Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2016/11

Risk and punishment revisited
Errors in variables and in the lab

Christoph Engel
Oliver Kirchkamp

MAX PLANCK SOCIETY


Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2016/11

Risk and punishment revisited
Errors in variables and in the lab

Christoph Engel / Oliver Kirchkamp

July 2016

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Risk and punishment revisited Errors in variables and in the lab*

Christoph Engel† Oliver Kirchkamp‡

9th July 2016

We provide an example for an errors in variables problem which might be often neglected but which is quite common in lab experimental practice: In one task, attitude towards risk is measured, in another task participants behave in a way that can possibly be explained by their risk attitude. How should we deal with inconsistent behaviour in the risk task? Ignoring these observations entails two biases: An errors in variables bias and a selection bias.

We argue that inconsistent observations should be exploited to address the errors in variables problem, which can easily be done within a Bayesian framework.

Keywords: Risk, lab experiment, public good, errors in variables, Bayesian inference.
JEL: C91, D43, L41

1. Introduction

When we run laboratory experiments and when we try to structure the results of these experiments, we sometimes combine two parts of an experiment. In one part of the experiment we elicit individual traits. These traits are used to explain behaviour in another part of the experiment. The way the analysis is done often implicitly assumes that the elicitation of the trait is free of any participant's errors. This presupposes that the trait will play itself out the same way whenever it is elicited and in whichever context it happens to matter. Differential psychology has long cast doubt on this assumption. Traits are unlikely to be stable

*Helpful comments by Ioanna Grypari and André Schmelzer on an earlier version are gratefully acknowledged.

†MPI for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, engel@coll.mpg.de, +49 228 91416-0, +49 228 91416 55.

‡University Jena, School of Economics, Carl-Zeiss-Str. 3, 07737 Jena, oliver@kirchkamp.de, +49 3641 943240, +49 3641 943242.

across situations (Ross, Nisbett and Gladwell, 2011). This suggests that traits will often only be imperfectly observed in post-experimental tests.

For the econometrician the problem of an explanatory variable that is only imperfectly observed is well known as one of "errors in variables". More technically, if we want to estimate $Y = \beta_0 + \beta_1 X + u$, but we can observe X only with an error, e.g. we observe $\xi \sim \mathcal{N}(X, \sigma_\xi)$, then estimating $Y = \beta_0 + \beta_1 \xi + u$ with standard OLS usually fails to provide an unbiased estimator for β_1. Already Adcock (1877) mentions the problem of errors in variables. Since then many authors have discussed this problem (see Gillard, 2010, for an overview). Errors in variables are, indeed, acknowledged in surveys in the field (see, e.g., Kimball, Sahm and Shapiro, 2008, who use survey data on risk tolerance). Deficiencies in the maximum likelihood approach to estimate models with errors in variables were pointed out e.g. by Neyman and Scott (1948) and Solari (1969). Lindley and El-Sayyad (1968) and Florens, Mouchart and Richard (1974) have proposed Bayesian inference to overcome these problems. During the last decades Markov chain Monte Carlo methods have become a powerful and accessible tool for Bayesian inference. Thus, the Bayesian approach lends itself to estimate models with errors in variables.

This brings us to laboratory experiments in economics: Should we worry about errors in variables in the lab? After all, when σ_ξ in the above problem is small, the bias will be small, too. Perhaps the situations we are studying as experimental economists are of the latter kind and the problem is more of academic than of practical interest?

In this paper we study an example which is meant to demonstrate that errors in variables do matter for lab data. The design, we think, is quite typical. In one part of the experiment we attempt to measure an attitude towards risk with the help of a Holt and Laury (2002) task.[1] In another part of the experiment we use this attitude to explain reactions to punishment in a public good game. In the Holt and Laury task, 18% of all participants behave "inconsistently", in that they switch more than once between the lottery with the smaller and the lottery with the larger spread. One of the options mentioned by Holt and Laury (2002) and used by many experimentalists, is to simply drop the data from such participants. We show why this solution can be problematic. We discuss a series of alternatives, and show how a simultaneous estimation of both decision processes, here within a Bayesian framework, offers an easy and effective solution. The simultaneous estimation has two advantages: First, one uses the data from all participants, and thereby avoids a selection bias.[2] Second, since the Holt and Laury task provides us with 10 separate choices per participant, we are also in a position to estimate, separately for each participant, the precision of the measure for her risk attitude. This allows us to address the errors in variables problem.

There are two reasons why a participant has been imprecise: (a) the measure is noisy, e.g. since the participant has been inattentive, or (b) the participant lacks confidence. With just the data from the Holt and Laury task, we cannot disentangle those reasons. But either way, the more the individual estimate of risk aversion is precise, the more it should matter for estimating the effect of risk aversion on choices in the public good. Even if it is not perfectly

[1] As is standard, participants were not admonished to switch at most once.

[2] Otherwise one does not estimate the effect of risk aversion on punishing behavior in the population, but the effect of risk aversion on the punishing behavior of only those individuals whose reactions to risky choices are highly consistent.

precise, it should not be ignored altogether. As our sample demonstrates, results indeed change substantially if one treats the results from the Holt and Laury task as an explanatory variable measured with error.

The remainder of the paper is organized as follows: Section 2 introduces the research question and the design of the experiment from which the data are taken and that we use to illustrate our methodological point. Section 3 discusses alternative methods for dealing with inconsistency in the measurement of risk attitudes. Section 4 uses simulations to assess the size of the bias due to errors in variables in a more general context. Section 5 concludes.

2. Research Question and Design of the Example Experiment

Public goods face individuals with an n-person, continuous action space prisoner's dilemma. Standard theory therefore predicts zero contributions. Through unraveling, this is also the prediction for a repeated game with known end. It is well known that experimental results look different. In a typical group, at the beginning participants on average contribute about half of their endowment. But contributions decay with repetition. Chaudhuri (2011), Ledyard (1995), Zelmer (2003) summarize this literature. Fehr and Gächter (2000) show that if participants are given the possibility to punish each other, at a cost, contributions stabilize at a high level.

The example experiment is interested in understanding how punishment is able to reduce free riding. Arguably, free riders hold standard preferences. In principle they should anticipate sanctions. If the expected value of the sanction is larger than the gain from defection, they should contribute the amount they expect punishers to enforce. Now in a typical public good experiment, punishment is meted out by other participants. Would-be free riders do not know with certainty which norm the punishers will try to enforce, and how severe the sanction will be. This explains why punishment may change the dynamics of the game. In the beginning of the game, participants tempted to free ride must rely on their beliefs. Experience makes it possible to update their expectations about the individually most profitable contribution level. In principle these adjustments should make contribution choices volatile, but should not induce a trend. This is different if punishers condition intervention on the overall cooperativeness in the group.

In the foregoing argument, experienced punishment is just information. Seeing another group member punished should be as important as being punished oneself. Severity is only relevant in a yes or no fashion: if free riding still pays (albeit a bit less), the free rider does not change her behavior; as soon as punishment makes free riding a bad deal, the participant contributes exactly the amount she expects to be enforced. There are several ways of explaining why free riders might exhibit a marginal reaction to severity. A purely cognitive explanation would be: severity informs them about the degree by which punishers are determined to enforce their norm. Severity of experienced punishment would be information about the expected certainty of future punishment. An alternative explanation is motivational. It requires that free riders are not completely selfish. They hold some rudimentary form of social

preferences. Yet on their own, these preferences are not strong enough to counteract the pull of the profit motive. Punishment compensates for the insufficient strength of the social preference. Engel (2014) shows that social preferences may make punishment effective even if its expected value is so low that a perfectly selfish individual would not be deterred. The more severe the sanction, the more likely it is to be strong enough.

