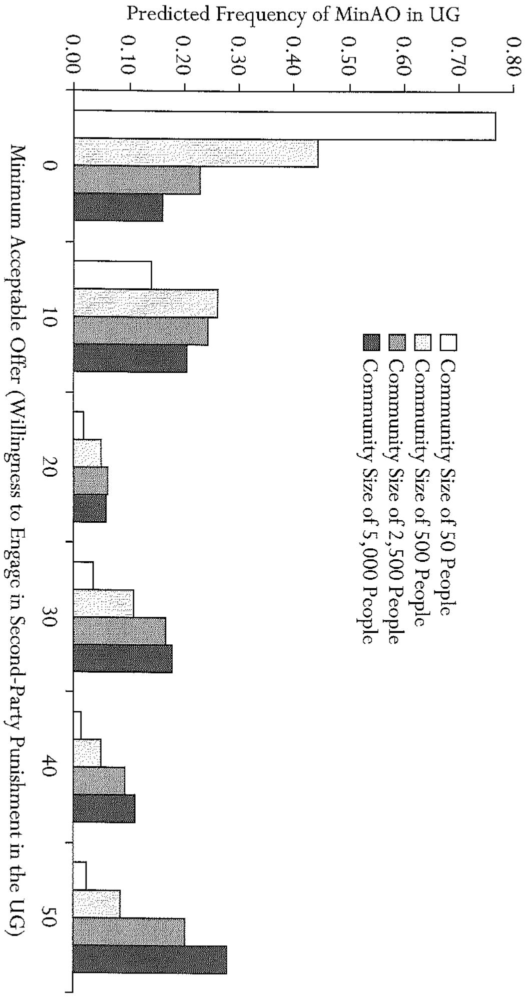
$*p < 0.1$; $**p < 0.05$; $***p < 0.01$

FIGURE 4.8 *Community Size Predicting Ultimatum Game Minimum Acceptable Offers*



Source: Project data.
Note: The coefficients used to create the plots are from an ordered logistic regression containing the eight other variables.

FIGURE 4.9 *LNCS Predicting Ultimatum Game Minimum Acceptable Offers*



Source: Project data.
Note: The coefficients used to create the plots are from an ordered logistic regression containing the eight other variables.

To check for the effects of cultural phylogeny and the use of clustered robust standard errors, compare models 1 and 2. When continental controls are added and clustered robust errors are used, the odds ratio for both CS (table 4.11) and LNCS (table 4.12) increases (a larger effect) and remains highly significant. The odds ratio for MI also withstands these checks in both tables. In contrast, the odds ratio for income drops below one, indicating a change in direction, while the odds ratio for wealth becomes significant because its standard error shrinks. For household size, the odds ratio increases toward one and becomes nonsignificant.

Models 3 to 7 in tables 4.11 and 4.12 examine the effect of alternative specifications for CS and LNCS. The odds ratios for both CS and LNCS remain large and highly significant across these models. To aid interpretation, the estimated association between CS (when LNCS is used) and the UG MinAO is presented graphically in figure 4.8. The effects of MI and household size are also robust across these specifications.

We also explored how dropping CS or LNCS, respectively, from these models influences what happens to the coefficients on market integration, income, and household size. The result is that the odds ratios for both MI and HS move toward one and become nonsignificant. The odds ratio for income also moves toward one, but remains significant. These analyses show that the significant negative relationship for MI and HS depends on keeping CS or LNCS in the regression (and only on this). In contrast, if CS or LNCS is the only predictor in the model, both remain highly significant, though the odds ratio for their coefficients does decrease.

Here, too, we checked the robustness of these findings by using modifications of our income and wealth control variables. First, in models replacing these variables with income and wealth variables derived from the prices of local consumables at the site, we found no qualitative differences from the results in tables 4.11 and 4.12. When CS is used, its coefficient varies from an odds ratio of 1.07 to 1.11. When LNCS is used, its coefficient varies from an odds ratio of 1.75 to 2.55. Second, we also explored the effects of standardizing our income and wealth variables. When CS is used, its coefficient varies from an odds ratio of 1.05 to 1.06. When LNCS is used, its coefficient varies from an odds ratio of 1.54 to 1.57. Again, these results vary little from what has already been observed, except that in these analyses income is not an important predictor of MinAO. All coefficients on CS and LNCS are highly significant across the board in both of these robustness checks.

Overall, both CS and LNCS are robust and potent predictors of MinAO in both the third-party punishment and ultimatum games. Larger communities punish more, controlling for our eight other variables. Both measures of pseudo-R-squared and AIC indicate that the LNCS size fits better than CS alone. Market integration predicts less punishment in most of our analyses, though never when CS is used to predict MinAO in the TPG. MI's effect requires that CS or LNCS be in the equation. If CS or LNCS is dropped, MI's effect dramatically decreases or reverses direction. Household size and wealth sometimes emerge as important predictors of the MinAO in the UG. The direction of the effect is such that more wealth or bigger households lead to lower MinAOs in the UG. Income also generally shows significant effects for the MinAO in the UG, with more income predicting more punishment (though income predicts less punishment when continental controls are included).

DISCUSSION

Here we review each of our major findings and consider its implications. Then we discuss our theoretical interpretations and consider alternative views of our findings. We close by addressing common concerns about the interpretation of experimental findings. Taking our four major findings in turn, we summarize and interpret each set of results and then consider its implications for understanding human sociality and the emergence of prosocial norms and institutions.

Summary of Major Findings

1. Fairness and Punishment Show Both Substantial Variability and Reliable Patterns Across Diverse Populations Replicating and extending our previous work with the ultimatum game (Henrich, . . . and Gintis 2004), we find substantially more variability across our diverse samples than is typically observed among people from industrialized societies. In all three experiments, mean offers and the standard deviations in offers vary substantially across our populations, especially when compared to the variation observed in subject populations from Western industrialized societies. With regard to people's willingness to engage in costly punishment, we found variation across populations not only in their willingness to punish *low* offers but also in their willingness to punish UG offers that were "too large." This phenomenon of hyper-fair rejections confirms initial observations made of the Au and Gnau data in our previous project (Tracer 2003, 2004) and is consistent with subsequent work in other societies (Bellemare et al. 2008; Güth et al. 2003; Wallace et al. 2007), including work done in Russia and China (Bahry and Wilson 2006; Hennig-Schmidt et al. 2008). Although not visible in the UG among typical undergraduate participants in the United States and Europe, such behavior has been detected using bargaining games that permit more finely tuned punishment (Andreoni et al. 2003; Huck 1999).

The behavioral variation captured in our data has important implications. We think much of the variation across populations captures the presence of social norms, but whatever one's preferred theory of human social behavior, the observed variability must be explained. However, despite our previous efforts (Henrich et al. 2005a), much work in economics and psychology (see, for example, Sanfey 2007) continues to proceed as if the results from students are generalizable to the species. To the contrary, our experiments show that Westerners consistently occupy the extreme ends of the behavioral distributions, just as they do in many other aspects of psychology (Henrich, Heine, and Norenzayan 2010). This variability suggests that Western students—and people from industrialized societies more generally—should not be used to make inferences about "human behavior" or social motivations.[19]

Alongside this variation, however, are also quite robust patterns across populations. In all populations, the probability of second- or third-party punishment (in the UG and TPG, respectively) declines as offers increase from zero to half of the total stake. Offers of 50 percent were always the most acceptable offer, although fifty-fifty offers often tied with other offers. Mean and modal offers across societies ranged across all three experiments from 10 percent to 50 percent, with very few offers above 50 percent. Such robust patterns, while they may seem broad, dramatically reduce the state space of possible explanatory theories.

Some social scientists have often suggested that cultural variability is so immense that essentially anything goes—that people are infinitely malleable (Pinker 2002). This is not consistent with our findings, especially when the additional populations from our previous project are added (giving us twenty-four different populations). Along these lines, it is important to consider what we did *not* observe at the population level. First, we did not observe nonmonotonic changes in the likelihood of punishment for offers from 0 percent to 50 percent, or from 50 percent to 100 percent; that is, people from a particular society did not punish offers of 20 percent a lot, offers of 30 percent a little, and offers of 40 percent a lot. Second, no population had mean or modal offers above 60 percent. Mean offers for immensely different populations in all three games occupy only about one-quarter of the possible spectrum, replicating our findings in phase 1. It is not the case that "anything goes" cross-culturally.

