

1. Supplementary Materials

1.1 Indicator of the prevalence of rule violations (and other institutional and cultural indicators)

We selected the countries of our subject pools to span a wide range according to relevant societal background indicators. Our approach avoids the problem of drawing inferences about human behaviour from only a small set of "WEIRD" societies¹ and allows learning about societal differences that smaller data sets from less varied societies would not permit. In particular, we aimed at societies that differ strongly (and hence give us strong power) with regard to our main variables of interest - corruption, the shadow economy, and the honesty of the political system in a country. Extended Data Table 1 summarises for the countries of our subject pools the detailed scores of the societal indicator data we use in this paper. Here, we provide the details about these indicators.

Index of the Prevalence of Rule Violations

Our aim is to construct one measure - the Index of the Prevalence of Rule Violations (PRV) - that captures several dimensions of the prevalence of rule violations at the societal level: corruption, tax evasion, and political fraud. The three underlying indicators we use and explain in detail below are (i) the Political Rights indicator by Freedom House as a proxy for the honesty of the political system, (ii) estimates by Buehn and Schneider for the shadow economy as a proxy for tax evasion and (iii) the World Bank's governance indicator 'control of corruption'.

For calculating PRV we use data from 2003, the earliest year where data on all three sources from which PRV is derived are available. We conducted the experiments between October 2011 and September 2015 with subjects who at the time of the experiments were on average 21.7 years old. That means that in 2003 they were between 11 and 14 years old and thus most likely had no influence on the PRV of their country.

We now describe these indicators and then how we constructed the PRV.

Control of Corruption. The World Bank Governance indicator 'Control of Corruption' (<http://info.worldbank.org/governance/wgi/wgidataset.xlsx>, accessed 28.10.2015) is our

measure of corruption. It captures “perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as ‘capture’ of the state by elites and private interests” (Kaufmann, et al.², p. 223). Thus, corruption is a form of cheating, of bending the rules for one's own purposes. ‘Control of corruption’ is a widely used aggregate indicator that is based on 15 representative and non-representative sources. Since corruption itself is hardly measurable the focus is on *perception of corruption* by public sector, private sector and NGO experts, as well citizens and company survey respondents. The variable ‘Control of Corruption’ is standardised and it ranges from -1.8 to 2.5. The country scores in our sample range from -1.0 to 2.2, where higher values indicate better control of corruption (i.e., less corruption). The world average is 0.0, and our country sample mean is 0.4.

The Shadow Economy. The shadow economy concerns the market-based production of legal goods and services with a deliberate goal to conceal this activity to avoid income, value added or other taxes, social security contributions, or to avoid compliance with regulations and administrative procedures. It is a significant problem for many economies³⁻⁵ and an important example for dishonest practices. Our data for the size of the shadow economy are from Buehn and Schneider⁵, Table 3. Their estimates are based on monetary indicators (cash holdings), labour market indicators (labour market participation rate and growth rates) and the state of the official economy (GDP per capita and growth rates). The size of the shadow economy is measured in percent of the size of the GDP. In 2003 the average country in our sample had a shadow economy of 28% (fairly close to the world average of 33%); the range is from 10% to 66%.

Political Rights. To capture the honesty of political processes, we use the ‘Political Rights’ scores by Freedom House (<https://freedomhouse.org/report/freedom-world-aggregate-and-subcategory-scores>, file ‘Aggregate Scores’, accessed 28.10.2015). The scores are based on expert judgments guided by a series of checklist questions grouped in three sub-categories. (A) *Electoral process* focuses on free and fair elections of the executive and legislative as well as a fair electoral framework. For example, underlying checklist questions ask whether the vote count is transparent and honestly reported, or

whether the registration of voters and candidates is conducted in an accurate, timely, transparent and non-discriminatory manner. (B) *Political Pluralism and Participation* is based on questions on the discrimination of (oppositional) political parties, the extent to which political choices are free from domination by powerful groups (e.g. by bribing, intimidation, harassment or attacks), and questions on minority voting rights; (C) *Functioning of Government* focuses on corruption, transparency, and the ability of elected officials to govern in practice. For example, checklist question ask whether allegations of corruption by government officials are thoroughly investigated and prosecuted, or whether the budget-making process is subject to meaningful legislative review and public scrutiny. The scores vary between -2 and 44 with higher scores referring to a higher level of political rights. Our sample mean is 28, somewhat higher than the world average of 24. The countries in our sample cover almost the whole range of values from 2 to 40.

Index of Prevalence of Rule Violations (PRV). Consistent with expectations, the three variables (Control of Corruption (CC), Shadow Economy (SE) and Political Rights (PR)), which measure different but related aspects of the prevalence of rule violations, are substantially and highly significantly ($P < 0.0001$) correlated (CC & SE: -0.7; CC & PR: 0.67; SE & PR: -0.4). This suggests there is an underlying component (which we call "prevalence of rule violations") that drives these correlations. We apply a Principal Component Analysis (PCA) to uncover this component (or the components). The purpose of the PCA is to reduce the dimensionality of a data set with correlated variables down to a number of variables (the principal components) that explain a significant fraction of the variation in the underlying data set and thereby succinctly summarize the data⁶. The fact that the three variables are sufficiently strongly, but less than perfectly, correlated makes a Principal Component Analysis (PCA) meaningful.

To perform the PCA and to construct our PRV we use country averages for the year 2003 of 159 countries for which we have data for all three indicators. The year 2003 is the first year where data is available for all three indicators. Our initial PCA with three components shows a clear distinction between the first component (eigenvalue of 2.13 – explaining 71% percent of the variance) and the second (eigenvalue of 0.67) and third

component (eigenvalue of 0.2) explaining 22% and 7% respectively. This result demonstrates that the three indicators share one component that explains a very large fraction of the variance. Based on this finding (and following the Kaiser criterion which suggests dropping components with eigenvalues below 1) we only retained the first component. For this component we calculated composite scores for each country, which constitute our PRV. For the sample of 159 countries PRV has a mean of about zero, a standard deviation of 1.46. PRV ranges from -3.1 to 2.8. Higher values indicate a higher prevalence of rule violations. For the countries in our sample, the mean is -0.7 and the range is from -3.1 to 2. The overlap in the range of PRV is considerable (from -3.1 to 2). Only 7 countries (4%) out of 159 have a PRV higher than '2' (the largest value in our sample). Thus, our sample is approximately representative of the world distribution.

Institutional Indicators

As indicators of the quality of institutions, we use two frequently used measures, Constraint on the Executive, Fairness of Electoral Process and Participation (which is a sub-indicator of Constraint on the Executive, and Government Effectiveness.

Constraint on the Executive. Our proxy for political institutions is Constraint on the Executive from the Polity IV data set (<http://www.systemicpeace.org/inscrdata.html>, accessed 13.06.2015). It measures the institutionalized limitations on the arbitrary use of power by the executive. Any “accountability groups” may impose such limitations to varying degrees. This can be legislatures; the ruling party in a one-party state; councils of nobles or powerful advisors in monarchies; the military in coup-prone polities; and in many states a strong, independent judiciary. ‘Constraint on Executive’ is closely related to protection from government expropriation: More constraints implies that the executive cannot simply expropriate, but is accountable to legislatures and other groups as well as the judiciary. ‘Constraint on Executive’ is a widely-used measure for the quality of institutions^{7,8}.

Conceptually this measure has the feature that it relies on expert opinions on (mostly) observable features of a well-defined set of political institutions. As Acemoglu and Johnson⁷ argue it corresponds to the procedural rules constraining state actions. Thus,

in contrast to subjective performance measures of institutions, this measure captures variation in institutions without being strongly associated with variation in law compliance of ordinary citizens. ‘Constraint on Executive’ is measured on a scale ranging from 1 (unlimited authority) to 7 (accountable executive, constrained by checks and balances). Thus, higher values correspond to better institutions. For our contemporary measure we used country averages of the years 1990 to 2000. The world sample varies from 1 to 7 with a mean of 4.5 while the countries in our sample exhibit a mean of 5.5 and vary from 2.8 to 7. The Spearman correlation between ‘Constraint on Executive (1990 to 2000)’ and PRV is -0.67 ($P < 0.0001$, $n = 150$). This is also the case when we use a historic measure ‘Constraint on Executive (1890 to 1900)’. The Spearman correlation is -0.60 ($P < 0.0001$, $n = 50$).

Fairness of Electoral Process and Participation. For this variable we excluded the subcategory ‘Functioning of Government’ from the ‘Political Rights’ indicator (the data are taken from <https://freedomhouse.org/report/freedom-world-aggregate-and-subcategory-scores>, file ‘Subcategory Scores’, accessed 28.10.2012). This leaves only the subcategories ‘Electoral Process’ and ‘Political Participation and Pluralism’. The rationale for constructing this measure is that compared to ‘Political Rights’ it captures to a larger degree variations in law enforcement without being as strongly associated with variation in law compliance of ordinary citizens. That is, it is more likely to measure the functioning of the law and less likely to be affected by the dishonesty of the generation of the parents of our participants. The correlation with PRV is -0.73 ($P < 0.001$).

Government Effectiveness. This is an indicator developed by the World Bank. It captures perceptions of the quality of public and civil service, the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government’s commitment to such policies (<http://info.worldbank.org/governance/wgi/wgidataset.xlsx>, accessed 28.10.2015). Aside from the quality of bureaucracy the measure reflects satisfaction with infrastructure, education, drinking water, sanitation and basic health services. As such, it can be used as a proxy for material security⁹. It is also correlated with behaviour in a die-rolling

experiment related to ours.¹⁰ We use the scores for the year 2000. The Government Effectiveness indicator is a standardised variable with world mean of approx. 0, standard deviation of approx. 1 and an empirical range of -2.3 to 2.2. The average for the countries of our sample is 0.5, ranging from -0.7 to 2.1. The Spearman correlation with PRV is -0.89 ($P < 0.001$; $n = 159$) and with GDP per capita 0.78 ($P < 0.001$; $n = 180$).

Cultural Indicators

As indicators of cultural differences between our subject pools we use two frequently used measures, Individualism/Collectivism and Value Orientations.

Individualism/Collectivism. Our measure for Individualism/Collectivism is due to Hofstede¹¹ (retrieved from <http://geert-hofstede.com/>, accessed 28.10.2015 - the results are qualitatively very similar when using the smaller subset published in Hofstede et al.¹²). ‘Individualism’ measures how important the individual is relative to the collective in a society. Collectivist societies are tightly knit and individuals act predominantly as loyal members of a lifelong and cohesive group or organization; individualist societies are more loosely knit and group boundaries are more permeable.