Thus far the argument assumes risk neutrality. Empirically, risk preferences are heterogeneous. The majority of a typical experimental population is risk averse. A risk averse subject evaluates the certainty equivalent more positively than a lottery with the same expected value. The subject has a preference for certainty. Consequently, for risk averse participants the information about certainty is even more relevant. They expect to loose even more utility from punishment. This yields the following hypothesis:

Hypothesis 1 *The more a participant is risk averse, the more she increases her contributions to a linear public good after having been punished in the previous period.*

To test this hypothesis, we reanalyze data generated for testing the interplay between social preferences and punishment. The data is taken from Engel (2014). 4 participants $i \in \{1, \ldots, 4\}$ in group k play in each round t a linear public good where profit π_{ikt} is given by (1)

$$\pi_{ikt} = e - c_{ikt} + \mu \sum_{i} c_{ikt} \tag{1}$$

The experiment uses standard parameters with endowment $e = 20$ and marginal per capita rate $\mu = .4$. To each group a fifth participant is randomly assigned. This participant gains a fixed period income of 25 tokens. She has an additional endowment of $20 \cdot \frac{1}{4}$ tokens that she can use to punish any of the active group members. Any token not used for punishment she keeps for herself. The fine to fee ratio is $1 : 12$. After the end of the first round, there is a surprise restart with another 10 rounds of the same game. Participants are rematched every period. Following the procedure that is standard in the experimental literature (see e.g. Charness, 2000; Montero, Sefton and Zhang, 2008) participants are assigned to unannounced matching groups of size 10, to preserve independence.

After the main experiment, a battery of post-experimental tests is administered. For the purposes of this paper, only the test for risk aversion is of interest. The experiment uses the test introduced by Holt and Laury (2002). Choices in the risk task for the 72 participants holding the active role are shown in Figure 1. Vertical reference lines denote participants with inconsistent choices.[3]

The experiment was conducted in the Cologne Laboratory for Economic Research in 2012. The experiment was implemented in zTree (Fischbacher, 2007). Participants were invited using the software ORSEE (Greiner, 2004). Of 90 participants 80 were students of various majors with a mean age 25.4. 44% were female. Participants on average earned 15.11 €

[3] If, for a given participant, a more risky choice (○) is below a safer choice ( ), this participant preferred the risky choice when the probability of the good outcome is small, but not when the probability of the good outcome is large. We call this choice inconsistent. We also call a choice inconsistent if the probability of the good outcome is $p = 1$ but still the lottery with the smaller spread (with the smaller payoff) is preferred over the lottery with the larger spread.

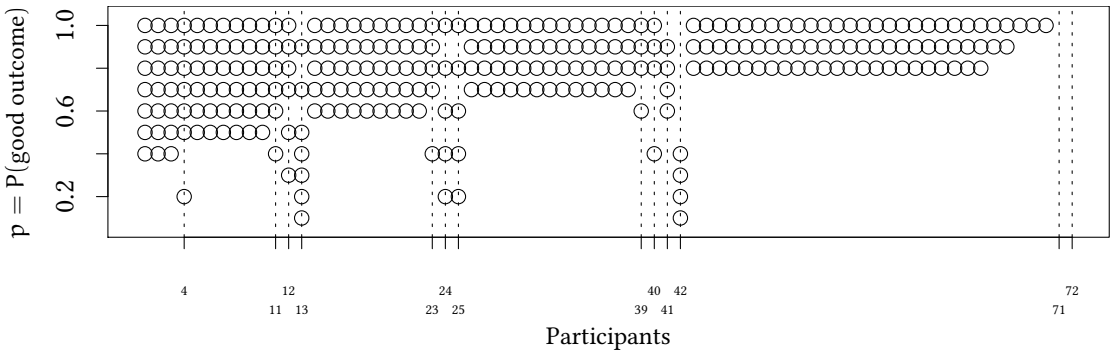

Figure 1: Choices in the risk task

The panel shows choices for each participant: ○ if the participant chose the lottery with the larger spread and nothing if the participant chose the smaller spread. Vertical reference lines denote participants with inconsistent choices (see Footnote 3). Participants are ordered by their risk attitudes, with the more risk seeking participants at the left.

(19.82$ on the days of the experiment), 14.80 € for active players, and 16.38 € for authorities. The experiment had 3 sessions of 30 participants (6 groups of 4 active participants; 6 passive authorities).

3. How to Deal with an Inconsistent Measure for Risk Attitudes?

3.1. The estimation problem

We eventually want to estimate the following model:

$$\Delta c_{ikt} = \beta_0 + \beta_\xi \xi_{ikt} + \beta_p p^c_{ik} + \beta_{\xi\times p}\, \xi_{ikt} \cdot p^c_{ik} + \nu_k + \nu'_{ik} + \epsilon_{ikt} \tag{2}$$

Δc_{ikt} is the change of contribution to the public good of individual i from matching group k at time t. ξ_{ikt} is the total punishment received by individual i from group k at time $t-1$, i.e. the punishment received in the previous period. p^c_{ik} is our measure for risk aversion of individual i from group k. ν_k is a random effect for group k. ν'_{ik} is a random effect for individual i from group k. ϵ_{ikt} is the residual. In line with Hypothesis 1 we expect the interaction term $\beta_{\xi\times p}$ to be positive.

To test our hypothesis, we need for each active participant a reliable measure p^c_{ik} of her risk aversion. This is what we use the test by Holt and Laury for. They design a task where participants choose between a (safe) lottery with a small spread, $p \cdot 2\$ + (1-p) \cdot 1.6\$$, and a (risky) lottery with a large spread, $p \cdot 3.85\$ + (1-p) \cdot .1\$$, where the probability of the good outcome is $p \in \{.1, .2, .3, \ldots, 1\}$. If we assume that preferences for money follow e.g. CRRA, i.e. $u(z) = z^{1-r}$, then the critical value of p^c where participants are indifferent between the more safe and the more risky choice is a monotonic function of their relative risk aversion r.

We can then either describe participants by r or by their critical value of p^c. In the following we will use p^c_{ik} to describe preferences of individual i in group k.[4]

Figure 1 shows behaviour in the risk task. Choices where a participant chose the lottery with the larger spread are denoted with a $\circ$, choices where the participant chose the lottery with the smaller spread are left blank. In the figure we have ordered participants from the most risk loving on the left to the most risk averse on the right. Ideally, we should expect that each participant i in group k can be characterized by a single switching point p^c_{ik} such that the following holds:

$$\text{choice}_{ik}(p) = \begin{cases} \text{risky} & \text{if } p > p^c_{ik} \\ \text{either safe or risky} & \text{if } p = p^c_{ik} \\ \text{safe} & \text{if } p < p^c_{ik} \end{cases} \tag{3}$$

We call a participant i in group k *consistent* iff

$$\max\{p|\text{choice}_{ik}(p) = \text{safe}\} < \min\{p|\text{choice}_{ik}(p) = \text{risky})\}. \tag{4}$$

For a consistent participant a p^c_{ik} can be found such that all choices can be rationalised with Equation 3. Indeed, 82% of the participants in this sample are consistent. We call a participant inconsistent if 4 does not hold, i.e. not all their choices can be rationalised with Equation 3. The choices of 18% of the participants are inconsistent.

