

# How to Deal with Inconsistent Choices on Multiple Price Lists\*

Christoph Engel<sup>†</sup>      Oliver Kirchkamp<sup>‡</sup>

21st October 2019

Economic experiments are often based on the claim that some heterogeneous behavioral trait affects response to treatment. This trait is measured in another part of the experiment, using a multiple price list. Frequently, choices on such a list are not perfectly consistent.

In this paper we argue that this inconsistency is a resource. It informs the researcher about the precision with which the trait is measured. Precision can be estimated with a measurement error model.

With data from an experiment that studies the relation between risk aversion and punishment we illustrate the approach, and show that it matters.

Keywords: Measurement error, risk, lab experiment, public good, Bayesian inference.  
JEL: C81, C9, D8

## 1. Introduction

When we run laboratory experiments and when we desire to test or find explanations for treatment effects, we sometimes combine two parts of an experiment. In one part of the experiment we measure individual traits. These measurements are used to explain behaviour in another part of the experiment. Clearly, one can not assume that these measurements are free of any errors. Traits are unlikely to be stable across situations (Ross, Nisbett and Gladwell, 2011). They will often only be imperfectly observed in post-experimental tests.

For the econometrician the problem of an explanatory variable that is measured only with some error is well known as one of “errors in variables”. If we want to estimate  $Y = \beta_0 + \beta_1 \xi +$

---

\*Helpful comments by Ioanna Grypari, André Schmelzer and by two anonymous reviewers on an earlier version are gratefully acknowledged. Data and methods used in the paper can be found at <http://www.kirchkamp.de/research/errorsInVar.html>

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

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

$u$ , but we can observe  $\xi$  only with an error, e.g. we observe  $X \sim \mathcal{N}(\xi, \sigma_X)$ , then estimating  $Y = \beta_0 + \beta_1 X + u$  with standard OLS provides a biased and inconsistent estimator for  $\beta_1$ . Already Adcock (1877) mentions the problem of measurement errors. Since then many authors have contributed to the discussion (see Gillard, 2010, for an overview).

In the above problem our estimate for  $\beta_1$  is biased towards zero. The measurement error makes it harder to find an effect if an effect exists. If, nevertheless, we find something, then the true effect must be even stronger. In this context measurement errors might be considered harmless.

The situation changes when further variables come into play. Let us assume we have two explanatory variables, i.e. we estimate  $Y = \beta_0 + \beta_1 \xi_1 + \beta_2 \xi_2 + u$ . Let us further assume that  $\xi_1$  and  $\xi_2$  are positively correlated and  $\xi_1$  can only be observed with an error, i.e. we only observe  $X_1 \sim \mathcal{N}(\xi_1, \sigma_{X_1})$ . Then estimating  $Y = \beta_0 + \beta_1 X_1 + \beta_2 \xi_2 + u$  with standard OLS underestimates  $\beta_1$  and overestimates  $\beta_2$ . Neglecting the error in measuring  $\xi_1$  might wrongly lead us to believe that  $\xi_2$  would matter, even when in reality  $\xi_2$  is irrelevant.

Hey and Orme (1994) point out the importance of error in understanding the relation between preferences and economic decisions. The relevance of measurement errors for risk preferences has been highlighted in a number of studies (Harrison et al. 2005; Loomes 2005; Sahn 2012, to name only a few). Kimball, Sahn and Shapiro (2008) use an error in variables model to interpret survey data on risk tolerance. Beauchamp, Cesarini and Johannesson (2017) employ a latent variable model to assess the predictive power of risk-attitudes. Gilen, Snowberg and Yariv (2018) replicate three influential experimental studies (Niederle and Vesterlund 2007, Friedman et al. 2014, Halevy 2007), pointing out that in all these studies accounting for measurement error changes the interpretation of the results.