Although broad patterns do exist across our populations, predictions based on the assumption of narrow economic self-interest (money maximization) fail in all populations studied and in all three experiments. Participants either engage in significant amounts of costly punishment, despite the one-shot nature of the experiments, or give too much, given the likelihood of punishment. This remains true even if standard versions of risk aversion are considered. Most importantly, this model fails *in different ways in different places,* suggesting that behavioral models of decisionmaking rooted in empirical work from industrialized societies are rather limited in their generalizability.

This conclusion should not be taken as an indictment of game theory in general, or of the use of utility models (Henrich et al. 2005b). We enthusiastically support the use of game theory, both evolutionary and classical versions, and many models already exist that broaden narrow assumptions (Bolton and Ockenfels 1999; Charness and Rabin 2002; Fehr and Schmidt 1999). For example, our team has developed a one-parameter model based on inequity aversion that, once calibrated to each population, performs remarkably well in explaining the broad patterns observed (Barr et al. 2009). Thus, the problem highlighted here is not so much with the common assumption of purely self-regarding motivations, but with a failure to endogenize preferences in a manner that recognizes the effects of social norms and institutions on the internalization of motivations and formation of beliefs (Bowles 1998).

2. Fairness Increases with Market Integration Data from all three games demonstrate that fairness (making more equal offers) in transactions with anonymous partners is robustly correlated with increasing market integration.

For UG offers, this finding largely replicates the findings from phase 1 of the project (Henrich et al. 2005a). The DG findings represent a substantial extension of this finding, since offer decisions in the DG do not face any threat of punishment. These DG results combined with our analyses of the UG and TPG offers controlling for the local mean MinAO (threat of punishment) support the idea that this variation is at least partially rooted in internalized motivations.

3. Fairness Increases with an Individual's Participation in a World Religion In the DG and UG, participation in a world religion (versus a local traditional religion) is associated with a six- to ten-percentage-point increase in offers, controlling for other economic and demographic variables, including market integration. However, we should be less confident in the WR finding, both because of the rather small proportion of our sample with WR = 0 and because of weaknesses that emerge in some of the robustness checks, though the WR findings hold up quite well in the face of many of the additional analyses. Also, for world religion, our data include only Islam and Christianity (in a variety of forms), so we do not know if our conclusions would hold up for other widely subscribed supernatural belief systems like Judaism, Buddhism, and Hinduism.[20] We did not find a significant effect of WR on TPG offers, though as we conjectured earlier, the threat of weak third-party punishment may have crowded out prosocial motivations among adherents to world religions, which may also explain why mean offers drop in the TPG relative to the DG and UG.

Together, our findings for MI and WR show that moving from a fully subsistence-based population participating in a traditional religion to a market-based economy with a world religion predicts an increase in offers of roughly twenty points, thus explaining most of the difference we observe in offers across populations. As we emphasize later, this means that whatever theory one deploys to explain the variation we observe, it needs to account for the strong relationship of fairness with market integration and world religions. Our evolutionary approach to norms and institutions provides one hypothesis for this linkage.

4. Willingness to Engage in Costly Punishment Increases with Community Size People from larger communities punish more. This effect emerges in both of our measures of second- and third-party punishment and is robust to the inclusion of demographic and economic control variables, including market integration. The predicted mode for MinAO, our measure of willingness to punish, shifts from 0 percent in populations of 50 to 500 to 50 percent in populations of 5,000 (see figures 4.6 to 4.9). These findings dramatically extend Frank Marlowe's (2004) initial insights derived from work among Hadza foragers, as well as broader analyses of settlement size from phase 1 (Henrich, . . . and Gintis 2004). Whatever one's preferred theory about the variation in punishment, it should explain why punishment varies with community size.

In addition to the relationship between CS and MinAO, two other relationships are worth highlighting here as well. First, greater punishment in the TPG predicts not only TPG offers but mean offers in the UG and DG (Henrich et al. 2006). Second, larger communities with greater punishment and fairness (for example, in the DG) are strongly associated with larger ethnic groups. Although only correlational, these observations are consistent with the idea that groups with more stable prosocial norms can spread at the expense of groups with less prosocial norms, or that groups with norms stabilized by costly punishment (as opposed to other reputation-based mechanisms) can expand and sustain an ethnic identity more readily than groups lacking these kinds of norms. Both of these interpretations are consistent with the theoretical picture developed in chapter 2: norms can spread as a consequence of their group-beneficial effects.

Finally, it is worth noting that despite the fact that we studied multiple communities within many of our two dozen sites and conducted extensive analyses over two phases of our project, no robust individual-level demographic or economic predictors emerged in their data to account for variation within sites. On this count, other researchers have made much of a few correlations that emerge after testing dozens of predictor variables (Lamba and Mace 2011, 2013), which generally turn out to be inconsistent across games and not robust across societies (Henrich et al. 2012). Our strategy is to focus on only the most robust and theoretically well-grounded aspects of our overall findings.

Theoretical Interpretations

Here we consider several different theoretical interpretations of our findings. We begin by briefly recounting the theoretical interpretation presented at length in chapter 2, which we label the *social norms for complex, market-integrated societies hypothesis.* Then we deal with three versions of what we call the *anonymity hypotheses.* Finally, we consider the *genetic differences hypothesis,* which explores whether recent work in behavioral genetics using experimental tools like the UG and DG suggests that the variation we observe arises from genetic differences among populations.

In presenting and clarifying these interpretations, it is important to distinguish motivations (or preferences) from beliefs. *Motivations* are the internalized states or the anticipation of such states—such as goals, wants, drives, desires, or preferences—that aggregate to propel behavior. *Beliefs* are mental representations or expectations about the state of the world or the likelihood of various future states. Beliefs map actions onto outcomes. Motivations and beliefs can be context-dependent, culturally transmittable, and non-independent (linked). To give a simple example, suppose John likes pizza but really wants to live a long time (he is motivated to eat pizza and to have a long life). However, because he believes that pizza is bad for his longevity (a belief about pizza-eating), he tries not to eat very much pizza. If John comes to believe pizza is actually good for his longevity, he will eat much more pizza. Theories of human behavior need not make this distinction, but since at least some of the theories dealt with here do make it, we need to keep this in mind.[21]

Social Norms for Complex, Market-Integrated Societies In chapter 2 we argue that understanding human social behavior and psychology requires recognizing both (1) that humans acquire, as a consequence of growing up in particular places, interrelated sets of beliefs, expectations, and internalized motivations, and (2) that competition among social groups, religions, and institutions has sculpted these group-beneficial social norms over the course of cultural evolution (Chudek and Henrich 2010). This recognition allows us to begin to construct an understanding of the formation of large-scale cooperative societies in which strangers, or those not engaged in long-term, durable relationships, can increasingly engage in reliable, mutually beneficial transactions and cooperative enterprises. Since the origins of agriculture, large and complex societies are likely to have prospered and spread to the degree that their norms and institutions effectively sustain successful interaction in ever-widening socioeconomic spheres.

Norms that enhance fairness and trust among strangers are likely to be causally interconnected with the diffusion of several kinds of institutions. Here we focus on: the expansion of both the breadth and intensity of market exchanges (Bowles 1998; Ensminger 1992; Smith 1759/2000) and the spread of universal religions with high moralizing gods. At its most efficient, market exchange requires trust, fairness, and cooperation among individuals engaged in infrequent or anonymous interactions. The greater the shared set of motivations and expectations related to trust, fairness, and cooperation among interactants, the lower the transaction costs, the greater the frequency of beneficial transactions, and the higher the long-term rewards. Although reliable exchanges among strangers are now commonplace, studies of nonhuman primates and small-scale societies suggest that during most of our evolutionary history transactions beyond the local group, and certainly beyond the ethnolinguistic unit, were fraught with danger, mistrust, and exploitation (Fehr and Henrich 2003). Thus, "market norms" may have evolved as part of this overall process of societal evolution to facilitate and govern mutually beneficial exchanges in contexts where established social relationships (for example, those based on kinship, reciprocity, or status) were insufficient.

At the same time, religious institutions, beliefs, and rituals may have coevolved with the norms that support complex societies. Competition among societies may have favored the spread of potent moralizing gods, along with the institutional and ritual machinery for instilling faith. These emerging religious systems may have helped incentivize prosocial behavior toward coreligionists (and the exploitation of non-coreligionists) using a range of supernaturally supplied rewards and punishments (Ensminger 1997; Norenzayan and Shariff 2008; Shariff et al. 2010), as well as rituals that built group solidarity (Henrich 2009; Sosis and Alcorta 2003). Thus, in contrast to the religions that probably dominated most of our evolutionary history, religions such as Christianity and Islam may be unusual in providing a moralizing god with omniscience and ample powers to reward and punish (Atran and Henrich 2010).