Individualism–collectivism is an important cultural distinction.¹³ Our interest stems from the idea that in collectivist societies morality tends to be limited to the in-group.^{14–16} Rule violations favouring the in-group may therefore be more frequent and sanctioned less by informal institutions in collectivist societies. Further, the strong reliance on the family or the clan may prevent the development of functioning formal institutions. Thus, the hypothesis is that individualist societies have a lower prevalence of rule violations. Existing research suggests that more collectivist societies tend to be more corrupt¹⁷, have weaker formal institutions^{8,18}, less innovation and weaker growth.^{19,20} Furthermore, research suggests that there is no robust effect on economic growth from cultural dimensions that are independent from the individualism–collectivism cultural trait.²¹

In the world sample the range is from 6 to 91 (average is 39), where higher values indicate more individualist societies. Our sample average is 46, and our range is from 6 to 89. Consistent with our hypothesis, the Spearman correlation with PRV is -0.65 ($P < 0.0001$, $n = 99$). That is, more collectivist societies tend to have higher PRV.

Value orientations. These indicators are due to Inglehart and co-workers²²⁻²⁴, who argue that societies can be characterized by two dimensions: ‘**traditional vs. secular-rational values**’ and ‘**survival vs. self-expression values**’. The data on value orientations are taken from Inglehart and Welzel²⁵.

‘Traditional vs. secular-rational values’ refers to people’s attitudes on topics like abortion, national pride, obedience, and respect for authorities. ‘Survival vs. self-expression values’ refers to attitudes on the importance of economic and physical security over self-expression and quality-of-life; homosexuality, happiness and trust.

Our sample averages on value orientations are very similar to the world averages and they also cover a wide range compared to the world sample. Both indicators, ‘Traditional vs. secular-rational values’ (Spearman's $\rho = -0.50$, $P < 0.0001$, $n = 86$) and ‘Survival vs. self-expression values’ ($\rho = -0.70$, $P < 0.0001$, $n = 86$) are highly correlated with PRV.

Survival values are interesting because they might be linked to material security, which can have an impact on quality of institutions, in particular Government Effectiveness as a proxy for material security.¹⁰ We find that survival values are correlated with Government Effectiveness ($\rho = 0.71$; $P = 0.0000$, $n = 92$).

Traditional values are also interesting because they are correlated with beliefs in heaven and hell (which have been found to correlate with national crime levels²⁶): Beliefs in heaven and hell (data taken from²⁶) are correlated with Traditional values (beliefs in heaven, $\rho = -0.82$, $P = 0.0000$, $n = 63$; beliefs in hell: $\rho = -0.73$, $P = 0.0000$, $n = 73$).

Economic Indicator

GDP per Capita. On obvious economic variable to characterise societies is the GDP per capita. Due to its long history and importance in country comparisons it is likely a comparable objective measure of economic prosperity. The data are taken from the World Economic Outlook of the International Monetary Fund (retrieved from <http://www.imf.org/external/pubs/ft/weo/2014/01/weodata/index.aspx>). Between 1990 and 2000 the mean GDP per capita (in \$1000, PPP) in the countries of our sample is 9.9 compared to 7.8 in the world sample. The range in our sample is from 0.7 to 23.7. The Spearman correlation of the GDP with PRV is -0.76 ($P < 0.0001$, $n = 159$).

1.2 Experimental design and procedures

The task we use is a "die-in-a-cup task" due to Fischbacher and Föllmi-Heusi²⁷ which we implement in a very similar way as them. It has several features, which makes it an ideal task for our purpose. It can be conducted quickly, is very easy to understand for participants, allows for gradual dishonesty and is a non-strategic setting. That is, payoffs do not depend on the action of other participants. As lying can never be verified at the individual level, it gives a benchmark for dishonesty when reputational concerns are absent and subjects are aware of this.

Because the task is very short (it lasts about 5 minutes), it was, as in²⁷, added at the end of an experimental session featuring unrelated experiments (we will discuss them in section 2.5 and show that spillover effects are unlikely). The following texts and three screen shots document how we explained the task to the subjects. Extended Data Figure 1 illustrates the typical laboratory setup in which decisions occurred.

Script and Screenshots of Instructions

Script: A local experimenter (always a native speaker) read the following text (translated into the local language):

"Before we start with the pay-out we ask you to answer a short questionnaire. All answers are treated as strictly confidential. For the following questionnaires you will receive an additional payment. However, this payment is not the same for every participant. You determine your own payment by throwing a die twice. You will find the instructions on the computer screen. As soon as you have read and understood the instructions, please press OK. Do not roll the die before you have read the instructions and you are told on the computer screen to roll the die. Let me repeat: do not throw the die until you are told on the computer screen."

Subsequently the six-sided dice are distributed in non-transparent plastic cups and the z-Tree treatment is started.

Instructions/Screenshots for the die-in-a-cup task

Instructions were only delivered on screen. Screenshots I-III show the interface subjects saw (the screenshots are from the English version). These texts were translated into the respective local language and payoffs were adapted to reflect local purchasing power. More details on the cross-societal implementation are described in section 1.3.

Screenshot I: Introducing the task

For the following questionnaires you will receive a small additional payoff. However, this payoff is not the same for every participant. You determine your own payoff by throwing your die twice as soon as you are asked to.

Your first throw decides on how much you receive. You can see the exact payoff from the following chart. It will remain on the screen until you entered your throw.

The second throw only serves to make sure that the die is working properly. You may of course throw the die more than twice. However, only the first throw counts.

If you have any questions, please raise your hand. If you are ready, please press OK.

Number thrown:	1	2	3	4	5	6
Resulting payoff (in £):	0.5	1.0	1.5	2.0	2.5	0.0

OK

Screenshot II: Prompting subjects to throw the die

Please throw the die now.

Please keep in mind the first number you have thrown.

If you have thrown the die, please press OK.

Number thrown:	1	2	3	4	5	6
Resulting payoff (in £):	0.5	1.0	1.5	2.0	2.5	0.0

OK

Screenshot III: Prompting subjects to report their number and resulting claim

Now please enter the number, i. e. the first number you have thrown.

Number thrown:

Resulting payoff:

Number thrown:	1	2	3	4	5	6
Resulting payoff (in £):	0.50	1.00	1.50	2.00	2.50	0.00

Testing the quality of the dice

To ensure that the dice used in a particular subject pool do not bias the results of die rolling, we used the same set of dice in all subject pools. In total, we used a set of 71 dice, from which we randomly sampled the actual number needed in a particular session.

To test the quality of dice we randomly sampled 35 dice ($\approx 50\%$ of dice) and tested them by rolling each of them in a cup for 120 times. We ran $\text{Chi}^2(5)$ tests to see whether the distribution of $\text{Chi}^2(5)$ -values is within the theoretically expected range, which we determined by simulated die rolls, using a computerised random number generator. For a fair unbiased die, all numbers are equally likely. Random variation does exist, but at $\alpha=0.05$ $\text{Chi}^2(5)$ -values should be below 11.07 with a probability of 95%. That is, in our sample of 35 dice we expect about 1.75 $\text{Chi}^2(5)$ -values exceeding 11.07. We observed three $\text{Chi}^2(5) > 11.07$, which is not significantly different from the expected value (binomial test, $P = 0.2542$). The empirical distribution of the 35 $\text{Chi}^2(5)$ -values is also not

significantly different from the simulated distribution (Kolmogorov-Smirnov test, $P = 0.3940$). We conclude that the dice we used are unbiased.

Rationale and indicators of (dis)honesty derived from the task

Because participants roll the die in private the reported numbers are never verifiable and subjects are aware of this. Reputational incentives are minimised. Thus, participants who did not roll a 5 have a financial incentive to report a higher number than they actually rolled (the material incentives are to report a 5) and no one will ever find out. In that sense the experiment taps into people's pure intrinsic preferences for honesty that are not confounded with considerations of weighing costs and benefits of dishonesty because the experiment sets the material and reputational costs essentially to zero.

While individual dishonesty is not detectable, aggregate behaviour is informative about dishonesty. If all people are honest and report the number they actually rolled, all six numbers should be equally distributed with frequency $1/6 \approx 16.7\%$. Expressed in terms of money units (MU), the expected claim therefore is 2.5 MU. If people are maximally dishonest, they would all report '5'.

Given previous results on this task, in particular by Fischbacher and Heusi-Föllmi²⁷ whose design and procedures we have adopted, and previous literature using the same²⁸⁻³⁰ or related^{10,31-36}, but also different tasks³⁷⁻⁴¹), we expect the results to be between these two extreme outcomes⁴¹. An important benchmark, discussed in detail in the main text, is Justified Dishonesty²⁸, according to which people will not report a number they actually did not roll, but they might bend the rules and report the better of the two rolls, rather than the first one, as the rules stipulate.

In addition to comparing reporting patterns to the Justified Dishonesty benchmark, our simple experiment allows us to look at four measures of honesty: the actual average claim; the fraction of high numbers (3, 4, 5); the estimated fraction of income maximizing (dishonest) people; and the estimated fraction of fully honest people:

1. The *actual average claim* ('Mean Claim') can be interpreted as the severity of dishonesty in a subject pool. In an honest subject pool mean claims are 2.5 MU, while in a fully dishonest subject pool claims are 5 MU. The average claim implied by the *Justified Dishonesty* benchmark is 3.47 MU.

2. The *fraction of high claims* ('High Claims', 3, 4, and 5). If subjects are honest, 50% of all reports should be 3, 4 or 5. However, subjects who roll a lower-paying number (6, 1, and 2) may have an incentive to lie and report a higher number. While income maximising people will report a 5 some people may shy away from exaggerating too much and increase their claims by reporting a 3 or a 4. Compared to the *actual average claim* it captures therefore to a greater extent the frequency of lies (and not severity) in a subject pool. The expected frequency of High Claims is 75% under the *Justified Dishonesty* benchmark.
3. The *fraction of income maximising people* ('Income Maximisers') can be estimated from the number of people who report a '5' assuming that all people who actually rolled a '5' report a 5 (expected to occur in 16.7% of cases). The fraction of income maximisers can therefore be calculated from the fraction s of reported 5's: $(s - 0.167) * (6/5)$. If only those who actually rolled a '5' report a 5, $s = 0.167$, the fraction of income maximising people is 0; if all people who did not roll a '5' (expected to occur for 83.3% of people) nevertheless report a 5, the fraction of income maximising people is estimated to be $(1 - 0.167) * 1.2 = 1$. The multiplication of $(6/5) = 1.2$ is necessary to account for income maximising people who actually rolled a 5 (see also Fischbacher and Heusi-Föllmi²⁷, footnote 5). The measure of income maximisers contains *homo economicus* types who will always report the number 5 irrespective of the number they actually rolled. However, it may also contain limited dishonest people (e.g., a person who reports the number 5 in case he rolled a 4, but reports 4 if he rolled a 3). Thus, the measure of income maximisers is an upper-bound estimate for *homo economicus* types. Under *Justified Dishonesty*, nobody is an income-maximiser (in the sense of always claiming 5), so this benchmark is not properly defined. With regard to the expected number of claims of 5 under justified dishonesty we expect claims of 5 whenever a 5 showed up in any of the two rolls, which is expected to happen in $11/36 \approx 30.6\%$ of cases.
4. The *fraction of unconditionally honest people* ('Fully Honest People') can be estimated from the fraction h of people who report a '6', which earns nothing ('No Claim'). People who are willing to report the truth even if it earns nothing are

arguably at least as likely to tell the truth if it earns something. Hence, we can assume that unconditionally honest people report any number they actually roll and honest reports can therefore occur across all die numbers. We can therefore use the fraction h of reports of 6 to estimate the fraction of fully honest people across all die rolls: $h \times 6$. We speak of ‘fully honest people’ to emphasise that people who did not truthfully report a claim of 0, would have truthfully reported a claim of 1 or higher. In this sense it is a lower bound for honesty. However, it would be an upper bound estimate for ‘fully honest people’ if some people reported a 6 and therefore claim 0 despite having rolled another number. Since the measure is based on a single claim (the one that earns nothing) this is a more noisy measure than ‘High claims’ and ‘Mean Claim’. Under *Justified Dishonesty* nobody is unconditionally honest. Thus, strictly speaking, Justified Dishonesty is not defined, but we expect to see No Claim only if subjects roll 6-6, which is expected to happen in $1/36 \approx 2.8\%$ of cases.