A certain amount of inconsistent choices is typical for this test. Some researchers react by using an alternative test that forces consistency. Eckel and Grossman (2008) directly ask participants for the switching point. Depending on the research question, this may be satisfactory. But we note that this method does not obtain information about the consistency of a participant's choice. Below we will argue that information about consistency may be useful.

Sometimes inconsistent choices might result from a sub-optimal design, and a re-run of an experiment might be recommended. However, an experimental design that mechanically forbids inconsistencies also prevents the researcher from observing natural and informative heterogeneity among participants. Should one just ignore this data? Or should one, instead, try to make sense of this data? If so, should one make a distinction between participants who gave consistent answers and those who did not? We think one should. If participants have difficulties answering this task in a consistent way, it is essential to learn as much as possible about these difficulties. At the least the fact that a participant had a hard time deciding which lottery to prefer informs us about the reliability of our measure of risk attitude. The inconsistency may even be more than confusion or fatigue; it may tell us something about the confidence a participant has in expressing her attitude towards risk.

Looking at Figure 1 again we see that the better part of these inconsistencies can be found among the more risk loving participants. Systematically dropping these observations might not only introduce a selection bias but might also make the remaining measure of risk attitudes appear unrealistically precise and, thus, lead to a bias due to errors in variables.

[4] We later estimate, for each participant, the precision of the measure. This can be done in a more straightforward way with p^c_{ik} as a measure of risk aversion.

3.2. No correction for errors in variables

In the example study, the aim is to explain punishment behaviour as a function of the attitude towards risk. The latter is described as a switching point p^c_{ik} in the Holt and Laury task. In Section 3.2 we comparatively assess four alternative approaches for dealing with inconsistent choices. All four approaches can be used to estimate Equation 2, but all assume that p^c_{ik} could be measured with infinite precision. As a result none of these four approaches addresses the errors in variables problem. In Section 3.3 we present two more approaches which estimate the decision process determining p^c_{ik} simultaneously with Equation 2. These approaches offer a solution for the errors in variables problem.

Drop inconsistent observations (DROP) This procedure would require to remove from our sample the 18% of the participants which are inconsistent according to 4. In Figure 2 these are the participants which are crossed out by a vertical dashed line. For the remaining 82% of our participants we define the switching point as follows:

$$\hat{p}^{c,D}_{it} = \frac{\max\{p|\text{choice}_{ik}(p) = \text{safe}\} + \min\{p|\text{choice}_{ik}(p) = \text{risky})\}}{2} \tag{5}$$

Figure 2 suggests that inconsistent behaviour could be more likely with risk seeking participants. The DROP procedure might, hence, selectively remove risk seeking participants from the sample. It also does not tell us anything about the precision of p^c_{ik}, i.e. it does not help us to address the errors in variables problem.

Counting the number of safe choices (COUNT) Holt and Laury (2002) propose to replace the switching point for inconsistent participants by simply counting the number of safer choices. To ease the comparison with the other measures we use the following linear transformation:

$$\hat{p}^{c,C}_{it} = \frac{1}{20} + \frac{1}{10}\sum_{p}[\text{choice}_{ik}(p) = \text{safe}] \tag{6}$$

Figure 2 shows the resulting estimates of risk preferences as a thick dotted line. This procedure addresses the selection bias but not the errors in variables problem.

A logistic regression to estimate switching points (LOGIS) We could describe the probability that individual i in group k chooses the lottery with the larger spread conditional on the probability p of the good outcome as a logistic function $\mathcal{L}$ of a linear function of p:

$$P(\text{risky}_{ik}|p) = \mathcal{L}\left(\beta_{0,ik} + \beta_{1,ik}p\right) \text{ where } p \in \{.1, .2, \ldots, 1\} \tag{7}$$

The value of p where the $P(\text{risky}_{ik}|p) = 1/2$, i.e. where individual i in group k chooses the more risky and the safer lottery with equal probabilities, is our estimated switching point $\hat{p}^{c,L}_{ik}$. It is given by

$$\hat{p}^{c,L}_{ik} = -\hat{\beta}_{0,ik}/\hat{\beta}_{1,ik}\,. \tag{8}$$

The dashed line in the bottom part Figure 2 shows for each individual the critical value $\hat{p}^{c,L}_{ik}$ obtained with this method.[5] As Figure 2 demonstrates, the results obtained with LOGIS are similar to COUNT, except for participants 13 and 42.[6] The top part of the same figure shows for each individual the coefficient $\hat{\beta}_{1,ik}$. When this coefficient is large then $P(\text{risky}_{ik}|p)$ is either close to 1 or close to 0 for most values of p. A large coefficient is, hence, a measure of consistency. When we use maximum likelihood to estimate Equation (7) we should expect that for consistent choices $\hat{\beta}_{1,ik} \to \infty$. Since numerical precision is limited we find for consistent choices in our estimation $432 \leqslant |\hat{\beta}_{1,ik}| \leqslant 447$ which is clearly smaller than $+\infty$, but already sufficiently large to make sure that the actual choices are made almost with certainty.[7] Still, we should keep in mind that it is only numerical imprecision which yields finite values where we should see a $+\infty$.

Looking at Figure 2 again we see two (related) problems:

1. For the 18% inconsistent choices we have $\hat{\beta}_{1,ik} \leqslant 13$. These choices are clearly more noisy than the 82% consistent choices with $\hat{\beta}_{1,ik} \geqslant 432$ but it is not obvious how to reflect this difference in precision in our estimate of Equation (2).

2. The estimation of Equation 7 yields for two participants (13 and 42) negative values for $\hat{\beta}_1$ (-6.1 and -445). These participants choose the safer lottery more frequently when the probability of the good outcome is larger. The LOGIS model does not tell us how one should interpret the data for these cases.

We will argue below that these 18% inconsistent participants can serve two purposes. First, although their observations are noisy, dropping them leads to a selection bias. Second, and more importantly, the noise of these observations allows us to address the errors in variables problem. If 18% of our participants clearly violate consistency we should perhaps not expect that the remaining 82% are ultimately precise. The inconsistent 18% will allow us to better assess the precision of the remaining 82% consistent observations.

Estimation results for DROP, COUNT and LOGIS Table 1 shows the estimation results for Equation 2 for different ways to deal with inconsistent observations. We see that, regardless which method we use here, the differences are not very large. We find $\hat{\beta}_{\xi\times p}$ somewhere between -1.17 and -0.876. Whichever approach we use, we have a negative effect. The more

[5] Note that LOGIS (the same way as the Bayesian methods) easily handles "inconsistent" participants. Figure 1 shows that we have 13 such participants in the dataset.

We have no participants who, independent of p, always choose the risky lottery. These participants would correspond to $\hat{p}^{c,L}_{ik} < 0$. We have two participants always choose the safe lottery. They correspond to $\hat{p}^{c,L}_{ik} > 1$.

[6] Since the logistic model is not fully identified it is only a convenient artefact of the numerical implementation to find a unique answer to the question for the optimal switching point. If a participant has chosen the safer lottery for all choices $p \leqslant .6$ and the more risky lottery for all choices $p \geqslant .7$, the logistic model will estimate a switching point just in the middle between .6 and .7 at almost exactly .65.