This process had been rolling along for thousands of years, and suddenly our team arrives on the scene with our suite of experiments designed to tap the norms that govern impersonal exchange. The kinds of cues that spark these norms involve money, the primary medium in such exchanges, and a lack of relevant information about one's relationship to the other party (anonymity).

This approach takes seriously the idea that norm-learning could result in internalized motivations both to adhere to—or fulfill—a normative expectation (endogenous preferences) and to think poorly of those individuals who fail to follow the norm. Consider this toy example: Suppose a learner grows up in a place in which a certain social behavior is expected in a certain context. Most people in our learner's world typically adhere to the norm, and those who occasionally do not suffer reputational damage or punishment. As a consequence of growing up in this environment, our learner internalizes a motivation to adhere to the social norm. Then our learner enters an experiment that taps the norm he has internalized. This occurs even though, in this experimental context, the learner fully believes that no one will ever know what he did and no reputational damage can ensue. In some sense, the learner's prosocial behavior is related to growing up in a society in which reputation and punishment matter, but not because he mistakenly believes reputation matters in the experimental context. Another learner from a society that does not damage the reputations of those who fail to follow this particular social norm might have no guide in the experiment, except for self-interest and uncertainty (and perhaps empathetic altruism).

Anonymity Hypotheses Some researchers have proposed that the prosociality observed in behavioral experiments results from purely selfish motivations or genetically evolved mechanisms combined with an uncertainty about the anonymity in the experiments (Baumard, Boyer, and Sperber 2010; Delton et al. 2010). There are at least three versions of this idea. The first version proposes that individuals' beliefs regarding the reality of the anonymity in our experiments may have been influenced either by their experience in their daily lives or by calibrations to the local social environment (based on "cues") that reflect some ancestral world (and an associated evolved module for figuring out repeated interactions).[22] To clarify, using our earlier toy example, suppose a learner grows up in a society in which a certain prosocial behavior is sustained through reputation. But this fully self-interested learner does not internalize motivations; he only learns the rules and expectations of his group. He behaves prosocially *only* because of his assessment of the reputational effects or punishment. Then this learner enters our experiment. If he believes that the experiment is anonymous (that is, he perceives accurately), he will behave in a purely selfish manner, consistent with economic models that assume pure self-interest. However, since the experiment is a novel situation, he might bring beliefs and expectations derived from daily life into his game decision. By applying inaccurate beliefs about the effects of his play on his own future material payoffs, this learner ends up behaving more prosocially than predicted by models based on narrow self-interest. A parallel toy example can be provided for an evolved-module executor who is impervious to cultural learning.

Applying this reasoning predicts that people from societies with more actual ephemeral interactions and real anonymity (Western societies, for example) ought to be the best able to handle and understand the anonymity and behave purely selfishly. In these larger societies, people actually do have lots of interactions with strangers and in contexts lacking expectations of future interactions. If people's experience-derived beliefs or cue-calibrated modular expectations about anonymity are in fact driving prosociality in these experiments, we would expect the smaller-scale societies to be *more* prosocial in the experiments, not less (Henrich et al. 2005b).

Our findings reveal precisely the opposite pattern from that suggested by these lines of hypothesizing. Empirically, the smaller-scale, least anonymous, and most face-to-face societies are generally less prosocial (New Guinea is the exception), while larger, more complex, and more anonymous societies are more prosocial. That is, the societies that live more like our ancestors in those small groups are actually less prosocial in these experiments. Results from both phases of our project confirm that market integration is highly correlated with offers, even in the DG and even when the local threat of punishment is statistically controlled for in the UG and TPG. Previously, we showed that anonymity in everyday exchanges (anonymous roles like "cashier"), market integration, and societal complexity are all highly correlated (Henrich, . . . and Gintis 2004). Moreover, if we treat CS or LNCS as a proxy for anonymity and face-to-face interactions in our populations, then our work shows that community size is not negatively associated with prosocial behavior, as predicted by this anonymity hypothesis. For offers, CS is never significant. However, if MI is removed from the regression, CS always has a *positive* coefficient, and it sometimes approaches significance. Overall, our findings do not support this version of the anonymity hypothesis, and this version cannot explain the empirical patterns we do find for MI, CS, and WR.

A second version of the anonymity hypothesis proposes that individuals from smaller communities may have avoided punishing because they feared that their actions might be found out and interpreted as an aggressive move against player 1. The idea is that the likelihood of being found out increases in smaller communities, thus yielding the observed relationship between community size and MinAO. When seen in the light of the offer data, we think this interpretation is largely consistent with our theory. To see this, first, recall that CS was not an important predictor of offers in any of the experiments. This means that somehow this anonymity concern was a factor only on the punishment side and did not influence offer decisions. Individuals from many smaller communities (though not all), such as in Fiji (chapter 9), entered the experiment and gave fairly, but were rarely willing to punish. Following our theoretical approach, this probably reflects local prosocial norms stabilized by reputational mechanisms, such as results from being dropped from dyadic helping networks (Panchanathan and Boyd 2004), and does not involve costly punishment. Costly punishment is, in fact, often frowned upon in these small communities because it can generate cycles of reprisals. Individuals who live in communities with norms maintained by such reputational mechanisms ought to show both fair behavior and an unwillingness to engage in costly punishment. As is the case for many norms (Henrich and Henrich 2007), habitually performed and frequently observed behaviors are partially internalized so that actions in daily life reflect some combination of internalized motivations (not to punish, in this case) and beliefs about, for example, the consequences of punishing. Thus, in our view, the unwillingness to punish in our experiments reflects the rules of daily life and arises from some combination of beliefs and motivations imported into the experiment from routine practices and interactions.[23]

Relevant to assessing this anonymity hypothesis, we performed these experiments among university students in the United States, as mentioned earlier (chapter 9). In an effort to approximate the small communities from which we drew our subjects, we randomly selected students from the same small freshman dormitory. If the size of the pool from which players are drawn influences their assessment of future potential anonymity and thereby causes them to punish less, then these students should have punished less than students in typical experiments, and potentially a lot less, since the pool sizes varied from one freshman dormitory to the entire university. Although a carefully controlled comparison is not possible, the results for punishment among these student subjects do not appear different from those observed in student samples drawn from larger, more anonymous populations. This finding suggests that merely manipulating the size of the pool of frequently interacting potential subjects is not driving the impact of the community size variable.

A third version of the anonymity hypothesis is that prosocial behavior in our experiments results from efforts to manage one's reputation with the experimenter (Levitt and List 2007). Inspiration for this effect comes from work using the dictator game involving experimenter-blind treatments (Cherry, Frykblom, and Shogren 2002; Hoffman et al. 1994; List and Cherry 2008) and work manipulating anonymity cues (Haley and Fessler 2005). Most experiments are single-blind, meaning that a player's behavior will never be known by the other players (and everyone is told this). In double-blind experiments, neither the other players nor the experimenters can figure out what a specific player did. Among students, protocols that make the double-blind transparent (giving maximum confidence in anonymity vis-à-vis the experimenter) cause dictator game offers to decline. The approach could explain our variation if different populations varied in how much they wanted to impress the experimenter, and in a manner that happened to strongly correlate with MI, WR, and CS.

To put such experiments in a broader context, it is important to recognize two things. First, most of these effects have been limited to *students in dictator games*. Both students (Henrich, Heine, and Norenzayan 2010) and DGs are notorious for being relatively easy to manipulate using framing effects (Fehr and Schneider 2010). Second, a norm-learning approach offers a ready interpretation to these findings. Some student players find that the DG is ambiguous in terms of which norms apply, since it lacks the strategic conflict of other games. This ambiguity causes them to seize on otherwise minor framing effects to figure out which norms to apply to the game. The double-blind procedures provide strong cues—the game equivalent of neon signs—proclaiming: "No one is looking, and the experimenter wants you to behave self-interestedly; don't feel guilty about it." Psychologists call this well-known phenomenon an "experimenter demand effect." Attempting to remove concerns about the impacts of experimenter knowledge may actually dramatically raise experimenter demand effects by signaling to the subject how he or she is "supposed" to behave (cuing the subject as to what the norms are).[24]

Our project sought to address these concerns by running double-blind experiments among nonstudent subjects in four populations. Among the Orma and Samburu, we found no measurable impact of the double-blind treatment (chapter 5). Moreover, in the United States the standard finding using the DG with students, which reveals a large drop in mean offers, does not readily extend to nonstudent populations. Among U.S. nonstudent adults, we found a substantially diminished effect in the double-blind treatment compared to typical student findings. Among the Sanquianga in Colombia (chapter 16), we did find a strong impact in a quasi-double-blind treatment. Across our four comparative experiments, the two with no detectable effects for the double-blind treatments were conducted by non-coethnics of the participants. The U.S. experiment and the Sanquianga experiment were done, not by locals, but by conationals. This may suggest that people do not care what out-group others think of them. If that is the case, it reinforces our results, since, with a couple of exceptions, our experiments were administered by out-group non-coethnics. This could also suggest that the large effects of double-blind procedures are an artifact of relying on Western student subjects with Western experimenters whose approval they seek.