Procedures

We conducted all experiments in computer laboratories at the respective universities or institutions (see Section 1.4, Table S1). All participants gave informed consent. We took great care that the participants made their decision in private. In all laboratories we set up partitions that ensured maximal anonymity. Additionally, we distributed the die in a cup. This ensures that only participants and no one else can find out the actual number rolled. Distributing the die in a cup is also very natural. During the experiment all experimenters stayed at the experimenter computer (invisible to the participants since they were sitting behind their partitions). Extended Data Figure 1 illustrates the typical setup as seen by the participants.

We used the experimental software z-Tree⁴² to conduct our experiment. Z-tree was adapted so that messages on-screen appeared in the local language (the z-Tree codes we used are available upon request). Subjects typed in their reported number *and* the respective earnings in a computer interface (see screenshot III). Subjects could only continue when the entries matched. This ensured that participants understood the financial consequences of their report.

We conducted the experiments between October 2011 and September 2015 according to established methods of cross-cultural experimental economics (see Section 1.3 for the details). We aimed at recruiting participants according to standardised procedures subject to constraints at local universities. We did not ask potential participants about their nationality during recruitment to not introduce experimenter effects. For this reason we did not only recruit nationals of the respective country. Out of the 2790 participants we had to drop 222 observations for which the nationality did not coincide with the country the experiment was conducted. (We kept data from one session in China, where we do not have questionnaire data about nationality. However, based on existing data only 1 percent foreigners participated in China.) This leaves us with 2568 participants who are nationals of the country of the experiment.

In every country participants took part in a public goods game (with and without punishment; PG for short) prior to the die-in-a-cup task. To check for robustness, in several countries we additionally varied the experiment prior to our task and we also conducted stand-alone experiments (that is, the die-in-a-cup task without a prior experiment) as a test for potential spillovers from the PG. Based on a total of 721 subjects, who participated in a different or no previous experiment, and a regression analysis we do not find an impact of the experiment preceding our die-in-a-cup task on reporting behaviour (see Section 2.5 for details).

1.3 Further methodological details of the cross-societal implementation

The implementation of our experiment closely followed the methodological standards introduced by Roth et al.⁴³ and adapted to a multiple country setup by Herrmann et al.⁴⁴. We aimed at only varying the societal background (i.e., country where the experiment was conducted) while minimising any variation from other sources. We took several steps to ensure this goal:

- We selected societies that span a wide range of cultural and economic backgrounds, in particular indicators of the prevalence of rule violations that are most relevant to our study (Section 1.1 and Extended Data Table 1).

- We ran the experiment with undergraduate students. This minimises variability in the socio-demographic composition of the subject pools. Our participants share a similar age, level of education and upper middleclass background (Section 1.4). Young people have the further decisive advantage for our purposes that they will have had very limited possibilities of engaging in political fraud, in evading taxes, and in acts of corruption. The societal environment as we measure it (Section 1.1) is hence uninfluenced by our subjects.
- Subject to local practices, we recruited participants according to a standardised procedure intended to obtain an approximately representative selection of local students who also did not know one another. In universities that host an experimental economics lab and use online recruiting systems we relied on local recruitment practices. In other universities we recruited participants by approaching students at random all around campus. We aimed to recruit students that did not have prior experience in economic experiments. Students who agreed to participate received a text message or an email to remind them of participation.
- We aimed at getting at least 50 participants per subject pool. The average subject pool size is 112; the range is from 54 to 244 participants per subject pool (table S1).
- We implemented a standardised protocol of how to conduct the experiment to minimise experimenter effects (see below).

Instructions and protocol

Subjects received instructions displayed on-screen in the local language. Instructions were originally written in English (see screen shots I-III) and, as is common practice in cross-cultural experiments, a native speaker translated the instructions into the relevant local language. A second native speaker provided a back translation⁴⁵. Translations were fine-tuned until they converged.

We conducted all sessions according to a standardised protocol. To further minimise experimenter effects, which may originate from subtle differences in the implementation of the protocol, we had one author (J.S.) present in the background in almost all sessions. He trained local research assistants and ensured consistent implementation of the protocol. Since subjects may react differently to a foreign experimenter, local native

speakers led the experiment and made announcements. J.S. carefully trained the local experimenters to minimise variation between subject pools. During the experiments J.S. stayed in the background and did not interact with the participants. Thus, participants were not aware of the international dimension of this research and participants most likely perceived it as a locally run experiment.

Currency effects and stakes

Simply converting earnings according to the nominal exchange rates would neglect differences in local purchasing power (PP) in the different subject pools. To minimise differences in stake sizes we collected information about the hourly wage of a typical student job as well as the costs of a lunch at the cafeteria of the university. Maximal earnings amounted to about 1/4th of an hourly wage of a typical student job or about the price of half a lunch at a student cafeteria (recall that the experiment lasts only a few minutes). We adjusted stake sizes using those measures as they likely reflect local PP. Table S1 reports the nominal \$-earnings of reporting the maximal claim. These vary from \$0.7 dollar in Vietnam to \$4.2 in the Netherlands. This reflects the fact that a dollar in Ho Chi Minh City has a considerably higher PP for a student than in the Netherlands.

Even if small PP differences in the adjusted stake sizes remain they are unlikely to have an impact on our measures of dishonesty; the existing literature does not find stake size effects in laboratory experiments on dishonesty.⁴¹ Fischbacher and Heusi-Föllmi²⁷ did not find a significant difference when they tripled financial incentives in their die-in-a-cup experiment. Similarly, Mazar et al.³⁸ show that quadrupling the incentives had no significant impact on dishonesty in their matrix task. As further evidence, in one non-western country (China) we conducted in addition to our normal incentives ($n = 138$) a low-stakes treatment ($n = 99$), where we divided standard incentives by 1/4. Average payments are 3.51 MU when subjects are faced with standard incentives and 3.59 MU with low incentives. A Rank sum test does not reveal significant differences in reported numbers ($P = 0.8398$). This is consistent with evidence that in simple tasks involving social preferences, stake size effects are absent^{46–48}. Thus, within the range of tripling or quadrupling (small) monetary incentives stake sizes are unlikely to impact dishonesty.

Further evidence comes from regression analyses we report in Extended Data Table 2 and Table S3 in Section 2.2. We included the variable ‘middleclass’, which reflects whether a participant reported its family’s income at the age of 16 to be below average (compared to other families in its country) or not. The variable ‘middleclass’ is not significantly related to participants’ reported claims. That is, even though the financial incentives to be dishonest are relatively bigger for poorer participants within a country, they are not significantly more likely to report higher numbers.

1.4 Subject pool details

Our subject pools consisted of samples of undergraduate students from local universities. We chose undergraduate students because we aimed for subjects who are homogenous across subject pools with respect to their socio-economic background. Large variability in the socio-demographic composition may introduce confounding factors with the subject pool differences we are interested in.

Student samples have the advantage that participant pools are comparable: subjects are of similar age, typically from an upper or middleclass background and have a very similar level of prior education. They tend to be intelligent and used to problem solving. All of this qualifies them as a very useful subject pool if large random samples are not feasible and the main purpose is to test conceptual arguments⁴⁹.

Another important reason for using young subjects is that due to their youth they have had limited chances of influencing the honesty of their social environment, that is, the social environment (which we measure with PRV in 2003, when our subjects were between 11 and 14 years old – see Section 1.1) is arguably uninfluenced by our subjects.

Although student subject pools are homogenous in many respects, some potential variation remains, which we measured with the help of a post-experimental questionnaire. Table S1 lists the countries, cities and universities the students were recruited from, the number of participants in each country, subjects' average nominal earnings in \$, as well as the control variables we elicited. We use these variables as controls in the regression analyses (Extended Data Table 2; Tables S3, S4).

Country	City	Student's University	Total # of subjects	Nominal max. earnings in \$	Age	% female	% middle class	% urban	% economics student	% religious	% known in a session	% belief in fairness	Ind. norms of honesty
Austria	Vienna	University of Vienna	66	4.0	24.7	45	71	44	18	17	2	64	7.8
China	Shanghai	Jiao Tong University	237	1.9*	22.1	35	60	10	0	12	5	82	9.5
Colombia	Bogota	Universidad Nacional	104	2.9	19.8	45	83	63	13	33	5	62	8.0
Czech Rep.	Prague	University of Economics	77	2.6	22.9	57	87	73	62	21	2	40	7.5
Georgia	Tbilisi	Tbilisi State University	97	1.8	19.1	57	77	49	0	78	15	30	8.7
Germany	Constance	University of Constance HTWG Konstanz	69	4.0	21.4	55	87	52	10	32	4	65	7.6
Guatemala	Guatemala City	Universidad Francisco Marroquin Universidad Rafael Landivar Universidad San Carlos de Guatemala Universidad Galileo Universidad del Valle de Guatemala	193	3.3	22.0	42	95	50	7	66	9	43	8.8
Indonesia	Yogyakarta	Gadja Mada University	76	1.1	20.7	64	87	38	41	92	10	43	9.3
Italy	Rome	La Sapienza LUISS Guido Carli Tor Vergata	82	3.5	26.6	55	89	77	18	38	1	38	8.3
Kenya	Nairobi	Nairobi University	92	1.2	22.3	29	60	40	47	89	15	29	8.5
Lithuania	Vilnius	Vilnius University	71	2.3	20.4	55	92	72	3	31	13	62	8.5
Malaysia	Semenyih	University of Nottingham, Malaysian Campus	64	1.6	19.7	42	80	42	3	69	8	45	7.0
Morocco	Meknes	Ecole Nationale d'Agriculture Meknès	138	1.8	20.9	65	88	48	1	87	50	34	8.5
Netherlands	Amsterdam Groningen	University of Amsterdam University of Groningen	84	4.2	21.8	49	93	62	30	21	3	80	7.3
Poland	Warsaw	University of Warsaw	110	2.1	23.1	61	86	82	30	63	2	57	7.1
Slovakia	Bratislava	Slovak Technical University Comenius University	87	2.6	22.9	53	83	56	2	43	13	43	7.1
South Africa	Cape Town	University of Cape Town	92	2.2	20.2	47	74	46	8	54	5	41	8.4
Spain	Granada	University of Granada	54	3.3	22.6	44	88	50	52	41	7	-	8.4
Sweden	Linköping	Linköping University	82	3.6	23.9	27	87	67	9	9	7	78	7.8
Turkey	Izmir	Izmir University of Economics	244	2.8	21.9	47	87	82	5	44	21	43	7.8
United Kingdom	Nottingham	University of Nottingham	197	4.1	19.3	49	79	48	13	33	1	58	8.2
Tanzania	Dar Es Salaam	University of Dar Es Salaam	140	1.6	23.2	29	60	34	22	84	11	41	7.2
Vietnam	Ho Chi Minh City	Economics University	112	0.7	20.0	64	88	25	60	27	10	65	7.2
Mean			112	2.7	21.8	49	82	53	20	47	10	52	8.0

Table S1 | Subject pool details, socio-demographics, beliefs in fairness of others, and personal norms of honesty. *In China, for a subset the maximal nominal earnings was \$0.5.