[7] If a participant is just indifferent at p^c, i.e. $\beta_0 + \beta_1 p^c = 0$, then the next actual choice in the experiment is made for $p = p^c + 1/20$ and $p = p^c - 1/20$. The probability of a safe or risky choice there is, hence, $\mathcal{L}(\beta_{1,ik}/20)$ and $\mathcal{L}(-\beta_{1,ik}/20)$. For $\beta_{1,ik} = 432$ we have $\mathcal{L}(432/20) \approx 1 - 4.16 \times 10^{-10}$, $\mathcal{L}(-432/20) \approx 4.16 \times 10^{-10}$.

	DROP			COUNT			LOGIS		
	β	2.5%	97.5%	β	2.5%	97.5%	β	2.5%	97.5%
0	-1.267	-2.776	0.171	-0.809	-2.025	0.423	-0.732	-1.962	0.496
ξ	1.356	0.663	2.050	1.208	0.699	1.702	1.190	0.673	1.692
p	1.130	-0.944	3.290	0.457	-1.322	2.190	0.339	-1.426	2.085
$\xi \times p$	-1.172	-2.198	-0.145	-0.909	-1.632	-0.166	-0.876	-1.608	-0.126
	σ^2	σ	$1/\sigma^2$	σ^2	σ	$1/\sigma^2$	σ^2	σ	$1/\sigma^2$
ν'_{ik}	0.000	0.000	Inf	0.000	0.000	Inf	0.000	0.000	Inf
ν_k	0.287	0.536	3.483	0.375	0.612	2.666	0.379	0.616	2.638
ϵ_{ikt}	10.381	3.222	0.096	10.296	3.209	0.097	10.301	3.210	0.097

Table 1: ME estimate of Equations 2.

a participant is risk averse (on a scale from 0 for very risk loving to 1 for very risk averse), the less she reacts to the amount of punishment she has received in the previous period.

To properly interpret this finding, note the large coefficient of β_ξ ($1.19 \leqslant \beta_\xi \leqslant 1.36$, depending on the model): Irrespective of the estimation procedure, a perfectly risk loving subject ($p^c = 0$) increases her contributions by more than 1 unit in response to any unit of punishment she has received in the previous period. The more the participant is risk averse, the less intense her reaction. Yet even a perfectly risk averse participant ($p^c = 1$) still exhibits a small increase of contributions in reaction to punishment ($0.184 \leqslant \beta_\xi + \beta_{\xi \times p} \leqslant 0.313$ depending on the model).

A Bayesian approach We do not want to enter a discussion on the comparative merits of the Bayesian versus the frequentist framework (Bayarri and Berger, 2004, or Kass, 2011 may provide a starting point for a discussion). Below we will employ the Bayesian approach as a flexible and straighforward method to obtain an estimate for the two decision processes simultaneously. To facilitate the comparison with the frequentist framework we base our estimations on vague priors. Bayesian estimation has been shown to work well in the context of errors in variables models for a long time and for a wide range of situations.[8] We will also demonstrate below that, as long as the frequentist and the Bayesian approach estimate the same model, the results are (of course) almost indistinguishable.[9] Very similar to Equations 7 and 8 above, we describe the probability to choose the risky lottery if the probability of the good outcome is p as follows:

$$P(\text{risky}_{ik}|p) = \mathcal{L}\left((p - p^c_{ik}) \cdot \sqrt{\tau_{ik}}\right) \text{ with } p \in \{.1, .2, \ldots, 1\} \tag{9}$$

The idea is the same as in Equations 7 and 8. The problem is now described in one equation. We explicitely introduce the parameter τ_{ik} to measures the precision of the choice of parti-

[8] Arminger and Muthén (1998), Dellaportas and Stephens (1995), Florens, Mouchart and Richard (1974) and Polasek and Krause (1993).

[9] To estimate all Bayesian models we use `JAGS` 4.0.0. Estimates are based on four chains with each 1000 samples for adaptation, 4000 samples for burnin, and, for each of the four chains, 10000 samples used for estimating the distribution. To estimate the mixed effects model we use `lme4` 1.1-12. Frequentist confidence intervals are based on a normal bootstrap with 1000 samples.

cipant ik. The amount contributed to the public good is as specified above in Equation 2. We assume the following (vague) priors:[10]

For the coefficients from Equation (2):

$$\beta_l \sim \mathcal{N}(0, 100) \text{ with } l \in \{0, \xi, p, \xi \times p\} \tag{10}$$

For the switching point from the risk task, Equation (9):

$$p^c_{ik} \sim \mathcal{B}(\alpha_c, \beta_c) \text{ with } \alpha_c \sim \Gamma(2, 1/2); \beta_c \sim \Gamma(2, 1/2) \tag{11}$$

For the precision of the switching point:

$$\tau_{ik} \sim \Gamma(m^2/d^2, m/d^2); \text{ with } m \sim \Gamma(1, 1); \ d \sim \Gamma(10, 0.1) \tag{12}$$

The group specific random effect in Equation (2):

$$\nu_k \sim \mathcal{N}(0, 1/\sqrt{\tau_\nu}); \text{ with } \tau_\nu \sim \Gamma(m^2_\nu/d^2_\nu, m_\nu/d^2_\nu); m_\nu \sim \Gamma(1, 1); \ d_\nu \sim \Gamma(1, 1) \tag{13}$$

The individual specific random effect in Equation (2):

$$\nu'_{ik} \sim \mathcal{N}(0, 1/\sqrt{\tau_{\nu'}}); \text{ with } \tau_{\nu'} \sim \Gamma(m^2_{\nu'}/d^2_{\nu'}, m_{\nu'}/d^2_{\nu'}); \\ m_{\nu'} \sim \Gamma(1, 1); \ d_{\nu'} \sim \Gamma(1, 1) \tag{14}$$

The residual in Equation (2):

$$\epsilon_{ikt} \sim \mathcal{N}(0, 1/\sqrt{\tau_\epsilon}); \text{ with } \tau_\epsilon \sim \Gamma(m^2_\epsilon/d^2_\epsilon, m_\epsilon/d^2_\epsilon); m_\epsilon \sim \Gamma(1, 1); \ d_\epsilon \sim \Gamma(1, 1) \tag{15}$$

Replicating LOGIS (B-LOGIS): Before we come to the results of the simultaneous estimation, let us replicate the result of the mixed effect estimation of Equation (2) with p^c_{ij} based on the LOGIS model within a Bayesian framework. As in the LOGIS case, we first estimate p^c_{ik} for each participant and then, as a separate problem, estimate Equation (2), but now using Bayesian inference, and the priors given by (10), (13), (14), (15). This procedure, which we call B-LOGIS, can not take into account errors in variables. Estimation results for the case where inconsistent observations are dropped are shown in Table 2. Here the value for $\beta_{\xi \times p}$ is -0.89, i.e. similar to the corresponding estimate of the mixed effects model based on the LOGIS estimate of p^c_{ik} ($\beta_{\xi \times p} = -0.876$).