Our previous work has also addressed the issue of experimenter anonymity by reasoning that if the variation we observe among populations results from different subject populations caring to differing degrees about what the experimenter might think, or believing different things about what the experimenter might do with the information, or possessing different beliefs about what the experimenter might want, we can control for this by including the number of months each experimenter had spent in the community prior to doing the experiments in our regression analyses on offers. The coefficient on this predictor was close to zero, and nonsignificant. Its inclusion had no influence on our findings (Henrich, . . . and Gintis 2004).

Finally, this anonymity hypothesis suffers because it cannot explain the strong relationship we find for MI, WR, and CS in our data.

Genetic Differences Across Populations

Another hypothesis is that some of the variation in game play that we observe across populations arises from genetic differences across populations. This idea receives some support from recent work showing how cultural evolution can shape—and indeed, has shaped—evolutionary processes to yield patterns of genetic variation in humans

(Laland, Odling-Smee, and Myles 2010; Richerson, Boyd, and Henrich 2010). Moreover, recent work combining tools from behavioral genetics with those of experimental economics has suggested that variation in behavioral game measures is partially explained by genetic variation. Using ultimatum game MinAOs, researchers have estimated a heritability of 0.42 and found no influence of common family environment among a large sample of Swedish twins (Wallace et al. 2007). Similar results obtain for the trust game, which measures trust and trustworthiness in anonymous interactions (Cesarini et al. 2008). Finally, within one population, there is work showing that high DG offers are associated with the longer microsatellite repeats for the arginine vasopressin 1a receptor (Knafo et al. 2008), a gene associated with bonding and affiliation in other species.

We think this evidence strongly indicates that some of the variation in game behavior *within* Western populations probably results from *individual* genetic differences, especially given the heritability of so many other aspects of behavior and personality. However, we emphasize the pitfalls of logically extending within-group findings on heritability to understanding differences between populations, thereby committing the ecological fallacy. To understand this, let us start with variation in height. Within modern, industrialized, especially Western populations, height is highly genetically heritable. Does this mean that variation in height among populations is likely to be the result of genetic differences among populations? No. Height is greatly influenced by a variety of factors (childhood nutrition and disease, in particular), most of which are relatively constant in Western populations, at least in the middle and upper classes. This relative uniformity within the West maximizes the role of genetic variation in determining total phenotypic differences, because it reduces the potential role of environmental variation. This can be seen in the rapid decrease in, or disappearance of, height difference in immigrant populations (compared to the average of the new host population), as well as in the much lower estimates of heritability in places like China and Africa. It is quite possible that while most of the variation in human height within Western populations is genetic, much of the variation among other populations in average height is environmental. Similarly, IQ is highly heritable among middle- and upper-class Americans, but not heritable much at all among those from the lowest socioeconomic strata in the United States (Nisbett 2009; Turkheimer et al. 2003). This is probably because middle- and upper-class Americans have squeezed out nearly all the social, cultural, and environmental factors that influence IQ, leaving only the genetic variation. Where these factors have not been squeezed out, genes are relatively unimportant.

The variation we observe within genetically well-mixed ethnolinguistic populations speaks against any simple story about genetic variation in our experimental findings. With the Orma (Ensminger 2004) and Tsimane' (Gurven 2004), we found variation at the level of communities, with different communities revealing quite different behavior in our experiments. (We interpret this as meaning different norms have become locally stable.) In the case of the Orma, these patterns are consistent with historically recent differences in degrees of market integration that accompany sedentarization; more subsistence-oriented nomadic herders were less fair-minded than nearby settled, market-dependent Orma. We also found differences among Quichua and Achuar inhabitants of the same community in Ecuador (Patton 2004), and between Mongolian and Kazakh participants in Mongolia (Gil-White 2004). This kind of group-level variation is quite unlikely to be accounted for by genetic variation.

Norms, like nutrition in the case of height, are often more uniform within populations, but vary dramatically between populations. Thus, we also suspect that the heritability estimates for behavioral games, like IQ and height, will vary greatly among populations depending on the nature and strength of the local norms—a norm-heritability interaction.

It is also important to recognize that single-gene correlation studies done in a single population have not stood up robustly in replications in other populations. For example, H. S. Kim and colleagues (2010) found that a particular serotonin receptor polymorphism (5-HTR1A) was associated with *increased* attention to focal objects among Americans, but that the same allele was associated with *decreased* attention to focal objects among Koreans—same gene, different effects in different populations. The potential for complex gene-environment interactions makes single-gene correlations from a single population necessarily preliminary.

Genetic differences between populations or groups would most likely account for the behavioral patterns we observe if they arose in response to stable differences in the culturally evolved social norms and institutions (formal and informal) found in different societies. Norms and institutions, in creating stable regularities in the local social environment, can theoretically produce conditions for natural selection to act on genes that make individuals better adapted to those particular norms and institutions (Henrich and Boyd 2001; Laland et al. 2010; McElreath, Boyd, and Richerson 2003; Richerson et al. 2010). This is an intriguing and provocative possibility, but there is no evidence at this point supporting a suspicion that such a culture-gene coevolutionary process has occurred.

What Do Experiments Measure and Do They Tell Us Anything About the Real World?

A variety of researchers have criticized the use of laboratory game experiments, arguing that the lack of real-world context makes them difficult to interpret and that results from these experiments are not associated with any real-world outcomes (Baumard and Sperber 2010; Levitt and List 2007; Rai and Fiske 2010).[25] We think that both of these important concerns can be addressed within the context of our theoretical approach, which explicitly theorizes about what the games measure, based on their salient contextual cues of cash and anonymity ("framing"), and how these link to real-world outcomes and measures.

Experimentalists in psychology have long recognized the importance of framing (contextual cues), though they have frequently lacked any general theory for explaining it. Sometimes experimentalists have attempted to sidestep this issue by maintaining that their experimental games are "frame-free," with the results measuring some dispositional social preferences. As chapter 2 makes clear, our approach to norms proposes that people acquire norms by making inferences about others' behavior in social contexts. Research among children shows that inferences from observed interactions result in quite context-specific behavioral tendencies, not dispositions. Children infer the rules of the game and quite energetically apply them to deviants (Rakoczy, Warneken, and Tomasello 2008; Rakoczy et al. 2010). But they do not apply them to different situations. For example, children can acquire altruistic preferences via imitation, but they do not readily extend that altruism to novel contexts (summarized in Henrich and Henrich 2007, ch. 2). Similarly, adults who would agree with admonitions against stealing or lying will also admit that stealing and lying are okay in some circumstances. In light of this, we interpret our experiments as context-specific measures of the presence of norms and motivations about how to treat someone for whom one has little information, in transactions involving cash. Based on our theory about what kind of norms the experiments tap, we arrived at our predictions about market integration, religion, and community size. We are exploiting the frame inherent in most economic games in testing our theory.

As mentioned, critics of the use of behavioral games have argued that many studies fail to show a relationship between game play and real-world behavior (Baumard and Sperber 2010; Levitt and List 2007; Rai and Fiske 2010). The problem with much of this criticism is that no theory is brought to bear to specify the real-world phenomena with which game play should be correlated. For example, Tage Rai and Alan Fiske (2010) note that behavior in Michael Gurven and Jeffrey Winking's (2008) experiments, for example, is not correlated with beer-making or well-digging among the Tsimane'. Why should it be? Our theoretical approach to social norms predicts that, if game play does tap norms evolved for interacting with strangers or anonymous others, then these games ought to be associated with things like market integration, social scale (community size), economic success, and other features related to the operation of large-scale societies—features that capture those elements of impersonal interactions not governed by personal relationships. We show that, looking across diverse populations, market integration is indeed highly correlated with experimental measures of prosocial behavior in bargaining games. Similarly, antisocial punishment is inversely correlated with gross domestic product and predicted by national measures of the strength of the rule of law and measures of norms of civic cooperation (Herrmann et al. 2008). Within populations, trust game measures of trustworthiness predict repaying loans in a microfinance program in Peru (Karlan 2005), and they predict alumni donations (Baran, Sapienza, and Zingales 2010). Dictator game offers are correlated with both donations to Hurricane Katrina victims (Kam, Cranmer, and Fowler, n.d.) and political participation (Fowler and Kam 2007). Also using the DG, Abigail Barr and Andrew Zeitlin (2010) show that Ugandan teachers' time allocations to teaching are negatively correlated with their DG money allocations to parents of pupils at their schools. They go on to show that, using the same theoretical model as a guide, the correlation can be improved by taking into account differences in reference points, including social norms, and norm enforcement across schools and teachers. Overall, once properly theorized, economic games have been shown to correlate highly with several predictable and important real-world phenomena.