Socio-demographics

Age and **% female** are standard variables. The average participant was 21.7 (s.d. 3.3) years old and subject pools' average age ranged from 19.1 years to 26.6 years; 98% of subjects are younger than 30 years. The median age is 21 years. Across all subject pools, 48% were female; the subject pool average range is from 27% to 65% females.

% middleclass is based on participants' personal judgment whether their family's income at the age of 16 was 'substantially below average', 'below average', 'average', 'somewhat above average', 'above average', or 'far above average'. As a proxy for socio-economic status the dummy variable 'middleclass' takes the value one if the subject answered at least 'average' and zero otherwise. Another reason to control for socio-economic status is evidence (in US samples) suggesting that higher social class increases unethical behaviour⁵⁰. Across all subject pools 81% of our subjects have a middle class background; the subject pool average range is from 60% to 95%.

% urban. The dummy '*urban*' is based on a question about the size of the city where participants had spent most of their lives. It takes the value of one if the city reported had at least 10'000 inhabitants and zero otherwise. This is intended to capture the degree of social anonymity subjects are typically used to. It is likely that social control is lower in large cities⁵¹ and this may matter for honesty. About half of our subjects (51%) grew up in a city with a size of 10'000 inhabitants or more; the subject pool average range is from 10% to 82%.

% economics students. We also control for the field of study of our subjects; in particular, we are interested in whether students of economics as a group behave any different from students of other fields^{52,53}. We therefore elicited the field of study and create a dummy variable for economics. The variable *% economics students* shows that 17% of our subjects were economists; the subject pool average range is from 0% to 62%.

% religious. We also asked subjects how religious they are (on a 7-point scale, ranging from 1=not at all, to 7=very religious). We include this variable because there is evidence

that religiosity matters for honesty^{29,54,55} although there are good arguments why such a link is not straightforward and may not be robust.⁵⁶⁻⁵⁸ We classify a subject as religious if he or she answers at least at 4 and this gives us our variable % *religious* (in a subset of sessions in Morocco and Turkey participants could give a binary answer only). The average is 46% and the subject pool average range is from 12% to 92%.

av. % known in a session. Through our recruitment we aimed at maximising anonymity across subjects. We controlled for the degree of anonymity we actually achieved by asking subjects at the end of the experiment how many other participants they had known. On average, subjects knew 11% of other participants; the typical average range is from 1% to 21%; in one very small university 50% knew at least one other participant.

Beliefs in the fairness of others and individual norms of honesty

In our analysis we not only control for the impact of socio-demographics for honesty in our experimental task but also include two variables that measure individual beliefs about the perceived fairness of others and individual norms of honesty.

Belief in fairness of others. To capture subjects' belief in the fairness of others we asked the World Values Survey fairness question: “Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair?” The binary variable takes the value zero for the answer “People would try to take advantage of you” and one for the answer “People would try to be fair”. The subject pool average range in our sample is from 29% to 82% and the average is 52%.

The variable ‘*Belief in fairness of others*’ is interesting because there is evidence that perceived unfairness increases dishonesty³³. Moreover, arguably this question measures beliefs about the cooperativeness of other people⁵⁹. In Extended Data Table 2 (column 3) we show that individual beliefs influence honesty in the sense that people who believe others are fair are less likely to report the income maximising number 5. While this is a correlation, recent research on the role of beliefs for social decision-making^{60,61} suggests that beliefs about the fairness of others can be causal for honesty.

Individual norm of honesty. This variable is based on our subjects' response to three questions on whether a particular act of dishonesty can be justified or not. This question is interesting because there is evidence that "moral firmness", that is, stronger moral convictions, increases honesty²⁹. The statements, which we took from the World Values Survey (WVS, <http://www.worldvaluessurvey.org/>) are: (i) "Claiming government benefits to which you are not entitled", (ii) "Avoiding a fare on public transport", and (iii) "Cheating on taxes if you have a chance". The subjects answered on a 10-point scale between "Never justifiable (=1) and "Always justifiable (=10). For each individual we took the average over the three questions and rescaled so that the value of 1 denotes low norms of honesty (always justifiable) and 10 high norms (never justifiable). The average in our sample is 8.1 with a subject pool average range from 7.0 to 9.5.

2. Supplementary Analyses

2.1 Supplementary Analyses to Fig. 1 and Extended Data Fig. 2

To compare actual data with the Full Honesty, Full Dishonesty and the Justified Dishonesty benchmarks, we used the Kolmogorov-Smirnov test for discrete data⁶², which accounts for the fact that our data are discrete, not continuous.

When we pool the data from the subject pools from low PRV countries and high PRV countries, respectively, we get the following distribution of claims of 0 to 5, respectively, which we can compare statistically to the *Justified Dishonesty* benchmark (** $P < 0.01$, ** $P < 0.05$, * $P < 0.10$, two-sided binomial tests):

- *Justified Dishonesty* (0 to 5, resp): 2.8%, 8.3%, 13.9%, 19.4%, 25%, 30.6%.
- Low PRV countries (1211 subjects; 14 countries):
 - In absolute numbers (0 to 5, resp): 120, 139, 127, 199, 287, 339.
 - In percent (0 to 5, resp): 9.9***, 11.5***, 10.5***, 16.4***, 23.7, 28.0**.
- High PRV countries (1357 subjects; 9 countries):
 - In absolute numbers (0 to 5, resp): 77, 95, 145, 207, 381, 452.
 - In percent (0 to 5, resp): 5.7***, 7.0*, 10.7***, 15.3***, 28.1**, 33.3**.

The low and high distributions are strongly and highly significantly different from one another ($\chi^2(5) = 40.21$, $P = 0.0000$). These data are the basis for Extended Data Fig. 2b and the statistical analyses surrounding it in the main text. Here, we report two more complementary analyses supporting the claims made in the text about the difference between subject pools coming from low and high PRV countries.

The first test concerns the *difference in the distance from the Justified Dishonesty benchmark CDF for low and high PRV countries*, respectively. As test statistic, we use Kolmogorov-Smirnov's d (reported as KSD in Extended Data Fig. 2a). d is defined as the maximal absolute difference between the empirical CDF and a benchmark CDF⁶³, in our case the CDF of Justified Dishonesty. For the test of differences we use the signed d , which is positive if the empirical CDF tends to lie above the benchmark CDF and negative otherwise (except for three subject pools (Shanghai, Meknes, Dar es Salaam), all d values are positive). The average $d_{low} = 0.1101$ and $d_{high} = 0.0143$ and the difference is significant ($z = 2.772$, $P = 0.0056$, two-sided rank sum test). We conclude that the CDFs

from low PRV countries tend to be above the Justified Dishonesty benchmark (and also above the CDFs from high PRV countries), whereas the CDFs from high PRV countries tend to be close to (or below of) the Justified Dishonesty benchmark CDF. Signed d is also significantly negatively correlated with PRV (Spearman's $\rho = -0.61$, $P = 0.0021$).

A second test concerns a *difference in an index of concentration of the claims made by subjects from low and high PRV countries*, respectively. As a summary statistic for each subject pool we calculate Simpson's Index of concentration⁶⁴, which is the sum of the squared frequencies of each of the possible six outcomes of the die roll. The Simpson Index $\Sigma = 0.167$ for the Full Honesty benchmark, where all six claims are equally likely; $\Sigma = 0.221$ for the Justified Dishonesty benchmark; and $\Sigma = 1$ for the Full Dishonesty benchmark. The more concentrated the claimed payments are on particular (typically high) claims the higher is Σ . For our subject pools, the average Σ is 0.222 and ranges from 0.180 to 0.317. Averages for subject pools from low and high PRV countries, resp., are $\Sigma_{low} = 0.206$ and $\Sigma_{high} = 0.248$. The difference is significant according to a two-sided Mann-Whitney test ($z = 3.276$; $P = 0.001$). This provides further support for the observation of Fig. 1 and Extended Data Fig. 2b that claims are more equally distributed in low PRV countries (that is, closer to the Full Honesty benchmark) than in high PRV countries. Σ is significantly correlated with PRV ($\rho = 0.65$, $P < 0.001$).

2.2 Supplementary regression analyses to Extended Data Table 2 (Societal and individual determinants of dishonesty)

Extended Data Table 2 reports the regression results including the coefficients of the socio-economic controls. We chose these controls because they are standard or there is evidence that they matter for honesty. As robustness checks we run fixed effects regressions and find qualitatively very similar estimates for the variables varying within country. Hausman tests (without clustered errors) indicate that the coefficients are not significantly different between fixed and random effects specification for all three models with binary dependent variables. In case of 'Claim' as the dependent variable the Hausman test is significant. However, also in this model significant levels remain the same for all coefficients. We also conducted the Breusch-Pagan Lagrange Multiplier test to differentiate between random effects regression and pooled OLS. It does not reject the

hypothesis that the variance of the unobserved country effect is zero for the models in columns 1, 2, and 4, and only weakly rejects for the model of column 3.

As a further robustness check we estimated non-linear models. Column 1 in Table S2 reports ordered probit regressions, and Columns 2, 3 and 4 report marginal effects from probit regressions. The marginal effects are similar in magnitude to the OLS regressions reported in Extended Data Table 2; also all significance levels remain the same.