3.3. Correcting for errors in variables

B-JOINT: Compared with LOGIS, the Bayesian framework allows us a simultaneous estimation of both decision processes. We can obtain an estimate of precision of p^c_{ik} that lends itself to meaningful interpretation. Furthermore, the simultaneous estimation of Equations (2) and (9) automatically weighs the individual estimate of risk attitude by its precision. As a result, the estimation takes into account the errors in variables problem. As above we rely on a "standard" mixed effects model here, with random effects only on the intercept. Priors are as given by Equations (10)-(15).

Estimation results are shown in Table 3. The left part shows results for the entire data set with 72 observations, the right part shows results only for the 59 consistent observations.

[10] We use $\mathcal{N}(\mu, \sigma)$ for the normal distribution, $\Gamma(\alpha, \beta)$ for the Gamma distribution and $\mathcal{B}(\alpha, \beta)$ for the Beta distribution. The second argument of $\mathcal{N}(\mu, \sigma)$ is the standard deviation. $\tau = 1/\sigma^2$ is the precision. The first argument of $\Gamma(\alpha, \beta)$ is shape α, the second is rate β.

B-LOGIS

	Mean	2.5%	97.5%
0	-0.7565	-2.0902	0.5848
ξ	1.2121	0.6880	1.7219
p	0.3559	-1.5321	2.2766
$\xi \times$ p	-0.8901	-1.6295	-0.1324
τ_ν	7.9661	2.1946	29.9933
τ'_ν	2.2340	0.6166	5.2709
τ_ϵ	0.0974	0.0875	0.1077

Table 2: Estimating Equations 2 and (9) separately in the Bayesian Framework.
No correction is made for errors in variables. Results are, as they should be, quite similar to the LOGIS or the COUNT model. We use 4 chains with each 1000 samples and a thinning interval of 100. We obtain an effective sample size of 3943 and a potential scale reduction factor of 1.0005 for $\xi \times$ p (Gelman and Rubin, 1992).

B-JOINT

	Mean	2.5%	97.5%
0	-1.7651	-2.9991	-0.5696
ξ	3.2480	2.2234	4.3251
p	1.7451	0.1925	3.3990
$\xi \times$ p	-3.6398	-5.2124	-2.2025
τ_ν	6.8940	2.1041	23.3325
τ'_ν	2.3300	0.6502	5.8009
τ_ϵ	0.1092	0.0974	0.1218

B-JOINT-CONSIST

	Mean	2.5%	97.5%
0	-2.4791	-4.0332	-1.0334
ξ	4.2483	3.0818	5.6085
p	2.7848	0.7883	4.9295
$\xi \times$ p	-5.1555	-7.1253	-3.4713
τ_ν	5.4334	1.6673	17.0042
τ'_ν	2.5509	0.6659	6.5969
τ_ϵ	0.1136	0.1000	0.1282

Table 3: Simultaneous estimation of Equations 2 and (9) in the Bayesian Framework.
The simultaneous estimation corrects for errors in variables. The B-JOINT model uses all data (left table). We sample from 4 chains with each 1000 samples and a thinning interval of 100. We obtain an effective sample size of 4334 and a potential scale reduction factor of 0.9998 for $\xi \times$ p. B-JOINT-CONSIST uses only consistent participants (right table). We sample from 4 chains with each 1000 samples and a thinning interval of 100. We obtain an effective sample size of 4105 and a potential scale reduction factor of 1.0010 for $\xi \times$ p.

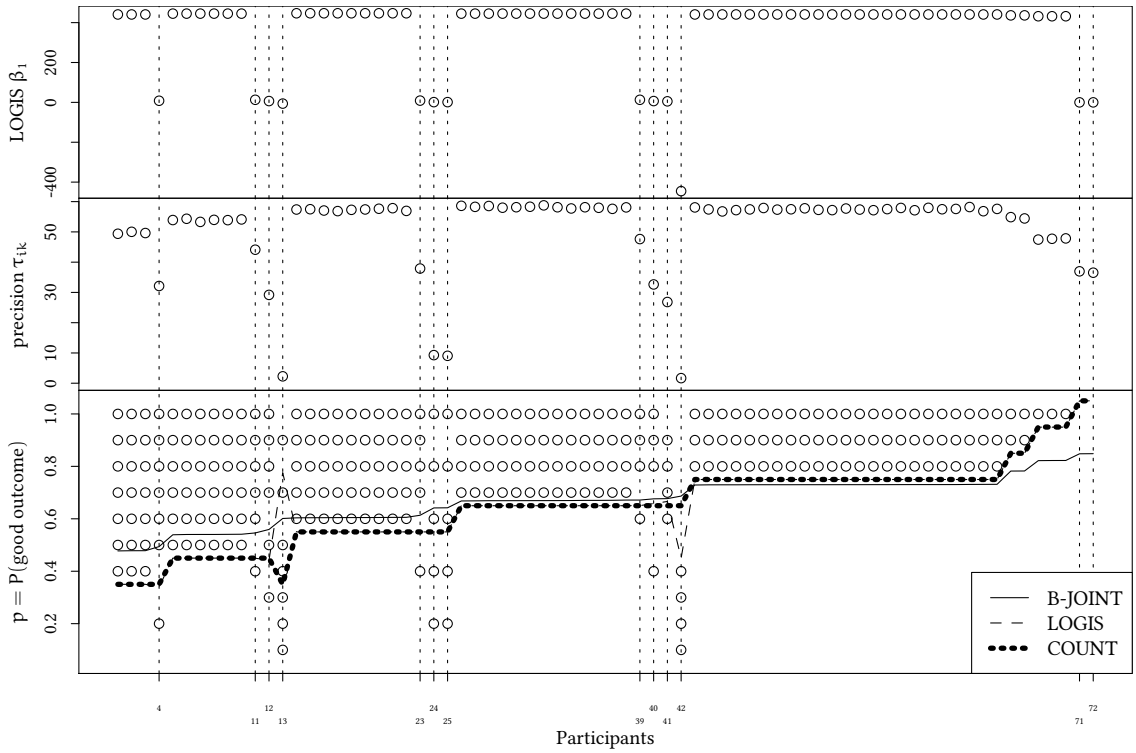

Figure 2: Choices, switching points p^c_{ik} and precision of choice τ_{ik}

The bottom panel shows for each participant the actual choices: ○ if the participant chose the more risky lottery. Participants are ordered by their median switching points p^c_{ik} as estimated from the B-JOINT model. The solid line denotes the median estimated switching points p^c_{ik} from B-JOINT. The dashed line shows the estimated switching points from LOGIS. Vertical reference lines denote participants with inconsistent choices, i.e. with more than one switching point. The panel in the middle shows the estimated values of the participant's precision, τ, from B-JOINT. The top panel shows the estimated value of β_1 from LOGIS.

Figure 6 in Appendix A shows posterior distributions for α_τ, β_τ and p^c_{ik}. These are, however, only intermediate results which we skip here. Figure 2 shows the predicted switching points p^c_{ik} as a solid line. The B-JOINT estimate for p^c_{ik} follows the estimates based on COUNT or LOGIS, in particular for the central values of p^c. For participants where LOGIS and COUNT estimate more extreme values of p^c, B-JOINT takes a more conservative approach. E.g. the extreme risk aversion of the rightmost participants in Figure 2 is not really in line with the distribution of the remaining values of p^c_{ik}. B-JOINT estimates, hence, a smaller precision τ_{ik}, and, accordingly, adjusts the value of p^c_{ik} more towards the centre of the distribution.