CONCLUSIONS: IMPLICATION FOR UNDERSTANDING THE OPERATION OF INSTITUTIONS

Our results show that prosocial behavior in situations involving cash and anonymity varies systematically across societies in patterns consistent with the emergence of social norms for governing social interaction among those not engaged in longer-term kin- or reciprocity-based relationships. These patterns of prosocial behavior, when combined with (1) findings showing that prosocial behavior (for example, costly punishment) activates neuronal reward systems, (2) developmental evidence showing the rather late emergence of this kind of behavior, and (3) research demonstrating the effectiveness of observational learning in transmitting both prosocial behavior and norms more generally, are consistent with the hypothesis that specialized norms have culturally evolved over thousands of years to facilitate successful interaction and exchange with individuals who would otherwise be outside of a reliable long-term social relationship. Growing up in a world with such norms, and the institutions and religious systems that buttress and bolster these norms, favors the internalization of context-specific prosocial motivations that facilitate exchange (with low transaction cost) and cooperation in large, relatively harmonious groups.

NOTES

1. Whether we have added ten or eleven new populations in phase 2 depends on whether one counts rural Missouri as a new population or as a replication of the U.S. control experiment we did among graduate students in Los Angeles during phase 1 (Henrich and Smith 2004).
2. In phase 1, our sample from Lamalera (Indonesia) posted a mean UG offer of 57 percent (Alvard 2004). Subsequent work using our protocol found a high mean UG offer of 61 percent among the Sukuma in Tanzania (Paciotti and Hadley 2003).
3. To make this calculation we dropped players who made more than one switch between rejecting and accepting offers that ranged from 0 percent to 50 percent—a complex pattern indicating that one number cannot capture an individual's preferences. For example, if a player rejected 0 and 10 percent, accepted 20 percent, and then rejected from 30 to 50 percent, the MinAO could be set at 20 percent. However, this is not very informative, since this person rejects not only offers below 20 percent but also offers between 30 and 50 percent. Empirically, few players were dropped because of this restriction. We were able to calculate MinAOs for 96 percent of responders. Of the eighteen individuals for whom we could not calculate an MinAO, five were Hadza, two were Yasawans, three were Dolgan or Nganasan, and eight were Sursurunga.

4. Mean offers in the UG around 25 percent have also been found among the Machiguenga (Henrich 2000) and Quichua (Patton 2004), while the Pimbwe revealed a mean offer of 15 percent (Paciotti and Hadley 2003).
5. Benedikt Herrmann, Christian Thöni, and Simon Gächter (2008) deployed a public goods game with punishment across a diverse swath of industrialized societies (including Russia) and found "antisocial punishment" (a willingness to punish the overly cooperative) in many of the non-Western populations, but found little among Western undergraduates. Antisocial punishment in public goods games may be related to hyper-fair punishment in the UG.
6. The best-fit estimates of *r* in student populations using risky decisionmaking experiments is 0.81 (Tversky and Kahneman 1992). More details on this approach can be found in McElreath and Camerer (2004).
7. Theoretically, the highest possible IMO in the TPG is only 20 percent. It is possible for offers of 30 percent to have exactly the same expected income and utility as offers of zero, but if even one person fails to punish at zero, or mistakenly punishes at 30 percent, offers of zero still maximize income and utility.
8. In small-scale societies, kinship systems have culturally evolved to provide an organizing framework within which all in-group social relationships are regulated by norms that appear to extend and exploit our evolved psychologies for kin and reciprocity altruism—see Henrich and Henrich (2007) and Alvard (2003, 2009). Sometimes through elaborate rituals, these systems frequently include ways to bring in those outside the kinship system—creating fictive kinship relations that bring prepackaged, established norms to bear on social relationships (Ensminger 2001). Often, interactions outside the systems are hostile and based on narrow self-interest and suspicion. With the possibility of larger-scale complex societies, we argue that norms regulating exchange beyond such kin- and reciprocity-based systems emerged and spread (see chapter 2 for our discussion of the mechanisms of that spread).
9. Injecting money into social interactions has distinct effects, at least among North Americans (Vohs, Mead, and Goode 2006, 2008). Giving someone money in exchange for something often signals a desire to avoid engaging in a longer-term nonmarket relationship (Heyman and Ariely 2004). Paying a date with cash, for example, after a satisfying evening suggests a different kind of interaction than providing exactly the same cash value in food, wine, and entertainment.
10. Technically, we predict more offers closer to fifty-fifty. However, since there were generally few offers greater than 50 percent, we used offers as the dependent variable instead of more complex formulations, such as the absolute value of 50 percent minus the offer (measured as a percentage of the stake). This formulation has its own problems, since the motivations favoring offers of 60 percent are probably not the same as those favoring offers of 40 percent, and we need not expect predictor variables to have the same symmetric effect on each side of 50 percent.
11. Note that for analyses using community size we drop the Accra sample. For those using wealth we had to drop both the Accra sample and the Dolgan/Nganasan sample owing to a lack of wealth data for these groups. For the same reason, the Gusii are dropped from analyses involving household size. Unless these variables are significant predictors in our baseline models, we then remove these predictor variables from our baseline analyses to show what happens when these samples are included (see chapter 3).
12. Versions of the supplementary analyses discussed here can be found in Henrich et al. (2010). However, since publishing that paper, we have discovered that there was a problem with the household size data for the Gusii (Chapter 12, this volume, available at: http://www.russellsage.org/Ensminger_Chapter12.pdf). This means that all the analyses discussed here deviate slightly from those found in Henrich et al. (2010) because the Gusii are now dropped from regressions using household size as a control variable. None of the substantive conclusions change.
13. Model 2 assumes *only* that our sites are statistically independent. We include this to address the concern that our participants might not represent fully statistically independent observations, since many participants were sampled from the same communities and populations. If true, this would mean that standard errors calculated assuming individuals are independently observed would be wrong. Although we think that this argument may misunderstand what statistical independence implies or requires, as our participants were alone in making their experimental decisions, we have taken this concern seriously and seek to address it in the analyses of each of our five game measures later in the chapter. We provide model 2 as a check on our baseline model (model 1), but we have not used clustered robust standard errors (clustering on site) in all our regressions because that would be excessively conservative.