	(1) Claim	(2) High Claim (Number 3, 4, or 5)	(3) Highest Claim (Number 5)	(4) No Claim (Number 6)
PRV in 2003	0.070*** (0.023)	0.029*** (0.007)	0.012 (0.010)	-0.015*** (0.005)
Individual norms of honesty	-0.038*** (0.013)	-0.013*** (0.005)	-0.014** (0.006)	0.003 (0.003)
Individual beliefs in fairness of others	-0.070 (0.054)	-0.012 (0.031)	-0.050** (0.021)	-0.005 (0.009)
Age	-0.001 (0.008)	-0.002 (0.003)	0.003 (0.004)	0.002 (0.001)
Female	-0.073* (0.039)	-0.020 (0.016)	-0.019 (0.020)	0.014 (0.012)
Middleclass	-0.032 (0.065)	-0.022 (0.032)	-0.001 (0.023)	0.003 (0.019)
Urban	-0.030 (0.034)	-0.027* (0.016)	-0.013 (0.015)	-0.006 (0.013)
Economics	0.062 (0.069)	0.044 (0.029)	-0.008 (0.032)	-0.023 (0.016)
Religious	-0.018 (0.062)	-0.029 (0.023)	0.023 (0.024)	0.016 (0.014)
% known in session	0.003 (0.002)	0.001 (0.001)	0.002** (0.001)	0.000 (0.001)
<i>N</i>	2284	2284	2284	2284
pseudo R^2	0.006	0.015	0.012	0.018

Table S2 | Probit regression analysis of societal and individual determinants of dishonesty. Displayed are marginal effects of an ordered probit regression (column 1) and from probit regressions (columns 2, 3, 4). Bootstrapped standard errors adjusted for clusters on country level are in parentheses. Dependent variables are the claimed amount (between 0 and 5; column 1), and the binary variables whether a high claim was made (column 2); whether the highest claim (reporting 5) was made (column 3); or whether no claim (number 6) was made (column 4). The regressions use 2284 observations. Data from Spain is missing because due to a technical problem we do not have data on 'belief in fairness'. In some Polish sessions ($n=50$) 'Religious' and '%-known' is missing, and questionnaire data is missing for one Guatemalan and one Chinese session. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

2.3 Robustness checks: association between intrinsic honesty and PRV, institutional quality, and cultural indicators

Analysis by PRV sub-indicators

The PRV sub-indicators ‘Control of Corruption’ and ‘Political Rights’ individually correlate with our experimental measure for dishonesty: mean claims are higher in countries with low control of corruption (Spearman’s $\rho = -0.65$, $P = 0.0009$; Table S3) and less political rights ($\rho = -0.67$, $P = 0.0004$). The correlation for ‘Shadow Economy’ ($\rho = 0.34$, $P = 0.1102$) has the expected sign but it is marginally insignificant.

The regressions in Table S4 (columns 1, 2, 3) show that this finding is robust to the inclusion of socio-economic controls. The results are similar (to our findings in the main text) when we focus on the other experimental measures of intrinsic (dis)honesty. Spearman tests show that High Claims are significantly correlated with the PRV sub-indicators Control of Corruption ($P = 0.0002$), Political Rights ($P < 0.0001$) and weakly with Shadow Economy ($P = 0.0580$); No Claim is significantly correlated with Control of Corruption ($P = 0.0158$) as well as Political Rights ($P = 0.0220$), while there is no correlation with Highest Claim. Regression results are available upon request.

Stability of sub-indicators over time

The sub-indicators change only slowly over time. The Spearman correlation between Control of Corruption in 1996 and 2012 is 0.85; for Shadow Economy between 1999 and 2007 it is 0.99, and for Political Rights in 2003 and 2012 it is 0.94.

Unsurprisingly, therefore, correlations between our experimental measure of intrinsic honesty (conducted between 2011 and 2015) and the sub-indicators of PRV measured at different points in time do not vary much. For example, the Spearman correlation between Control of Corruption in 1996 and Claim in our experiment is -0.60 ($P = 0.0023$). This is robust to regression estimates with socio-economic controls (see Table S3). In 1996 most of our participants were around 5 to 6 years old and therefore in no position to influence the measurement of corruption in their society. This underscores that subjects have been exposed to a rather stable environment with a particular level of PRV for a long period of time.

	(1) Claim	(2) High Claim (Number 3, 4, or 5)	(3) Highest Claim (Number 5)	(4) No Claim (Number 6)
Control of Corruption in 1996	-0.139*** (0.038)	-0.039*** (0.009)	-0.012 (0.012)	0.020*** (0.006)
Individual norms of honesty	-0.055*** (0.018)	-0.013*** (0.004)	-0.014** (0.006)	0.003 (0.002)
Individual beliefs in fairness of others	-0.073 (0.086)	-0.011 (0.031)	-0.051** (0.021)	-0.004 (0.009)
Constant	4.095*** (0.315)	0.932*** (0.072)	0.375*** (0.111)	-0.008 (0.044)
Socio-demographic controls	Chi ² (7)=13.14*	Chi ² (7)=16.41**	Chi ² (7)=7.28	Chi ² (7)=10.08
<i>N</i>	2284	2284	2284	2284
<i>R</i> ²	0.021	0.018	0.014	0.009

Table S3 | Regression analysis using Control of Corruption in 1996 instead of PRV. Bootstrapped standard errors adjusted for clusters on country level are in parentheses. Dependent variables are the claimed amount (between 0 and 5; column 1), and the binary variables whether a high number was reported (column 2); whether the highest number was reported (column 3); or whether number 6 was reported (column 4). The regressions contain 2284 observations. Data from Spain are missing because due to a technical problem we do not have data on ‘*belief in fairness*’. In some Polish sessions ($n = 50$) ‘*Religious*’ and ‘*%-known*’ is missing, and questionnaire data is missing for one Guatemalan and one Chinese session. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

Dishonesty and economic, institutional, and cultural indicators

In the Extended Data Figs. 3 and 4 we demonstrate the association between institutional and cultural indicators and intrinsic honesty. Table S4 focuses on those associations in a regression analysis, when controlling for several socio-economic variables and individual norms of honesty and beliefs in fairness of others. Table S4 shows that these associations are robust to the inclusion of those controls. Again individual norms of honesty are predictive of the claims. Socio-demographic variables are included in all regressions but are jointly significantly different from zero (at $P < 0.1$) only in model 2 (Chi²(7) = 18.9, $P < 0.008$), model 9 (Chi²(7) = 25.0, $P < 0.001$) and model 10 (Chi²(7) = 20.2, $P = 0.005$).

	(1) Claim	(2) Claim	(3) Claim	(4) Claim	(5) Claim	(6) Claim	(8) Claim	(9) Claim	(10) Claim
Control of corruption in 2003	-0.161*** (0.039)								
Shadow economy in 2003		0.006 (0.004)							
Political rights in 2003			-0.013*** (0.004)						
Constraint on Executive 1990 to 2000				-0.097*** (0.029)					
Gov. Effectiveness in 2000					-0.171*** (0.044)				
GDP per capita (in \$1000, 1990 to 2000)						-0.022*** (0.006)			
Individualism							-0.006*** (0.002)		
Traditional vs. secular-rational values								-0.100* (0.053)	
Survival vs. self-expression values									-0.151*** (0.048)
Individual norms of honesty	-0.057*** (0.018)	-0.046** (0.021)	-0.058*** (0.017)	-0.058*** (0.017)	-0.056*** (0.018)	-0.056*** (0.017)	-0.061*** (0.017)	-0.043* (0.023)	-0.057*** (0.018)
Individual belief in fairness of others	-0.074 (0.085)	-0.080 (0.085)	-0.120 (0.081)	-0.110 (0.080)	-0.079 (0.086)	-0.089 (0.083)	-0.096 (0.088)	-0.073 (0.092)	-0.086 (0.088)
Constant	4.122*** (0.301)	3.815*** (0.379)	4.304*** (0.329)	4.551*** (0.296)	4.122*** (0.314)	4.137*** (0.295)	4.336*** (0.297)	3.880*** (0.380)	4.031*** (0.327)
Socio-Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>N</i>	2284	2284	2284	2284	2284	2284	2193	2192	2192
<i>R</i> ²	0.022	0.015	0.021	0.022	0.021	0.022	0.021	0.017	0.020

Table S4 | Regression analysis of macro-level indicators and Claims. Dependent variable is claims (payout). Displayed are the coefficients from OLS regressions of macro-level indicators on claims with robust standard errors clustered on countries. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

2.4 Supplementary regression analyses to Extended Data Table 3 (Institutional and Cultural Determinants of PRV)

The Extended Data Table 3 reports determinants of PRV. In this section we give a more detailed account on the estimation approach. In an effort to shed light on causality we employ instrumental variables estimation following the economic literature that focuses on cultural and historic or ‘deep’ factors that influence economic prosperity and institutional quality. In line with the economic literature, our main focus is on two explanatory variables: *individualism-collectivism* and *institutional quality*.

Individualism-collectivism and PRV. When estimating the effect of individualism (as a proxy for limited morality in collectivist societies, see Section 1.1 ‘Cultural Indicators’)

on PRV we have to take reverse causality or omitted variables into consideration (i.e., a high prevalence of rule violations strengthens family ties and collectivism since individuals cannot rely on the rule of law). We use two instruments for Individualism, grammatical rule, and genetic distance.

Our first instrumental variable estimation strategy follows Licht et al.¹⁸ and Tabellini⁸ and uses *grammatical rules* concerning the use of pronouns as instruments for the cultural trait that is associated with limited morality (see Extended Data Table 3, column (8) and (10)). The underlying assumptions are that (i) there is a link between linguistic rules and deep features of culture and (ii) the instrument is not correlated with the error term in the explanatory equation (conditional on other control variables). The data on linguistic rules are taken from Tabellini⁸ (data retrieved from <http://didattica.unibocconi.it/mypage/index.php?IdUte=48805&idr=13301&lingua=ita>). The variable is based on two grammatical rules. The first rule governs the use of first and second person pronouns in conversations. Languages like English make the use of subject pronouns obligatory (e.g., ‘I see’), while in other languages, like Spanish, the use is not obligatory (e.g., it is possible to say ‘veo’ or ‘yo veo’). Kashima and Kashima⁶⁵ suggest that these rules reflect the conceptions of the person in different cultures. They argue that languages that forbid dropping the pronoun are typically those that emphasise the individual. The second grammatical rule concerns 2nd person differentiation. This differentiation exists in some languages (like in French the ‘tu’ and ‘vous’) and is associated with a hierarchy of power.

As a second instrument for Individualism we follow Gorodnichenko and Roland¹⁹ and use ‘*genetic distance*’ (Extended Data Table 2, columns (9) and (10)). Genetic distance is an indirect measure of cultural transmission. Parents transmit their genes to their children but also their culture. Genetic distance can therefore be seen as a proxy for very distant migration patterns and with it distant divergence in intergenerationally transmitted cultural traits. It is very unlikely that institutions influence genetic distance. We use the ‘genetic distance’ measure from Spolaore and Wacziarg⁶⁶ (data retrieved from <http://sites.tufts.edu/enricospolaoire/>). We take the genetic distance of each country to the USA, which is the most individualist country in our sample. The measure is based on neutral genetic markers, that is, markers that are not related to evolutionary fitness and

therefore should have no direct effect on behaviour and thus PRV. This identification strategy does not postulate a causal effect between genes and culture. It only exploits the correlation between cultural and genetic transmission. In column (10) of Extended Data Table 2 we use both ‘grammatical rules’ and ‘genetic distance’ as instrument for Individualism. This allows testing for overidentifying restrictions.