For individuals 13 and 42 (those, who choose the safer lottery more frequently when the probability of the good outcome was larger) LOGIS estimates with Equation (7) a negative slope β_1 and, hence, a meaningless switching point. For these two individuals the Bayesian model estimates a precision τ_{ik} very close to zero.

The top panel in Figure 2 shows the value of β_1 from Equation (7). The panel in the middle shows the estimated precision τ_{ik} from Equation (9). Comparing both panels, one sees that the B-JOINT estimates are more differentiated. The LOGIS estimates for β_1 are either close to

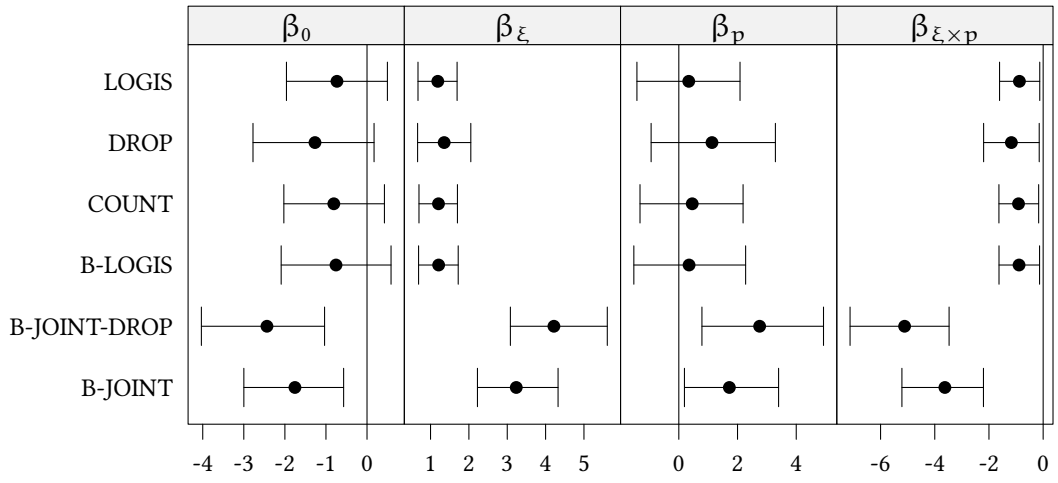

Figure 3: HPD and confidence invervals for Equation 2.
The figure shows 95% confidence intervals for the mixed effects model based on LOGIS, DROP and COUNT estimates for p^c_{ik}. The figure also shows 95% HPD intervals for three specifications of the Bayesian model: B-LOGIS, which is a replication of the B-LOGIS model in the Bayesian framework based on consistent choices only, B-JOINT-DROP, the joint model based on only consistent choices, and B-JOINT, the joint model for all choices.

positive or negative infinity, or close to zero. By contrast the B-JOINT estimates for precision τ_{ik} show a more detailed picture of deviation from utility maximising behaviour. For the consistent choices the estimated parameter for τ is rather large with a median value of 57.4. For the inconsistent choices τ covers a range from 1.68 to 47.6.

3.4. Selection bias versus errors in variables

While the results of B-JOINT are based on the entire dataset, including the inconsistent decision makers, we also estimate B-JOINT-CONSIST, based on the same model but using only data from the consistent decision makers. The comparison of the two models, B-JOINT and B-JOINT-CONSIST, allows us to decide whether our results are mainly driven by the correction for errors in variables or by avoiding selection bias. Both models take into account errors in variables. Both models come to substantial effect sizes for $\xi \times p$: -5.16 for B-JOINT-CONSIST, and -3.64 for B-JOINT. Not controlling for errors in variables in the DROP, COUNT, or LOGIT models yields effect sizes between -1.17 and -0.876. In other words: Correcting for errors in variables (and thereby weighting the individual measure of risk attitude with its precision) changes the effect size by 210%. Once errors in variables are taken into account, including inconsistent observations affects the effect size by only 30%.

Figure 3 compares the estimation results graphically. The figure shows confidence intervals for the uncorrected models and HPD intervals for the B-JOINT models. In particular when it comes to $\beta_{\xi\times p}$ we observe a big difference between the B-JOINT model and the

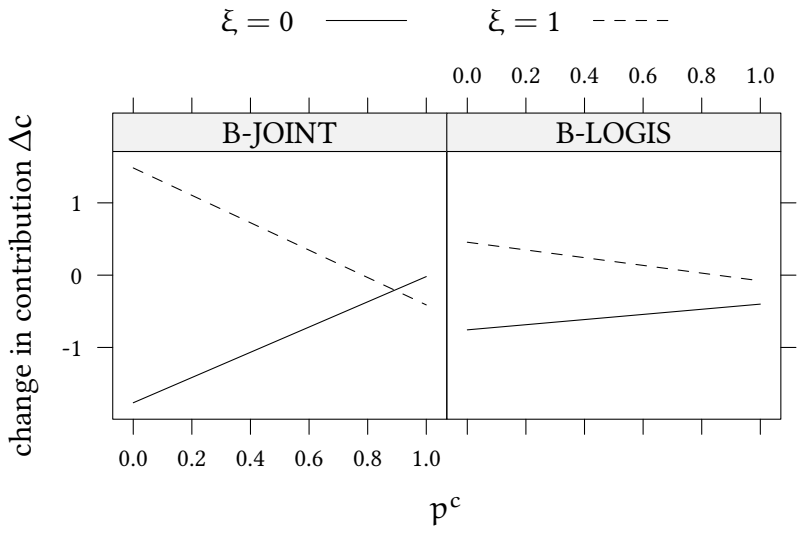

Figure 4: Predicted Δc depending on the estimation method.

uncorrected models.

Figure 4 illustrates the bias when not correcting for errors in the measurement of risk. The B-LOGIS model suggests that the effect of punishment is small (the intercepts of the two lines are much closer to zero), and that the effect of risk aversion on the sensitivity to punishment is less pronounced (the slopes of both lines are much flatter). When the bias is removed, one sees that the effect of punishment is large (intercepts are far away from each other) and more sensitive to risk aversion (both lines have a much steeper slope).

4. Simulation

Should one correct for errors in variables? The above result seems to suggest that such a correction is desirable, but how general is this finding? Here we simulate 100 times a sample that is similar to the one we studied above. Each sample has a size of 100 participants which come in 25 groups.

Behaviour in the risk task and in the public good game follows Equations (2) and (9). The parameters of the regression are random and in the same order of magnitude as in our experiment: $\beta_l \sim \mathcal{N}(0, 2)$ for $l \in \{0, \xi, p, \xi \times p\}$. The random effects have a similar variance: $\nu_k \sim \mathcal{N}(0, \sqrt{1/5})$, $\nu'_{ik} \sim \mathcal{N}(0, \sqrt{2/7})$, $\epsilon_{ikt} \sim \mathcal{N}(0, \sqrt{10})$. The risk aversion also follows a distribution similar to the one in our experiment: $p^c_{ik} \sim \mathcal{B}(6.98, 3.63)$, $\tau_{ik} \sim \Gamma(0.847, 0.2)$.

For each of the 100 simulations we obtain an estimate for the coefficients of Equation (2). Here we are specifically interested in $\beta_{\xi \times p}$. Figure 5 shows for both methods COUNT and B-JOINT quartiles of the difference between the estimates and the true values, $\hat{\beta}_{\xi \times p} - \beta_{\xi \times p}$. We see that B-JOINT performs fairly well. The difference $\hat{\beta}_{\xi \times p} - \beta_{\xi \times p}$ is close to zero. The estimates of COUNT are clearly biased. They are too large in the negative and too small in the positive domain. This bias is what we should expect if errors in variables are neglected.