Still, the number of economic experiments which take into account measurement errors is small. One reason might be that to correct for measurement errors one has to assess the magnitude of the error. This assessment requires multiple measurements. If the replication movement gains momentum, in the future more such datasets may become available, and can be used for estimating measurement error. Yet, multiple measurements are more expensive and more complicated than single measurements.

The experimental economics community has thus far paid little attention to a much less costly approach. Behavioral traits can often be measured with multiple price lists, like risk aversion in the standard test proposed by Holt and Laury (2002). Such a task by construction requires multiple measurements.

We argue that experimentalists should not force consistency, but should exploit the inconsistency in reactions to estimate the precision of the measurement. We readily admit that this procedure is not perfect. Arguably, consistency over time is even lower than consistency in multiple choices presented on the same computer screen. Yet such data enables experimenters at least to estimate the lower bound of the measurement error problem for the behavioral trait in question.

To demonstrate our approach we consider the following example: In one part of an experiment we measure participants' attitudes towards risk with the help of a Holt and Laury (2002) task. We then use the risk measure as an explanatory variable in a second task, here behaviour in a public good game with punishment. We show how a joint estimation of both

decision processes has two advantages: First, one uses the data from all participants, thus avoiding a selection bias.<sup>1</sup> Second, we can estimate, separately for each participant, the precision of the measure for her risk attitude. This allows us to take into account the measurement error. We can show that results change substantially if one respects the measurement error in the Holt and Laury task.

The remainder of the paper is organized as follows: Section 2 introduces the design of the example 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 the measurement error in a more general context. Section 5 concludes.

## 2. Design of the Example Experiment

In our example we revisit the data from a four-person repeated public good game with punishment (Engel, 2014). 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€ (19.82\$ on the days of the experiment), 14.80€ for players in the public good game, and 16.38€ for players with the power to punish (authorities). The experiment had 3 sessions of 30 participants (6 groups of 4 active participants; 6 passive authorities).

The aim of the experiment was to study the relation between attitudes to risk and the reaction to punishment in a public good game. Fehr and Gächter (2000) show that if participants in a public good game have the possibility to punish each other, contributions in the public good game stabilize at a high level.

In the experiment four (active) participants  $i \in \{1, \dots, 4\}$  in group  $k$  contribute  $c_{ikt}$  in round  $t$  to a linear public good. A fifth participant  $\alpha$  (an authority) has the power to impose a punishment  $\eta_{ikt}$  on each active participant. Profits  $\pi_{ikt}$  of the active participants and  $\pi_{akt}$  of the authority are given by (1) and (2):

$$\text{Profit of active participant } i: \quad \pi_{ikt} = 20 - c_{ikt} + .4 \sum_i c_{ikt} - 3\eta_{ikt} \quad (1)$$

$$\text{Profit of authority } \alpha: \quad \pi_{akt} = 25 + 20 - \sum_i \eta_{ikt} \quad (2)$$

The game is repeated for 11 periods. Participants are rematched every period to matching groups of size 10.

After the main experiment, Engel (2014) administers (among several other tests) a test for risk aversion, following Holt and Laury (2002). Holt and Laury design a task where participants choose between a (safe) lottery with a small spread,  $p \cdot 2\$ + (1 - p) \cdot 1.6\$$ , and a

---

<sup>1</sup>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.

(risky) lottery with a large spread,  $p \cdot 3.85\$ + (1 - p) \cdot .1\$$ , where the probability of the good outcome is  $p \in \{.1, .2, .3, \dots, 1\}$ .