Institutional Quality and PRV. The link between institutional quality and PRV is intuitive: strong formal institutions will limit rule violations. Empirical estimations, however, have to take account of reverse causality (a high prevalence of rule violation impairs the functioning of institutions) or omitted variables (e.g. deep cultural factors that lead to low quality formal institutions and a high prevalence of rule violations). Our instrumental variables estimation follows Acemoglu et al.⁶⁷ using settler mortality as an instrument. The idea is that colonisation strategies were different depending on settler mortality. In places with high settler mortality settlers installed extractive institutions (with low protection for private property) with the purpose of extracting and transferring resources of the colony to the coloniser. In places with low settler mortality settlers built inclusive institutions with a strong emphasis on private property and checks against government in power. These differences in institutions persisted after independence and are still reflected in current institutions. The exclusion restriction implied in this estimation strategy is that conditional on the other controls mortality rates have no effect on PRV today other than through their effect on institutions.

Following Tabellini⁸, in all regressions of the Extended Data Table 3 we control for primary education in 1930 (Benavot and Riddle⁶⁸) and legal origins (La Porta et al.⁶⁹). As further controls we include GDP per capita, Government Effectiveness, and Ethnolinguistic Fractionalisation in 1985 (due to Philip Roeder; downloaded from <http://weber.ucsd.edu/~proeder/elf.htm>; accessed 15.06.2015).

Extended Data Table 3, columns (1-6), demonstrate that the coefficients for institutional quality (Constraint on Executive) as well as Individualism are significant and robust to the inclusion of other controls. Better institutions (as proxied by Constraint on Executive) and more individualist societies are associated with less PRV. Remarkably, the measure of past Constraint on Executive (average of 1890 to 1900) is also highly

significantly associated with current PRV (column 2). A potential transmission channel is culture, that is, “those customary beliefs and values that ethnic, religious and social groups transmit fairly unchanged from generation to generation” (Guiso et al.⁷⁰, p. 23). As such, distant events and/or (institutional) environments transmitted via culture can have lasting effects on societies. Our further controls reveal that Primary Education in 1930 (column 3), GDP (column 4), and Government Effectiveness (column 5) are negatively associated with PRV. We do not find an association with Ethnolinguistic Fractionalization and PRV.

Our IV estimations in columns (7-10) suggest that the quality of institutions and individualism causally affect PRV. The coefficient for the quality of institutions (Constraint on Executive) is highly significant when using settler mortality as an instrument (column 7). When using the ‘grammatical rule’ as an instrument for Individualism the coefficient stays weakly significant (column 8), while it stays significant when using ‘genetic distance’ as an instrument (column 9), or using both simultaneously as instruments (column 10). The test for overidentifying restriction in column 10 suggests that the instruments for individualism were correctly excluded.

Robustness checks. As robustness checks we replicated regressions by Acemoglu et al.⁶⁷ and Tabellini⁸ using the exact same indicators and controls as in their original work - the only differences being that we substituted our PRV measure as the dependent variable (they used GDP and bureaucratic quality, resp). Table S5 follows Acemoglu et al.⁶⁷ and uses ‘protection against risk of expropriation’ as a proxy for the quality of institutions (data are taken from <http://economics.mit.edu/faculty/acemoglu/data/ajr2001>). This proxy measures the risk of expropriation of private foreign investment by government, where a higher score means less risk of expropriation. It also controls for continents and latitude (distance to the equator normalized between 0 and 1). Controlling for latitude does not change the significant levels of ‘Average protection against expropriation’ both in the OLS regression and the instrumental variables estimation (columns 1 and 3). Adding continent dummies does not change the significance level in the OLS regression, while it reduces the significance level of ‘Average protection against expropriation’ in the IV

estimation though (column 2 and 4). Thus, overall the impact of the institutional variable ‘Average protection against expropriation’ seems to be quite robust.

	(1) PRV	(2) PRV	(3) PRV IV: Settler Mortality	(4) PRV IV: Settler Mortality
Average protection against expropriation 1985-1995	-0.57*** (0.09)	-0.48*** (0.10)	-0.97*** (0.30)	-0.79* (0.43)
Latitude	-2.79*** (0.80)	-2.07*** (0.76)	-1.14 (1.39)	-1.18 (1.44)
Africa dummy		0.48** (0.23)		0.31 (0.27)
Asia dummy		0.08 (0.26)		0.26 (0.32)
“Other” continent dummy		-1.16*** (0.25)		-0.68 (0.69)
Constant	4.53*** (0.50)	3.66*** (0.60)	6.86*** (1.84)	5.53** (2.61)
<i>N</i>	62	62	62	62
<i>R</i> ²	0.634	0.697	0.460	0.610
<i>1st stage F-stat</i>			9.8***	5.1***

Table S5 | Expropriation Risk and PRV following Acemoglu et al.⁶⁷. Dependent variable is PRV in 2003. The explanatory variables closely follow Acemoglu et al.⁶⁷. The measure for institutions is ‘Average protection against expropriation 1985-1995’. In column (1) and (3) we control for latitude (normalised between 0 and 1) and column (2) and (3) contains continent dummies. In column (3) and (4) we instrumented Expropriation Risk (bold coefficients) with ‘Settler Morality’. Robust standard errors in parentheses. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

In Table S6 we replicate Tabellini’s approach⁸ for our PRV measure. Instead of ‘Individualism’ we use his indicator ‘Trust & Respect’ as the measure for limited morality. The indicator ‘Trust & Respect’ is based on the two WVS question: (1) ‘Generally speaking, would you say that most people can be trusted or that you need to be very careful’ and (2) whether or not people answered that ‘tolerance and respect for other people’ is an important quality for children to learn at home. The other covariates are very similar to Extended Data Table 2. It is apparent from Table S6 that the alternative measure “Trust & Respect” is quite robust as a factor predicting PRV. Further, the other covariates behave very similarly compared to the Extended Data Table 3.

	(1) PRV	(2) PRV	(3) PRV IV: Gram. Rule	(4) PRV IV: Gram. Rule	(5) PRV IV: Gram. Rule
Trust & Respect	-1.80*** (0.33)	-1.18*** (0.23)	-1.63* (0.87)	-1.33** (0.61)	-1.63*** (0.59)
Primary Education in 1930	-0.04*** (0.00)	-0.01* (0.00)	-0.03*** (0.01)	-0.01** (0.00)	-0.02*** (0.00)
Income in 1980-2000		-1.02*** (0.15)		-0.92*** (0.16)	
Constraint on Executive 1960-2000					-0.26*** (0.07)
Constant	0.52* (0.26)	8.53*** (1.17)	0.40 (0.32)	7.74*** (1.23)	1.27*** (0.42)
Control for UK and French Legal origins	Yes	Yes	Yes	Yes	Yes
<i>N</i>	63	63	50	50	50
<i>R</i> ²	0.716	0.827	0.749	0.851	0.801
<i>1st stage F-stat</i>			14.1***	10.7***	11.2***

Table S6 | ‘Trust & Respect’ and PRV following Tabellini⁸. Dependent variable is PRV in 2003. The explanatory variables closely follow Tabellini⁸. Column (1) to (5) contain the variable Primary Education in 1930 and dummy variables for UK and French Legal origin. Column (2) and (4) contain a control for income between 1980-2000 and column (5) a control for institutions (Constraint on Executive 1960-2000). In column (3) to (5) we instrumented ‘Trust & Respect’ (bold coefficients) with grammatical rules on the use of pro-nouns. Robust standard errors in parentheses. * $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$.

2.5 Testing for potential spillovers from preceding experiment

Because our die-in-a-cup experiment is very short we attached it – like Fischbacher and Föllmi-Heusi²⁷ whose procedures we follow closely - at the end of experimental sessions. In all subject pools the die-in-a-cup experiment was conducted after public goods experiments, which follow the procedures of Herrmann et al.⁷¹ (the only exception being Spain where the public goods game had a different parametrisation). First, subjects played a standard four-player linear public goods game for ten periods in a partner matching protocol followed by another ten periods of a four-player public goods game with punishment (composed of the same groups). We refer to these experiments as PG.

Here we check whether there are spillovers from this first experiment on our die-in-a-cup experiment. This is a possibility because of the findings of Peysakhovich and

Rand⁷² and in particular Houser, et al.³³ who find spillovers from being a recipient in a dictator game on subsequent dishonest behaviour in a similar task to ours: subjects are more likely to cheat in reporting the outcome of a coin flip when they received only a very small or no amount from the dictator and they claimed to have been treated unfairly.

Several aspects of our procedures likely mitigate the probability of spillovers:

1. The two experiments are clearly separated. Like Fischbacher and Föllmi-Heusi²⁷, our experiment is administered by telling subjects that they get additional money for filling in a questionnaire. Furthermore, while participants were told their total point earnings in each round of the PG, they did not receive feedback on their actual total earnings from the whole PG before the die-in-a-cup experiment (and never received feedback about earnings in their local currency before the die-in-a-cup experiment).
2. Compared to the one-shot dictator game in Houser, et al.³³, our repeated PG with and without punishment is more complex. One important difference is that participants are in a symmetric situation taking repeated choices. Thus, they are not the sole receivers at the discretion of others - payoffs partly depend on their own choices and they can react to the behaviour of others. In case of the standard PG (without punishment) they can adjust their contribution in all but the last round, while in the PG with punishment they can also punish others. Being able to take matters in their own hands by choosing contribution levels and retaliation likely reduces the need to compensate (due to a feeling of unfairness) in the die-rolling experiment.

Even though our procedures and design make spillovers unlikely, in this section we are testing for them. We follow two approaches: (i) *exogenous variation of preceding experiments*: in several countries we tested whether stand-alone or different previous experiments lead to systematically different reporting behaviour; (ii) *regression approach*: we tested whether individuals' experience in the PG is predictive of subsequent behaviour in the experiment on intrinsic honesty.

Exogenous variation of preceding experiments

The countries where we varied the preceding experiments are Austria (a risk task; $n = 21$); China (social learning; $n = 118$); Guatemala (stand alone; $n = 94$); Morocco (other social preferences; $n = 79$); Poland (other social preferences; $n = 50$); Sweden (stand-alone, $n = 26$), Turkey (stand-alone, $n = 83$; and other social preferences, $n = 95$) and United Kingdom (other social preferences I, $n = 75$; other social preferences II, $n = 26$; and stand-alone, $n = 58$). These countries cover a wide variation regarding average payment reported (UK second lowest, Morocco second highest).

The cleanest test for spillovers is to see whether behaviour without exposure to a previous experiment is different from behaviour after having experienced the PG. In four countries (Guatemala, Sweden, Turkey and UK) we conducted stand-alone experiments. The rank sum test in each country does not reject the hypothesis that the two samples (stand-alone vs after PG) have the same distribution. The P -value is 0.8841 in the United Kingdom, 0.4158 in Turkey, 0.2229 in Sweden and 0.2396 in Guatemala. The Fisher exact test also does not reject the hypothesis that they are from the same distribution (UK: $P = 0.3010$; Turkey: $P = 0.1820$; Sweden: $P = 0.500$; Guatemala: $P = 0.7770$).