14. Owing to findings in Bahry and Wilson (2006), we also ran models to look for any nonlinear effects of age, using age-squared, and found none.
15. Comparing UG and DG offers across all populations at the individual level, the Mann-Whitney nonparametric test obtains a *p*-value of 0.11. Thus, while not significant at conventional levels, it is probably the case that UG offers are higher than DG offers. When we run this test on each population, only the Sursurunga and Gusii show a significant difference at conventional levels. (The Maragoli reveal a *p*-value of 0.056.)
16. A person's DG offer may have set a reference point for the UG, given the temporal proximity of play in the two games. The effect of world religion could have been brought into the UG *via* the DG, though note that the coefficient on WR in the UG is generally larger than in the DG.
17. It is worth noting that all of the important results highlighted here can also be found using linear regression analyses that assume MinAO is a continuous, normally distributed variable.
18. If LNCS is dropped from the specification, the direction and effect of MI switches to positive while remaining significant. In fact, without LNCS in the model, the effect of MI changes direction and is significant across all of these specifications. MI and LNCS are correlated 0.6. The converse is not true: if LNCS is kept in the specification but MI is dropped, the effect of LNCS remains large (odds ratio about 2) and highly significant.
19. Methodologically, if we had used the standard method of eliciting minimum acceptable offers directly from responders in the ultimatum game, rather than our method of eliciting the full vector of decisions across all possible offers, we would not have observed the phenomenon of hyper-fair rejections. The standard strategy method design presupposes something about human preferences that turns out not to be very accurate for six of our fourteen populations. This is important because experimental designers sometimes use their own locale-dependent intuitions about what a sensible response is when they construct an experiment. In doing this, they can obscure variability.
20. Our theoretical approach to understanding religion proposes that religions that integrate beliefs in powerful moralizing gods who are willing to incentivize proper behavior with commitment-inducing rituals can help strengthen the acquisition and maintenance of prosocial norms among the faithful. What is theoretically relevant is not whether a system of religious belief is a world religion or is monotheistic, but instead whether it possesses varying degrees of these elements (Atran and Henrich 2010).
21. Neither beliefs nor motivations, in the technical sense meant here, should be confused with verbal expressions. The stuff people say may or may not reflect what they believe or want, and the part of the brain connected to the mouth may not even be aware of what a person believes or is motivated to do. So it is possible that people would be unable to express their beliefs and motivations to you, even if they wanted to.
22. This approach is called the *big mistake, or mismatch, hypothesis*. It proposes that humans do not have culturally evolved social norms, or at least that such norms do not explain game play. Instead, social behavior is governed (entirely) by a set of evolved psychological mechanisms that arose from the selective forces created by the action of reputation and reciprocity in small groups that have characterized much of human history (Burnham and Johnson 2005; Dawkins 2006; Hagen and Hammerstein 2006). An individual (not a learner, but a "module activator") enters the experiment and faces a social dilemma. He fails to fully interpret the anonymity of the situation—and thus the within-group fitness-maximizing behavior—because his social psychology misfires in this novel environment. There are two subvariants of this hypothesis that diverge at this point. This misfiring either activates other-regarding social motivations that propel prosocial behavior in small groups (like reciprocal dyads) or produces faulty beliefs about the anonymity of the experimental situation such that self-interest alone motivates prosocial behavior. Interestingly, proponents of this hypothesis often extend the logic to explain all larger-scale cooperation in modern societies—making the kind of larger cooperation we observe in the world today an out-of-equilibrium evolutionary "mistake" resulting from our long history in smaller-scale societies. Although this argument has superficial plausibility for some evolutionary researchers, it generally fails both theoretical and empirical tests (Cimino and Delton 2010; Chudek, Zhao, and Henrich 2013; Fehr and Henrich 2003; Henrich and Henrich 2007).
23. The fear that might deter a potential punisher from inflicting costs in the experiment can arise only if a certain community would judge this action as aggressive (and inappropriate). This depends entirely on local rules about appropriateness. For example, in some communities it is perfectly fine—encouraged even—to punch someone as hard as you can in the face, as long as you are in a boxing ring. In this light, one can just as easily run the anonymity hypothesis the other way. If individuals do not believe the anonymity, or believe it less in smaller communities, they should take the opportunity to demonstrate their commitment to fairness, their toughness, or their facility with "hard bargaining" to their fellow community members by dishing out larger punishments. Community size should negatively predict MinAOs. In short, this anonymity hypothesis depends on local norms that influence how "punishing" is understood and judged. Thus, it falls broadly within our explanatory framework based on norms.

24. Unconsciously priming Christians with "God" before playing a double-blind DG increases fairness. God is apparently way more important than the experimenter, and there is no experimental treatment that blinds him/her/it (Shariff and Norenzayan 2007).
25. Levitt and List (2007) also argue that experiments suffer from a lack of attention to self-selection of the individuals into experiments and to the nature and extent to which one's actions are scrutinized by the experimenter. The first concern does not apply to us, since we used random samples of adults from our communities with little attrition. The second has been dealt with in our description of our use of double-blind experiments and our efforts to control for the experimenters' field experience.

REFERENCES

Alvard, Michael. 2003. "Kinship, Lineage, and an Evolutionary Perspective on Cooperative Hunting Groups in Indonesia." *Human Nature: An Interdisciplinary Biosocial Perspective* 14(2): 129–63.

———. 2004. "The Ultimatum Game, Fairness, and Cooperation Among Big Game Hunters." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. Oxford: Oxford University Press.

———. 2009. "Kinship and Cooperation." *Human Nature: An Interdisciplinary Biosocial Perspective* 20(4): 394–416.

Andreoni, James, Marco Castillo, and Ragan Petrie. 2003. "What Do Bargainers' Preferences Look Like? Experiments with a Convex Ultimatum Game." In *American Economic Review* 93(3): 672–85.

Atran, Scott, and Joseph Henrich. 2010. "The Evolution of Religion: How Cognitive By-products, Adaptive Learning Heuristics, Ritual Displays, and Group Competition Generate Deep Commitments to Prosocial Religions." *Biological Theory* 5(1): 1–13.

Bahry, Donna L., and Rick K. Wilson. 2006. "Confusion or Fairness in the Field? Rejection in the Ultimatum Game Under the Strategy Method." *Journal of Economic Behavior and Organization* 60(1): 37–54.

Baran, Nicole, Paola Sapienza, and Luigi Zingales. 2010. "Can We Infer Social Preferences from the Lab? Evidence from the Trust Game." Working Paper 15654. Cambridge, Mass.: National Bureau of Economic Research.

Barr, Abigail, Chris Wallace, Joseph Henrich, Jean Ensminger, Clark Barrett, Alexander Bolyanatz, Juan-Camilo Cardenas, Michael Gurven, Edwins Laban Gwako, Carolyn K. Lesorogo, Frank W. Marlowe, Richard McElreath, David Tracer, and John Ziker. 2009. "Homo Aequalis: A Cross-Society Experimental Analysis of Three Bargaining Games." Department of Economics Discussion Paper 422. Oxford: Oxford University (February).

Barr, Abigail, and Andrew Zeitlin. 2010. "Dictator Games in the Lab and in Nature: External Validity Tested and Investigated in Ugandan Primary Schools." Working Paper 2010-11. Oxford: Oxford University, Centre for the Study of African Economies.

Baumard, Nicholas, Pascal Boyer, and Dan Sperber. 2010. "Evolution of Fairness: Cultural Variability." *Science* 329(5990): 388–89.

Baumard, Nicholas, and Dan Sperber. 2010. "Weird People, Yes, but Also Weird Experiments." *Behavioral and Brain Sciences* 33(2-3): 24–25.

Bellemare, Charles, and Sabine Kröger. 2007. "On Representative Social Capital." *European Economic Review* 51(1): 183–202.

Bellemare, Charles, Sabine Kröger, and Arthur van Soest. 2008. "Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities." *Econometrica* 76(4): 815–39.

Bolton, Gary, and Axel Ockenfels. 1999. "A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90(1): 166–94.

Bowles, Samuel. 1998. "Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions." *Journal of Economic Literature* 36(1): 75–111.

———. 2008. "Policies Designed for Self-interested Citizens May Undermine 'the Moral Sentiments': Evidence from Economic Experiments." *Science* 320(5883): 1605–9.

Burnham, Terence C., and Dominic D. Johnson. 2005. "The Biological and Evolutionary Logic of Human Cooperation." *Analyse und Kritik* 27: 113–35.

Camerer, Colin. 2003. *Behavior Game Theory: Experiments in Strategic Interaction*. Princeton, N.J.: Princeton University Press.

Camerer, Colin, and Ernst Fehr. 2004. "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. Oxford: Oxford University Press.

———. 2006. "When Does 'Economic Man' Dominate Social Behavior?" *Science* 311(5757): 47–52.

Cancho, Ramon Ferrer I., Ricard V. Solé, and Reinhard Köhler. 2004. "Patterns in Syntactic Dependency Networks." *Physical Review E* 69(5):051915.

Carpenter, Jeffrey, Stephen Burks, and Eric Verhoogen. 2005. "Comparing Students to Workers: The Effects of Social Framing on Behavior in Distribution Games." In *Field Experiments in Economics,* ed. Jeffrey Carpenter, Glenn W. Harrison, and John A. List. Greenwich, Conn.: JAI Press.

Carter, John R., and Michael D. Irons. 1991. "Are Economists Different, and If So, Why?" *Journal of Economic Perspectives* 5(2): 171–77.

Cesarini, David, Christopher T. Dawes, James H. Fowler, Magnus Johannesson, Paul Lichtenstein, and Björn Wallace. 2008. "Heritability of Cooperative Behavior in the Trust Game." *Proceedings of the National Academy of Sciences of the United States of America* 105(10): 3721–26.

Charness, Gary, and Matthew Rabin. 2002. "Social Preferences: Some Simple Tests of a New Model." *Quarterly Journal of Economics* 117(3): 817–69.

Cherry, Todd L., Peter Frykblom, and Jason F. Shogren. 2002. "Hardnose the Dictator." *American Economic Review* 92(4): 1218–21.

Chudek, Maciej, and Joseph Henrich. 2010. "Culture-Gene Coevolution, Norm-Psychology, and the Emergence of Human Prosociality." *Trends in Cognitive Sciences* 15(5): 218–26.