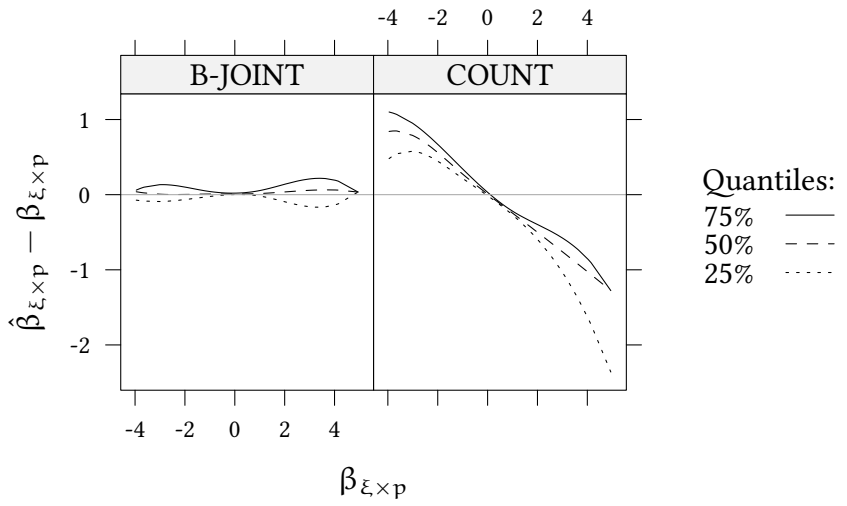

Figure 5: Simulation results
The figure shows 25%, 50%, and 75% quantiles of a B-spline (df=5) for different values of $\beta_{\xi\times p}$ and for the different models.

5. Conclusion

We have the impression that a lot can and should be learned from behaviour which looks inconsistent, i.e. which does not fit the model the experimenter has in mind. We have tried to make use of seemingly inconsistent data in two ways: directly, by not dropping these observations, thereby avoiding selection bias, and indirectly by taking more seriously the lack of precision of all, thereby addressing the errors in variables problem.

We have seen that, even in a situation where errors in variables look small, the difference between a model that neglects the error in variables and one that takes this error into account can be substantial. The aim of this paper is to convince the experimental community that it makes sense to take errors in variables seriously, and that these errors can be handled in a meaningful, and in a feasible way.

But the reanalysis of the example data set also yields a message that is relevant for criminal policy: the experience of having been punished has the most profound effect on individuals who are risk seeking. Regardless whether we neglect or take into account errors in variables we always find strong evidence against hypothesis 1. The size of the effect depends, however, on whether errors are taken into account. For criminal policy, this is welcome news. It has been claimed theoretically that criminals must in equilibrium be risk-seeking (Becker, 1968). Empirical evidence is only correlational, but supports the point (Cochran, Wood and Arneklev, 1994; De Li, 2004; LaGrange and Silverman, 1999). Hence those individuals whose behavior society is most interested to change by the experience of punishment are actually most sensitive to this experience.

We have also seen that observations which are not, or not perfectly consistent with theories of rational decision making, should not be cast away. Otherwise one risks to estimate

effects that suffer from selection bias and one foregoes the opportunity to address the errors in variables problem. The fact that the Holt and Laury task asks each participant to take multiple risky choices is not a nuisance. It enables the researcher to assess the precision of his or her instrument.

References

Adcock, R. J. (1877). "Note on the Method of Least Squares". In: *The Analyst* 4.6, pp. 183–184.

Arminger, Gerhard and Bengt O. Muthén (1998). "A Bayesian Approach to Nonlinear Latent Variable Models Using the Gibbs Sampler and the Metropolis-Hastings Algorithm". In: *Psychometrica* 63.3, pp. 271–300.

Bayarri, M. J. and J. O. Berger (2004). "The interplay of Bayesian and frequentist analysis". In: *Statistical Science* 19.1, pp. 58–80. DOI: 10.1214/088342304000000116.

Becker, Gary (1968). "Crime and Punishment: An Economic Approach". In: *Journal of Political Economy* 76.

Charness, Gary (2000). "Self-Serving Cheap Talk: A Test Of Aumann's Conjecture". In: *Games and Economic Behavior* 33.2, pp. 177–194.

Chaudhuri, Ananish (2011). "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature". In: *Experimental Economics* 14.1, pp. 47–83.

Cochran, John K., Peter B. Wood and Bruce J Arneklev (1994). "Is the religiosity-delinquency relationship spurious? A test of arousal and social control theories". In: *Journal of Research in Crime and Delinquency* 31.1, pp. 92–123.

De Li, Spencer (2004). "The impacts of self-control and social bonds on juvenile delinquency in a national sample of midadolescents". In: *Deviant Behavior* 25.4, pp. 351–373. DOI: 10.1080/01639620490441236.

Dellaportas, Petros and David A. Stephens (1995). "Bayesian Analysis of Errors-in-Variables Regression Models". In: *Biometrics* 51.3, pp. 1085–1095.

Eckel, Catherine C. and Philip J. Grossman (2008). "Forecasting risk attitudes: An experimental study using actual and forecast gamble choices". In: *Journal of Economic Behavior & Organization* 68.1, pp. 1 –17. DOI: http://dx.doi.org/10.1016/j.jebo.2008.04.006.

Engel, Christoph (2014). "Social preferences can make imperfect sanctions work: Evidence from a public good experiment". In: *Journal of Economic Behavior & Organization* 108.C, pp. 343–353.

Fehr, Ernst and Simon Gächter (2000). "Cooperation and Punishment in Public Goods Experiments". In: *American Economic Review, American Economic Association* 90.4, pp. 980–994.

Fischbacher, Urs (2007). "z-Tree: Zurich Toolbox for Ready-made Economic Experiments". In: *Experimental Economics* 10.2, pp. 171–178.

Florens, J. P., M. Mouchart and J. F. Richard (1974). "Bayesian inference in Error-in-Variables Models". In: *Journal of Multivariate Analysis* 4, pp. 419–452.

Gelman, Andrew and Donald B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences". In: *Statistical Science* 7.4, pp. 457–472. DOI: 10.1214/ss/1177011136.

Gillard, Jonathan (2010). "An overview of linear structural models in errors in variables regression". In: *REVSTAT–Statistical Journal* 8.1, pp. 57–80.

Greiner, Ben (2004). "An Online Recruitment System for Economic Experiments". In: *Forschung und wissenschaftliches Rechnen*. Ed. by Kurt Kremer and Volker Macho. Vol. 63. GWDG Bericht. Ges. für Wiss. Datenverarbeitung. Göttingen, pp. 79–93.

Holt, Charles A. and Susan K. Laury (2002). "Risk Aversion and Incentive Effects". In: *The American Economic Review* 92.5, pp. 1644–1655.

Kass, Robert E (2011). "Statistical inference: The big picture". In: *Statistical science: a review journal of the Institute of Mathematical Statistics* 26.1, p. 1.

Kimball, Miles S, Claudia R Sahm and Matthew D Shapiro (2008). "Imputing Risk Tolerance From Survey Responses". In: *Journal of the American Statistical Association* 103.483, pp. 1028–1038. DOI: 10.1198/016214508000000139.