Likewise we do not find significant differences when applying the rank sum test in the case of Austria ($P = 0.3903$); Morocco ($P = 0.2949$); and Poland ($P = 0.6562$). The same is the case for the Fisher exact test (Austria: $P = 0.346$; Morocco: $P = 0.2820$; Poland: $P = 0.455$). For China, even though the means are similar (3.66 MU after PG vs. 3.43 MU after Social Learning), the rank sum test is weakly significant ($P = 0.0798$).

Lastly, we apply a Kruskal-Wallis test in the case of United Kingdom, where we have four conditions and Turkey, where we have three conditions. In the case of Turkey the Kruskal-Wallis test does not reject the hypothesis that the data is drawn from the same distribution ($P = 0.2614$), while in the United Kingdom $P = 0.0976$. Overall these findings show that spillovers are unlikely.

Regression Analysis

To shed further light on potential spillovers we conduct a regression analysis. This allows us to check whether our main result, namely that PRV is predictive of intrinsic honesty, holds when we control for several variables that captures a person's experience

in the previous PG. The results are reported in Table S7, columns (1)-(3). It is apparent that the PRV is a highly significant predictor for reported payments even after controlling for previous experiences in the PG.

Having established that our main results holds when controlling for previous experience, in what follows we explain our regression approach in more detail and discuss how and if we expect to see correlations between experience in the PG and intrinsic honesty. Correlations between the two experiments can be either due to (i) experiences made in the previous experiment (spillovers) or (ii) underlying individual traits or (firm) beliefs unaffected by the experience in the PG. Two considerations mark the starting point for our analysis.

- (i) *Spillovers*: Following Houser et al.³³, one may expect potential spillovers to likely manifest themselves as follows: Participants, who contributed more to the PG than their members may feel treated unfairly by others' free riding. Thus the hypothesis is that – if there are spillovers – we expect individuals with higher relative contributions (conditional on others' average contributions) to lie more. Another potential source of feeling treated unfairly may work through received punishment – which we control in our regression as well.

For several reasons the last period is particularly interesting: Except for the last period(s) contributions in the repeated PG game may be driven by strategic reputational concerns. Thus, cooperative subjects may experience a higher fraction of low contributions in the last period. Further, the very last period is also the period where participants have no option to retaliate (e.g., cannot decrease contribution or, in the PG with punishment, cannot counter-punish in the next period). Lastly, compared to the early periods the last period in the PG with punishment may be more salient in memory. One can therefore expect spillovers to be more important in the last period compared to earlier periods.

- (ii) *Underlying traits*: The first decision individuals take is unaffected by any previous experimental experiences. As such it constitutes a measure for cooperative attitudes. Finding a positive correlation between lying and cooperative choices in the very first period suggests that this correlation is driven by more fundamental underlying

preferences. For example, there may be a ‘moral’ type, who has strong preference for cooperation *and* honesty. Thus, in contrast to spillovers (where we expect a negative correlation) underlying traits predict a positive correlation between contributions in the PG and honesty. This is especially the case for the very first period.

Nevertheless, initial choices may lead to systematically different experiences during the course of the experiment. In our regressions we control for several factors participants experience during the course of the PG. This and the fact that the first period is the farthest past in memory, makes it unlikely that the coefficient of the first period contributions capture experiences in the PG.

Table S7 reports the results of ordered probit regressions with Claim as the dependent variable. Our starting point is individual (average per period) earnings in the PG (column 1 controlling for PRV and column 4). Even though this is only a coarse measure for experience in the PG it allows us to see whether there are income effects. There is a weakly significant effect of income in the PG on dishonesty when controlling for a country’s PRV (column 1), while we find no effect without the PRV control (column 2). This effect is positive – contrary what is predicted if subjects would use the die-rolling-experiment to compensate for low earnings in the PG.

To get more detailed insights we include several variables capturing behaviour and experiences in the PG. (The first part refers to variables capturing experiences in the PG without punishment while the second part to those in the PG with punishment.) To describe the PG without punishment (first part) we include subjects’ very first decisions (‘Contribution Period 1’). We also control for the three other group members’ average contributions in period 1 of the PG without punishment (‘Others Av. Contr. Period 1’). Further controls for the standard PG are the individual’s average contributions in periods 2 to 10 (‘Av. Contribution in Period 2 to 10’) and the three group members’ average contributions in these periods (‘Others Av. Contr. in Period 2 to 10’). Similarly, in the PG with punishment we control for the individual’s average contributions (‘Average Contr. (other Periods)’ and the average contributions of the group members (‘Others Av. Contr. (other Periods)’). In columns 2 and 5 ‘Others Av. Contr. (other Periods)’ is the average contributions over all periods, while for column 3 and 6 it is the average of periods 1 to 9.

Additionally, we also control for the average amount of pro-social punishment ('Av. Received Pro-Pun (other periods)') and anti-social punishment ('Av. Received anti-Pun (other periods)') received.

It is apparent from table S7 that none of the variables capturing experiences and behaviour in the PG are significant when controlling for countries PRV (column 2 and 3). For example, others' behaviour in the first period of the standard PG does not have a significant impact on the number reported in the die-in-a-cup experiment. Neither do average contributions (in both the standard PG and the one with punishment) as well as pro- or antisocial punishment received. In column 3 we estimate a model where we additionally include variables that capture behaviour and experiences in the last period of the PG with punishment. The last period in the PG with punishment is the freshest in memory; strategic, reputational incentives are absent; and subjects cannot react to behaviour of others in future periods. As is apparent from column 3 these additional variables are not significant predictors of Claim. This is evidence that spillovers are not very likely.

In columns (5) and (6) we do not control for country effects. Interestingly, the coefficients for others average contributions in the PG without punishment ('Others Av. Contr. in Period 2-10') are significantly positively related to Claim. Thus, those who experienced high contributions of their other group members are more likely to make a higher Claim. This is in contrast to what spillovers due to fairness concerns (or cooperative norms revealed by others high contributions) predict.

	(1) Claim	(2) Claim	(3) Claim	(4) Claim	(5) Claim	(6) Claim
PRV in 2003	0.078*** (0.027)	0.072*** (0.027)	0.072*** (0.026)			
Average Profit	0.008* (0.004)			-0.002 (0.006)		
PG without punishment						
Contribution Period 1		-0.007 (0.005)	-0.007 (0.005)		-0.008 (0.005)	-0.009 (0.005)
Av. Contribution in Period 2-10		-0.005 (0.008)	-0.005 (0.008)		-0.003 (0.008)	-0.003 (0.008)
Others Av. Contr. Period 1		0.003 (0.008)	0.002 (0.008)		-0.002 (0.008)	-0.002 (0.008)
Others Av. Contr. in Period 2-10		0.006 (0.007)	0.006 (0.007)		0.013** (0.006)	0.012* (0.007)
PG with Punishment						
Average Contr. (other Periods)		-0.006 (0.009)	-0.002 (0.011)		-0.009 (0.010)	-0.005 (0.012)
Contribution Period 10			-0.005 (0.006)			-0.005 (0.006)
Others Av. Contr. (other Periods)		0.009 (0.010)	0.011 (0.010)		0.001 (0.008)	0.002 (0.010)
Others Av. Contr. Period 10			-0.001 (0.008)			0.001 (0.008)
Av. Received pro-pun, (other periods)		-0.004 (0.020)	0.001 (0.023)		-0.001 (0.020)	0.003 (0.024)
Received Pro-Pun., Period 10			-0.007 (0.012)			-0.007 (0.011)
Av. Received Anti-Pun, (other periods)		-0.006 (0.028)	-0.005 (0.033)		0.004 (0.032)	0.004 (0.037)
Received Anti-Pun., Period 10			0.000 (0.014)			0.001 (0.014)
<i>N</i>	1741	1741	1741	1741	1741	1741
<i>Wald Chi²</i>	11.42***	42.7***	101.72***	0.10	22.59***	59.81***
<i>Pseudo R²</i>	0.003	0.004	0.004	0.0000	0.002	0.002

Table S7 | Relation of experience in PG and behaviour in the die-in-a-cup experiment. Dependent variable is Claim in the die-rolling experiment. Data from Spain is missing because the parametrization of the PG was different. Bootstrapped standard errors adjusted for 22 clusters in countries are reported in parentheses (* $P < 0.10$, ** $P < 0.05$, *** $P < 0.01$).

3. References

1. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behavioral and Brain Sciences* **33**, 61 (2010).
2. Kaufmann, D., Kraay, A. & Mastruzzi, M. The worldwide governance indicators: Methodology and analytical issues. *Hague Journal on the Rule of Law* **3**, 220 (2011).
3. Schneider, F. & Enste, D. H. Shadow economies: Size, causes, and consequences. *Journal of Economic Literature* **38**, 77 (2000).
4. Feld, L. P. & Schneider, F. Survey on the shadow economy and undeclared earnings in oecd countries. *German Economic Review* **11**, 109 (2010).
5. Buehn, A. & Schneider, F. Shadow economies around the world: Novel insights, accepted knowledge, and new estimates. *International Tax and Public Finance* **19**, 139 (2012).
6. Everitt, B. S., *Making sense of statistics in psychology*. (Oxford University Press, Oxford, 1996).
7. Acemoglu, D. & Johnson, S. Unbundling institutions. *Journal of Political Economy* **113**, 949 (2005).
8. Tabellini, G. Institutions and culture. *Journal of the European Economic Association* **6**, 255 (2008).
9. Hruschka, D. J. & Henrich, J. Institutions, parasites and the persistence of in-group preferences. *PLoS ONE* **8**, e63642 (2013).
10. Hruschka, D. et al. Impartial institutions, pathogen stress and the expanding social network. *Human Nature* **25**, 567 (2014).
11. Hofstede, G., *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. (Sage, Thousand Oaks, CA, 2001).
12. Hofstede, G., Hofstede, G. J. & Minkov, M., *Cultures and organizations: Software of the mind. 3rd ed.* (McGraw-Hill USA, New York, 2010).
13. Greif, A. Cultural beliefs and the organization of society: A historical reflection on collectivist and individualist societies. *Journal of Political Economy* **102**, 912 (1994).
14. Banfield, E. C., *The moral basis of a backward society*. (The Free Press, Glencoe, IL, 1958).
15. Platteau, J.-P., *Institutions, social norms, and economic development*. (Harwood Academic Publishers, Amsterdam, 2000).
16. Tabellini, G. The scope of cooperation: Values and incentives. *Quarterly Journal of Economics* **123**, 905 (2008).
17. Mazar, N. & Aggarwal, P. Greasing the palm: Can collectivism promote bribery? *Psychological Science* **22**, 843 (2011).
18. Licht, A. N., Goldschmidt, C. & Schwartz, S. H. Culture rules: The foundations of the rule of law and other norms of governance. *Journal of Comparative Economics* **35**, 659 (2007).
19. Gorodnichenko, Y. & Roland, G. Individualism, innovation, and long-run growth. *Proceedings of the National Academy of Sciences* **108**, 21316 (2011).
20. Mungiu-Pippidi, A. Corruption: Good governance powers innovation. *Nature* **518**, 295 (2015).