Chudek, Maciej, Wanying Zhao, and Joseph Henrich. 2013. "Culture-Gene Coevolution, Large-Scale Cooperation, and the Shaping of Human Social Psychology." In *Signaling, Commitment, and Emotion,* ed. Richard Joyce, Kim Sterelny, and Brett Calcott. Cambridge, Mass.: MIT Press.

Cimino, Aldo, and Andrew Delton. 2010. "On the Perception of Newcomers." *Human Nature* 21(2): 186–202.

Currie, Thomas E., and Ruth Mace. 2009. "Political Complexity Predicts the Spread of Ethnolinguistic Groups." *Proceedings of the National Academy of Sciences of the United States of America* 106(18): 7339–44.

Dawkins, Richard. 2006. *The God Delusion*. Boston: Houghton Mifflin.

Delton, Andrew W., Max M. Krasnow, Leda Cosmides, and John Tooby. 2010. "Evolution of Fairness: Rereading the Data." *Science* 329(5990): 389.

Ensminger, Jean. 1992. *Making a Market: The Institutional Transformation of an African Society.* Cambridge: Cambridge University Press.

———. 1997. "Transaction Costs and Islam: Explaining Conversion in Africa." *Journal of Institutional and Theoretical Economics (Zeitschrift Fur Die Gesamte Staatswissenschaft)* 153(1): 4–29.

———. 2001. "Reputations, Trust, and the Principal Agent Problem." In *Trust in Society,* ed. Karen Cook. New York: Russell Sage Foundation.

———. 2004. "Market Integration and Fairness: Evidence from Ultimatum, Dictator, and Public Goods Experiments in East Africa." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. New York: Oxford University Press.

Fehr, Ernst, and Joseph Henrich. 2003. "Is Strong Reciprocity a Maladaption?" In *Genetic and Cultural Evolution of Cooperation,* ed. Peter Hammerstein. Cambridge, Mass.: MIT Press.

Fehr, Ernst, Karla Hoff, and Mayuresh Kshetramade. 2008. "Spite and Development." *American Economic Review* 98(2): 494–99.

Fehr, Ernst, and Klaus Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114(3): 817–68.

Fehr, Ernst, and Frédéric Schneider. 2010. "Eyes Are on Us, but Nobody Cares: Are Eye Cues Relevant for Strong Reciprocity?" *Proceedings of the Royal Society B: Biological Sciences* 277(1686): 1315–23.

Fiske, Alan. 1992. "The Four Elementary Forms of Sociality: Framework for a Unified Theory of Social Relations." *Psychological Review* 99(4): 689–723.

Fowler, James H., and Cindy D. Kam. 2007. "Beyond the Self: Social Identity, Altruism, and Political Participation." *Journal of Politics* 69(3): 813–27.

Frey, Bruno, and Reto Jegen. 2004. "Motivation Crowding Theory." *Journal of Economic Surveys* 15(5): 589–611.

Gil-White, Francisco. 2004. "Ultimatum Game with an Ethnicity Manipulation: Results from Khovdiin Bulgan Cum, Mongolia." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. New York: Oxford University Press.

Gurven, Michael. 2004. "Does Market Exposure Affect Economic Game Behavior? The Ultimatum Game and the Public Goods Game Among the Tsimane' of Bolivia." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. Oxford: Oxford University Press.

Gurven, Michael, and Jeffrey Winking. 2008. "Collective Action in Action: Prosocial Behavior In and Out of the Laboratory." *American Anthropologist* 110(2): 179–90.

Güth, Werner, Carsten Schmidt, and Matthias Sutter. 2003. "Fairness in the Mail and Opportunism in the Internet: A Newspaper Experiment on Ultimatum Bargaining." *Economic Review* 4(2): 456–75.

Hagen, Edward H., and Peter Hammerstein. 2006. "Game Theory and Human Evolution: A Critique of Some Recent Interpretations of Experimental Games." *Theoretical Population Biology* 69(3): 339–48.

Haley, Kevin, and Daniel M. T. Fessler. 2005. "Nobody's Watching? Subtle Cues Affect Generosity in an Anonymous Economic Game." *Evolution and Human Behavior* 26(3): 245–56.

Harbaugh, William T., and Kate Krause. 2000. "Children's Altruism in Public Goods and Dictator Experiments." *Economic Inquiry* 38(1): 95–109.

Harbaugh, William T., Kate Krause, and Steven G. Liday. 2002. "Bargaining by Children." Economics Working Paper 2002-4. Eugene: University of Oregon.

Hennig-Schmidt, Heike, Zhu-Yu Li, and Chaoliang Yang. 2008. "Why People Reject Advantageous Offers: Non-monotone Strategies in Ultimatum Bargaining: First Results from a Video Experiment in the People's Republic of China." *Journal of Economic Behavior and Organization* 65(2): 373–84.

Henrich, Joseph. 2000. "Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining Among the Machiguenga." *American Economic Review* 90(4): 973–80.

———. 2008. "A Cultural Species." In *Explaining Culture Scientifically,* ed. Melissa Brown. Seattle: University of Washington Press.

———. 2009. "The Evolution of Costly Displays, cooperation, and Religion: Credibility Enhancing Displays and Their Implications for Cultural Evolution." *Evolution and Human Behavior* 30(4): 244–60.

Henrich, Joseph, and Robert Boyd. 2001. "Why People Punish Defectors: Weak Conformist Transmission Can Stabilize Costly Enforcement of Norms in Cooperative Dilemmas." *Journal of Theoretical Biology* 208(7): 79–89.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, eds. 2004. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* Oxford: Oxford University Press.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, Michael Alvard, Abigail Barr, Jean Ensminger, Natalie S. Henrich, Kim Hill, Francisco Gil-White, Michael Gurven, Frank W. Marlowe, John Q. Patton, and David Tracer. 2005a. "'Economic Man' in Cross-Cultural Perspective: Behavioral Experiments in Fifteen Small-Scale Societies." *Behavioral and Brain Sciences* 28(6): 795–855.

———. 2005b. "Models of Decision-Making and the Coevolution of Social Preferences." *Behavioral and Brain Sciences* 28(6): 838–55.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2004. "Overview and Synthesis." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. Oxford: Oxford University Press.

Henrich, Joseph, Robert Boyd, Richard McElreath, Michael Gurven, Peter J. Richerson, Jean Ensminger, Michael Alvard, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Colin F. Camerer, Juan-Camilo Cardenas, Ernst Fehr, Herbert M. Gintis, Francisco Gil-White, Edwins Laban Gwako, Natalie Henrich, Kim Hill, Carolyn Lesorogol, John Q. Patton, Frank W. Marlowe, David P. Tracer, and John Ziker. 2012. "Culture Does Account for Variation in Game Behavior." *Proceedings of the National Academy of Sciences of the United States of America* 109(2): E32–33.

Henrich, Joseph, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan-Camilo Cardenas, Michael Gurven, Edwins Laban Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David P. Tracer, and John Ziker. 2010. "Market, Religion, Community Size, and the Evolution of Fairness and Punishment." *Science* 327(5972): 1480–84.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. "The Weirdest People in the World?" *Behavior and Brain Sciences* 33(2-3): 1–23.

Henrich, Joseph, and Richard McElreath. 2002. "Are Peasants Risk-Averse Decision Makers?" *Current Anthropology* 43(1): 172–81.

Henrich, Joseph, Richard McElreath, Jean Ensminger, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan-Camilo Cardenas, Michael Gurven, Edwins Laban Gwako, Natalie Henrich, Carolyn Lesorogol, Frank Marlowe, David Tracer, and John Ziker. 2006. "Costly Punishment Across Human Societies." *Science* 312(5781): 1767–70.

Henrich, Joseph, and Natalie Smith. 2004. "Comparative Experimental Evidence from Machiguenga, Mapuche, and American Populations." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Herbert Gintis, Ernst Fehr, and Colin Camerer. Oxford: Oxford University Press.

Henrich, Natalie, and Joseph Henrich. 2007. *Why Humans Cooperate: A Cultural and Evolutionary Explanation.* Oxford: Oxford University Press.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008. "Antisocial Punishment Across Societies." *Science* 319(5868): 1362–67.

Heyman, James, and Dan Ariely. 2004. "Effort for Payment: A Tale of Two Markets." *Psychological Science* 15(11): 787–93.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon Smith. 1994. "Preferences, Property Rights, and Anonymity in Bargaining Games." *Game and Economic Behavior* 7(3): 346–80.

Huck, Steffen. 1999. "Responder Behavior in Ultimatum Offer Games with Incomplete Information." *Journal of Economic Psychology* 20(2): 183–206.