LaGrange, Teresa C. and Robert A. Silverman (1999). "Low Self-Control and Opportunity: Testing the General Theory of Crime as an Explanation for Gender Differences in Delinquency". In: *Criminology* 37.1, pp. 41–72. DOI: 10.1111/j.1745-9125.1999.tb00479.x.

Ledyard, John O. (1995). "Public Goods: A Survey of Experimental Research". In: *The Handbook of Experimental Economics*. Ed. by John H. Kagel and Alvin E. Roth. Princeton NJ: Princeton University Press, pp. 111–194.

Lindley, Dennis Victor and G. M. El-Sayyad (1968). "The Bayesian Estimation of a Linear Functional Relationships". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 30.1, pp. 190–202.

Montero, Maria, Martin Sefton and Ping Zhang (2008). "Enlargement and the balance of power: an experimental study". In: *Social Choice and Welfare* 30.1, pp. 69–87.

Neyman, Jerzy and Elizabeth L. Scott (1948). "Consistent Estimates Based on Partially Consistent Observations". In: *Econometrica* 16.1, pp. 1–32.

Polasek, Wolfgang and Andreas Krause (1993). "Bayesian regression model with simple errors in variables structure". In: *The Statistician* 42, pp. 571–580.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Ross, L., R.E. Nisbett and M. Gladwell (2011). *The Person and the Situation: Perspectives of Social Psychology*. Pinter & Martin Limited.

Solari, Mary E. (1969). "The "Maximum Likelihood Solution" of the Problem of Estimating a Linear Functional Relationship". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 31.2, pp. 372–375.

Zelmer, Jennifer (2003). "Linear Public Goods Experiments: A Meta-Analysis". In: *Experimental Economics* 6.3, pp. 299–310.

A. Posteriors for α_τ, β_τ and p^c_{ik}

Figure 6 shows the prior and posterior distribution for p^c and for the parameters $\alpha_\tau = m^2/d^2$, $\beta_\tau = m/d^2$ which determine the distribution of τ_{ik}. According to Equation (11) p^c_{ik} follows a Beta distribution. The vague prior assumes that the parameters α_c and β_c for this distribution are from a Gamma distribution (so that a priori p^c_{ik} follows an almost uniform

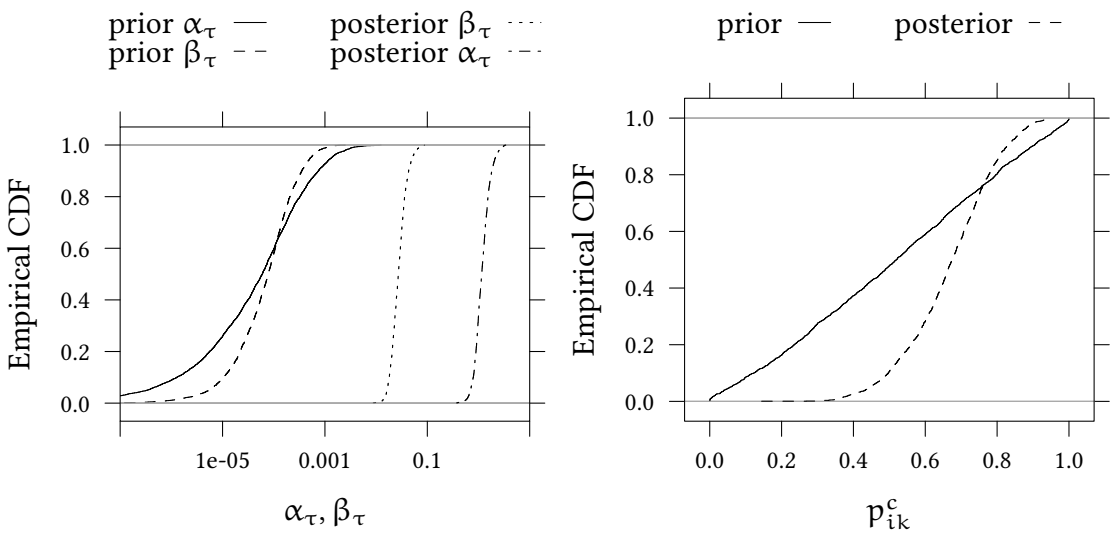

Figure 6: Posteriors for α_τ, β_τ and p^c_{ik}

distribution). The median of the posterior parameters are $\alpha = 7.89$ and $\beta = 3.95$, i.e., as we also see in the Figure, participants do avoid the extreme values of p^c and, not surprisingly, are more risk averse than risk loving.

According to Equation (12) we assume that the precision τ_{ik} is drawn from a Gamma distribution. The parameters of this distribution are endogeneous. The median of the posterior shape parameter is $\alpha = 1.16$ and the median of the posterior rate parameter is $\beta = 0.027$. Figure 7 shows the posterior distribution of τ_{ik} as well as the median values of τ_{ik} for the individual participants. Conceptually, this is not entirely trivial. Often we assume that "consistent" choices are infinitely precise, i.e. $\tau = \infty$. However, if some choices, here 18% of all participants, are inconsistent, i.e. contain a substantial lack of precision ($1.68 \leqslant \tau \leqslant 47.6$), it would be foolish to assume that the remaining 82% choices are infinitely precise.

How can we assess the precision of choices? In Figure 7 we see how the estimator uses the 18% inconsistent observations as a handle to estimate the left part of the distribution of τ. On the right side of the distribution the value of 57.4 for the median consistent decision maker results from the discrete steps in the Holt and Laury (2002) task which implies a finite precision for the consistent choices.

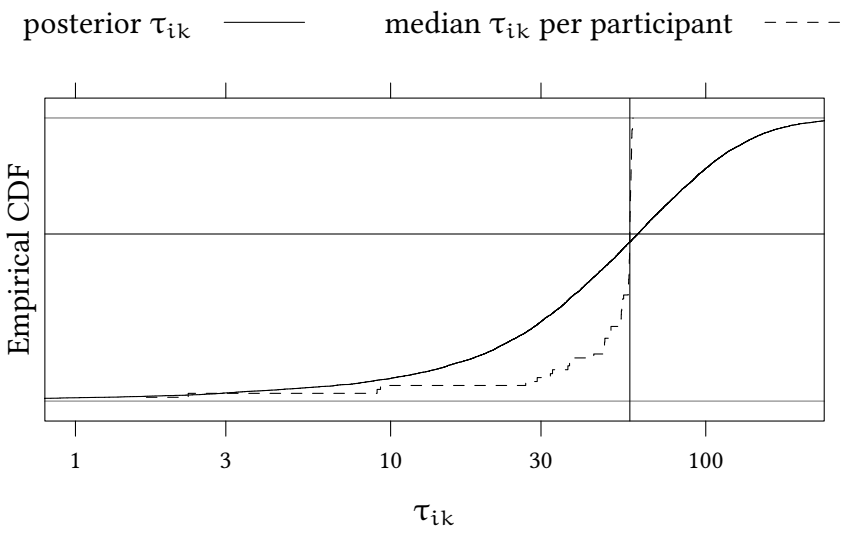

Figure 7: Precision of choices for τ_{ik}

The solid line show the posterior distribution of τ_{ik} as in Equation 12. The dotted line shows the distribution of the median of τ_{ik} taken for each participant.