21. Gorodnichenko, Y. & Roland, G. Which dimensions of culture matter for long-run growth? *American Economic Review* **101**, 492 (2011).
22. Inglehart, R., *Modernization and postmodernization: Cultural, economic and political change in 43 societies*. (Princeton University Press, Princeton, NJ, 1997).
23. Inglehart, R. & Baker, W. E. Modernization, cultural change, and the persistence of traditional values. *American Sociological Review* **65**, 19 (2000).
24. Inglehart, R. & Norris, P., *Rising tide: Gender equality and cultural change around the world*. (Cambridge University Press, Cambridge, 2003).
25. Inglehart, R. & Welzel, C., *Modernization, cultural change, and democracy. The human development sequence*. (Cambridge University Press, Cambridge, 2005).
26. Shariff, A. F. & Rhemtulla, M. Divergent effects of beliefs in heaven and hell on national crime rates. *PLoS ONE* **7**, e39048 (2012).
27. Fischbacher, U. & Föllmi-Heusi, F. Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association* **11**, 525 (2013).
28. Shalvi, S., Dana, J., Handgraaf, M. J. J. & De Dreu, C. K. W. Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes* **115**, 181 (2011).
29. Shalvi, S. & Leiser, D. Moral firmness. *Journal of Economic Behavior & Organization* **93**, 400 (2013).
30. Weisel, O. & Shalvi, S. The collaborative roots of corruption. *Proceedings of the National Academy of Sciences* **112**, 10651 (2015).
31. Hugh-Jones, D., Honesty and beliefs about honesty in 15 countries. *School of Economics Working Paper 2015-01*, University of East Anglia, (2015); available at <https://www.uea.ac.uk/economics/working-papers-series>.
32. Bucciol, A. & Piovesan, M. Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology* **32**, 73 (2011).
33. Houser, D., Vetter, S. & Winter, J. Fairness and cheating. *European Economic Review* **56**, 1645 (2012).
34. Abeler, J., Becker, A. & Falk, A. Representative evidence on lying costs. *Journal of Public Economics* **113**, 96 (2014).
35. Pascual-Ezama, D. et al. Context-dependent cheating: Experimental evidence from 16 countries. *Journal of Economic Behavior & Organization* **116**, 379 (2015).
36. Cohn, A., Fehr, E. & Marechal, M. A. Business culture and dishonesty in the banking industry. *Nature* **516**, 86 (2014).
37. Gneezy, U. Deception: The role of consequences. *American Economic Review* **95**, 384 (2005).
38. Mazar, N., Amir, O. & Ariely, D. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research* **45**, 633 (2008).
39. Lundquist, T., Ellingsen, T., Gribbe, E. & Johannesson, M. The aversion to lying. *Journal of Economic Behavior & Organization* **70**, 81 (2009).
40. Erat, S. & Gneezy, U. White lies. *Management Science* **58**, 723 (2012).
41. Rosenbaum, S. M., Billinger, S. & Stieglitz, N. Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology* **45**, 181 (2014).

42. Fischbacher, U. Z-tree: Zurich toolbox for readymade economic experiments. *Experimental Economics* **10**, 171 (2007).
43. Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M. & Zamir, S. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review*. **81**, 1068 (1991).
44. Herrmann, B., Thöni, C. & Gächter, S. Supporting online material for antisocial punishment across societies. *Science* **319**, available at <http://www.sciencemag.org/cgi/data/319/5868/1362/DC1/1> (2008).
45. Brislin, R. W. Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology* **1**, 185 (1970).
46. Camerer, C. F. & Hogarth, R. M. The effects of financial incentives in experiments: A review and capital-labor-production framework. *J. Risk Uncertain.* **19**, 7 (1999).
47. Kocher, M. G., Martinsson, P. & Visser, M. Does stake size matter for cooperation and punishment? *Economics Letters* **99**, 508 (2008).
48. Fehr, E., Tougareva, E. & Fischbacher, U. Do high stakes and competition undermine fair behaviour? Evidence from Russia. *Journal of Economic Behavior & Organization*, <http://dx.doi.org/10.1016/j.jebo.2013.09.005> (2013).
49. Gächter, S. (Dis)advantages of student subjects: What is your research question? *Behavioral and Brain Sciences* **33**, 92 (2010).
50. Piff, P. K. et al. Higher social class predicts increased unethical behavior. *Proceedings of the National Academy of Sciences* **109**, 4086 (2012).
51. Gächter, S. & Herrmann, B. The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review* **55**, 193 (2011).
52. Carter, J. R. & Irons, M. D. Are economists different, and if so, why? *The Journal of Economic Perspectives* **5**, 171 (1991).
53. Frank, B. & Schulze, G. G. Does economics make citizens corrupt? *Journal of Economic Behavior & Organization* **43**, 101 (2000).
54. Utikal, V. & Fischbacher, U. Disadvantageous lies in individual decisions. *Journal of Economic Behavior & Organization* **85**, 108 (2013).
55. Arbel, Y., Bar-El, R., Siniver, E. & Tobol, Y. Roll a die and tell a lie – what affects honesty? *Journal of Economic Behavior & Organization* **107**, Part A, 153 (2014).
56. Norenzayan, A. Does religion make people moral? *Behaviour* **151**, 365 (2014).
57. Henrich, J. et al. Markets, religion, community size, and the evolution of fairness and punishment. *Science* **327**, 1480 (2010).
58. Shariff, A. F. & Norenzayan, A. Mean gods make good people: Different views of god predict cheating behavior. *International Journal for the Psychology of Religion* **21**, 85 (2011).
59. Thöni, C., Tyran, J.-R. & Wengström, E. Microfoundations of social capital. *Journal of Public Economics* **96**, 635 (2012).
60. Smith, A. Estimating the causal effect of beliefs on contributions in repeated public good games. *Experimental Economics* **16**, 414 (2013).

61. Costa-Gomes, M. A., Huck, S. & Weizsäcker, G. Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect. *Games and Economic Behavior* **88**, 298 (2014).
62. Jann, B. Multinomial goodness-of-fit: Large-sample tests with survey design correction and exact tests for small samples. *The Stata Journal* **8**, 147 (2008).
63. Siegel, S. & Castellan, J. N. J., *Nonparametric statistics for the behavioral sciences*, 2nd. Ed. (MacGraw Hill, New York, 1988).
64. Simpson, E. H. Measurement of diversity. *Nature* **163**, 688 (1949).
65. Kashima, E. S. & Kashima, Y. Culture and language: The case of cultural dimensions and personal pronoun use. *Journal of Cross-Cultural Psychology* **29**, 461 (1998).
66. Spolaore, E. & Wacziarg, R. The diffusion of development. *Quarterly Journal of Economics* **124**, 469 (2009).
67. Acemoglu, D., Johnson, S. & Robinson, J. A. The colonial origins of comparative development: An empirical investigation. *American Economic Review*. **91**, 1369 (2001).
68. Benavot, A. & Riddle, P. The expansion of primary education, 1870-1940: Trends and issues. *Sociology of Education* **61**, 191 (1988).
69. La Porta, R., Lopez-de-Silanes, F. & Shleifer, A. The economic consequences of legal origins. *Journal of Economic Literature* **46**, 285 (2008).
70. Guiso, L., Sapienza, P. & Zingales, L. Does culture affect economic outcomes? *Journal of Economic Perspectives* **20**, 23 (2006).
71. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362 (2008).
72. Peysakhovich, A. & Rand, D. G. Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, in press (2015).

4. Human Subjects Approval

The experiments were approved by the research ethics committee of the Nottingham School of Economics and run with the permission of local institutions.

5. Acknowledgments

The experiment would not have been possible without numerous people who substantially supported this project by facilitating permissions, translations, setting up the laboratories, recruiting and conducting the experiments. We would like to thank Antonio Alonso Aréchar, Benjamin Beranek and Lucas Molleman for setting up and conducting the experiment in Spain, Turkey, China respectively; Monika Büttler and Christian Thöni for facilitating various permissions. In chronological order of places where we conducted the experiments we would like to thank Arti Adji, Ni Made Sukartini, and Bernardinus Purwanto at the Gadjadara University in Indonesia; Camilla Jensen and Daniel Schoch at the University of Nottingham in Malaysia; Hai Anh, Pham Khanh Nam, Thanh Nghi

and Minh Thanh at the Economic University in Ho Chi Minh City, Vietnam; Kristina Berger in Indonesia, Malaysia and Vietnam; Andrés Salamanca Lugo, Olga Manrique and Immanuel Schulz at the Universidad Nacional in Colombia; Giovanna Gusti, Johannes Haushofer, Cynthia Akinyi Ndar and Amos Odero at the BUSARA center in Kenya; Diego Aycinena, Jorge Andres Chang Urrea and Lucas Rentschler at Universidad Francisco Marroquin in Guatemala; Ana Beatriz Aguilar, Swalehe Ally, Lars Ivar Oppedal Berge, Ranveig Falch, Juda Lyamai and Bertil Tungodden for facilitating our experiment in Tanzania; Bianca Bohmer, Justine Burns, Andrew Grand, Kim Ingle and Salma Kagee at the University of Cape Town in South Africa; Benjamin Beranek, Ayla Oğuş Binatlı at the Izmir University of Economics in Turkey; Ana Albert, Wieland Müller, Owen Powell and Jean-Robert Tyran at the University of Vienna in Austria; Antonio Alonso Aréchar, Benjamin Beranek, Felix Kölle, Simone Quercia and Till Weber at the University of Nottingham in the UK; Miroslav Zajíček and Tomas Miklanek at the University of Economics in the Czech Republic; Bert Hoekstra, Aaron Kamm, Ben Loerakker, Jeroen Nieboer, Arthur Schram, Joep Sonnemans and Violet Swakman at the University of Amsterdam and the University of Groningen in the Netherlands; Jeffrey V. Butler, and Gloria Di Caprera at the Einaudi Institute for Economics and Finance, and Francesco Fallucchi, and Veronica Mandaglio in Italy; Tsotne Chikobava, Aleksi Kupradze, and Rati Mekvabishvili at the Tbilisi State University in Georgia; Khadija Amhaouch, Youness Barakat, Hoja Eddomazri, Said Elyounossi, Fatima Lambarraa, and Mostafa Lambarraa at the Ecole Nationale d'Agriculture Meknès in Morocco; Vytautas Butkus, Emilija Timmis, Gytis Valatka and Liutauras Vičkačka at the Vilnius University in Lithuania; Michal Wiktor Krawczyk, Magdalena Smyk and Natalia Starzykowska at the University of Warsaw in Poland; Veronika Chobotová, Tatjana Kluvánková-Oravská, Urban Kovac, and Mari Shioya at the Centre for Transdisciplinary Studies of Institutions, Evolution and Policies in Slovakia; Urs Fischbacher and Jonas Gehrlein at the University of Konstanz in Germany; Lucas Molleman, Jun Feng, Xiangdong Qin and Luqing Yu at the Jiao Tong University in China; Ali Ahmed, Lisa Norrgren, Elsa Söderholm, Hanna Swahnberg, Gustav Tinghög at Linköping University in Sweden.