Kam, Cindy D., Skyler J. Cranmer, and James H. Fowler. N.d. "When It's Not All About Me: Altruism, Participation, and Political Context." Available at: http://jhfowler.ucsd.edu/altruism_participation_and_political_context.pdf (accessed October 2013).

Karlan, Dean S. 2005. "Using Experimental Economics to Measure Social Capital and Predict Financial Decisions." *American Economic Review* 95(5): 1688–99.

Kelly, Raymond C. 1985. *The Nuer Conquest.* Ann Arbor: University of Michigan Press.

Kim, H. S., D. K. Sherman, S. E. Taylor, J. Y. Sasaki, T. Q. Chu, C. Ryu, et al. 2010. "Culture, Serotonin Receptor Polymorphism (5-HTR1A), and Locus of Attention." *Social Cognitive and Affective Neuroscience* 5(2-3): 212–18.

Knafo, Ariel, S. Israel, Ariel Darvasi, Rachel Bachner-Melman, F. Uzefovsky, L. Cohen, E. Feldman, E. Lerer, E. Laiba, Y. Raz, L. Nemanov, I. Gritsenko, C. Dina, G. Agam, B. Dean, G. Bornstein, and R. P. Ebstein. 2008. "Individual Differences in Allocation of Funds in the Dictator Game Associated with Length of the Arginine Vasopressin 1a Receptor RS3 Promoter Region and Correlation Between RS3 Length and Hippocampal mRNA." *Genes Brain and Behavior* 7(3): 266–75.

Laland, Kevin N., John Odling-Smee, and Sean Myles. 2010. "How Culture Shaped the Human Genome: Bringing Genetics and the Human Sciences Together." *Nature Reviews Genetics* 11(2): 137–48.

Lamba, Shakti, and Ruth Mace. 2011. "Demography and Ecology Drive Variation in Cooperation Across Human Populations." *Proceedings of the National Academy of Sciences of the United States of America* 108(35): 14426–30.

———. 2013. "The Evolution of Fairness: Explaining Variation in Bargaining Behaviour." *Proceedings of the Royal Society B: Biological Sciences* 280(1750): 20122028.

Levitt, Steven D., and John A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21(2): 153–74.

List, John A., and Todd L. Cherry. 2008. "Examining the Role of Fairness in High Stakes Allocation Decisions." *Journal of Economic Behavior and Organization* 65(1): 1–8.

Marlowe, Frank W. 2004. "Dictators and Ultimatums in an Egalitarian Society of Hunter-Gatherers, the Hadza of Tanzania." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. Oxford: Oxford University Press.

Marlowe, Frank, J. Colette Berbesque, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan-Camilo Cardenas, Jean Ensminger, Michael Gurven, Edwins Laban Gwako, Joseph Henrich, Natalie Henrich, Carolyn Lesorogol, Richard McElreath, and David Tracer. 2008. "More 'Altruistic' Punishment in Larger Societies." *Proceedings of the Royal Society B: Biological Sciences* 275(1634): 587–90.

McElreath, Richard, Robert Boyd, and Peter J. Richerson. 2003. "Shared Norms and the Evolution of Ethnic Markers." *Current Anthropology* 44(1): 122–29.

McElreath, Richard, and Colin Camerer. 2004. "Appendix: Estimating Risk Aversion from Ultimatum Game Data." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. Oxford: Oxford University Press.

Nisbett, Richard E. 2009. *Intelligence and How to Get It: Why Schools and Cultures Count.* New York: W. W. Norton and Co.

Norenzayan, Ara, and Azim F. Shariff. 2008. "The Origin and Evolution of Religious Prosociality." *Science* 322(5898): 58–62.

Paciotti, Brian, and Craig Hadley. 2003. "The Ultimatum Game in Southwestern Tanzania." *Current Anthropology* 44(3): 427–32.

Panchanathan, Karthic, and Robert Boyd. 2003. "A Tale of Two Defectors: The Importance of Standing for the Evolution of Indirect Reciprocity." *Journal of Theoretical Biology* 224(1): 115–26.

———. 2004. "Indirect Reciprocity Can Stabilize Cooperation Without the Second-Order Free Rider Problem." *Nature* 432(7016): 499–502.

Patton, John Q. 2004. "Coalitional Effects on Reciprocal Fairness in the Ultimatum Game: A Case from the Ecuadorian Amazon." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. Oxford: Oxford University Press.

Pinker, Steven. 2002. *The Blank Slate: The Modern Denial of Human Nature.* New York: Viking.

Rai, Tage, and Alan P. Fiske. 2010. "ODD (Observation and Description-Deprived) Psychological Research." *Behavioral and Brain Sciences* 33(2-3): 46–47.

Rakoczy, Hannes, Katharina Hamann, Felix Warneken, and Michael Tomasello. 2010. "Bigger Knows Better: Young Children Selectively Learn Rule Games from Adults Rather Than from Peers." *British Journal of Developmental Psychology* 28(4): 785–98.

Rakoczy, Hannes, Felix Warneken, and Michael Tomasello. 2008. "The Sources of Normativity: Young Children's Awareness of the Normative Structure of Games." *Developmental Psychology* 44(3): 875–81.

Richerson, Peter J., Robert Boyd, and Joseph Henrich. 2010. "Gene-Culture Coevolution in the Age of Genomics." *Proceedings of the National Academy of Sciences of the United States of America* 107(supplement 2): 8985–92.

Roes, Frans L. 1995. "The Size of Societies, Stratification, and Belief in High Gods Supportive of Human Morality." *Politics and the Life Sciences* 14(1): 73–77.

Roes, Frans L., and Michel Raymond. 2003. "Belief in Moralizing Gods." *Evolution and Human Behavior* 24(2): 126–35.

Sahlins, Marshall. 1961. "The Segmentary Lineage: An Organization of Predatory Expansion." *American Anthropologist* 63(2): 322–45.

Sanfey, Alan G. 2007. "Social Decision-Making: Insights from Game Theory and Neuroscience." *Science* 318(5850): 598–602.

Shariff, Azim F., and Ara Norenzayan. 2007. "God Is Watching You: Priming God Concepts Increases Prosocial Behavior in an Anonymous Economic Game." *Psychological Science* 18(9): 803–9.

Shariff, Azim, Ara Norenzayan, and Joseph Henrich. 2010. "The Birth of High Gods: How the Cultural Evolution of Supernatural Policing Agents Influenced the Emergence of Complex, Cooperative Human Societies, Paving the Way for Civilization." In *Evolution, Culture, and the Human Mind,* ed. Mark Schaller, Ara Norenzayan, Steve Heine, Toshi Yamaguishi, and Tatsuya Kameda. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Smith, Adam. 2000. *The Theory of Moral Sentiments.* New York: Prometheus Books. (Originally published in 1759.)

Sosis, Richard, and Candace Alcorta. 2003. "Signalling, Solidarity, and the Sacred: The Evolution of Religious Behavior." *Evolutionary Anthropology* 12: 264–74.

Sutter, Matthias, and Martin Kocher. 2007. "A Trust and Trustworthiness Across Different Age Groups." *Games and Economic Behavior* 59(2): 364–82.

Tracer, David. 2003. "Selfishness and Fairness in Economic and Evolutionary Perspective: An Experimental Economic Study in Papua New Guinea." *Current Anthropology* 44(3): 432–38.

———. 2004. "Market Integration, Reciprocity, and Fairness in Rural Papua New Guinea: Results from Two-Village Ultimatum Game Experiments." In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies,* ed. Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis. Oxford: Oxford University Press.

Turkheimer, Eric, Andreana Haley, Mary Waldron, Brian D'Onofrio, and Irving I. Gottesman. 2003. "Socioeconomic Status Modifies Heritability of IQ in Young Children." *Psychological Science* 14(6): 623–28.

Tversky, Amos, and Daniel Kahneman. 1992. "Advances in Prospect Theory—Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5(4): 297–323.

Vohs, Kathleen D., Nicole L. Mead, and Miranda R. Goode. 2006. "The Psychological Consequences of Money." *Science* 314(5802): 1154–56.

———. 2008. "Merely Activating the Concept of Money Changes Personal and Interpersonal Behavior." *Current Directions in Psychological Science* 17(3): 208–12.

Wallace, Björn, David Cesarini, Paul Lichtenstein, and Magnus Johannesson. 2007. "Heritability of Ultimatum Game Responder Behavior." *Proceedings of the National Academy of Sciences of the United States of America* 104(40): 15631–34.

Wright, Robert. 2009. *The Evolution of God.* Boston: Little, Brown and Co.