In this experiment, Engel (2014) is interested in the impact of punishment on future cooperation. Punishment in the most recent period is the most vivid experience and should have the strongest effect. The more they are averse to risk, the stronger this signal should guide choices in the subsequent period: risk averse participants lose more utility when punished again. Engel (2014) investigates 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.*

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

Here, we first explain in Section 3.1 why the standard test for risk aversion by Holt and Laury makes error in the measurement of this behavioral trait visible. In Section 3.2 we define the way in which risk aversion matters for the research question of the example experiment. In Section 3.3 we discuss three different ways of estimating the effect of risk aversion on choices that ignore measurement error. These methods proceed in two steps. In the first step, for each participant they generate a point estimate of their risk aversion. In the second step, they use this estimate as an explanatory variable. In Section 3.4 we contrast these approaches with our suggested approach. Our approach has two advantages: (1) for each participant, we do not only estimate the trait; we also generate an estimate of the precision of this measure. (2) we jointly estimate (a) the individual behavioral trait, (b) the precision of this estimate, (c) the impact of the trait on reaction to treatment. As always, joint estimation is more precise, as individual idiosyncrasies are considered in their relation to each other. This is also why the approach tends to have more statistical power. At the end of Section 3.4 we briefly compare selection versus measurement error.

#### 3.1. Measuring risk aversion

To test Hypothesis 1, we need for each active participant a measure of her risk aversion. Choices in the risk task for the 72 participants holding the active role are shown in Figure 1. Choices where a participant chose the lottery with the larger spread (risky) are denoted with a  $\circ$ , choices where the participant chose the lottery with the smaller spread (safe) are left blank.

We assume that preferences for money follow CRRA, i.e.  $u(z) = z^{1-r}$ . In this case 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  $p_{ik}^c$  to describe preferences of individual  $i$  in group  $k$ .

In Figure 1, we have ordered participants from the most risk loving on the left to the most risk averse on the right. If we could describe decision makers by a single switching point  $p_{ik}^c$

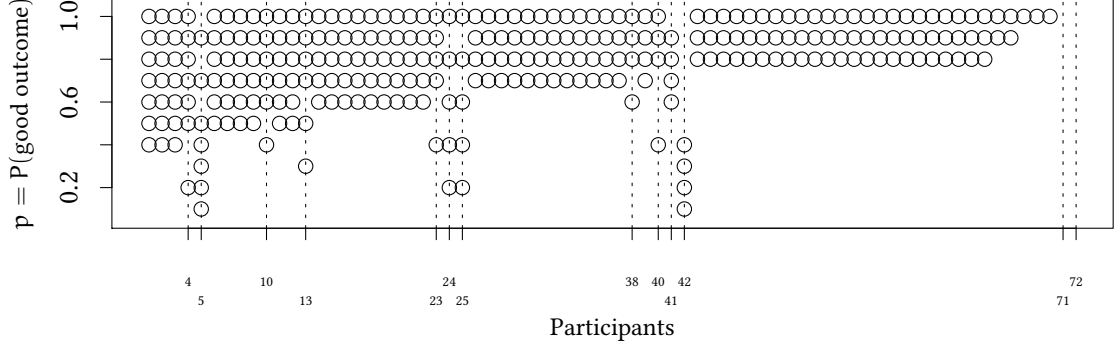


Figure 1: Choices in the risk task

The panel shows choices for each participant:  $\circ$  if the participant choses the lottery with the larger spread (risky) and nothing if the participant chose the smaller spread (safe). Vertical reference lines denote participants with inconsistent choices (see (4)). Participants are ordered by their risk attitudes, with the more risk seeking participants at the left.

then the following holds:

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

We call a participant  $i$  in group  $k$  *consistent* if there is a  $p_{ik}^c$  which rationalises all choices, i.e.

$$\max\{p | \text{choice}_{ik}(p) = \text{safe}\} \leq p_{ik}^c \leq \min\{p | \text{choice}_{ik}(p) = \text{risky}\}. \quad (4)$$

We call a participant *inconsistent* if (4) does not hold, i.e. not all their choices can be rationalised with Equation (3). In Figure 1 vertical reference lines denote participants with inconsistent choices. The choices of 18% of the participants in Engel (2014)'s experiment are inconsistent.

One way to formalise consistent and inconsistent choices in the Holt and Laury task is the logistic model. The probability of a risky choice of individual  $ik$  in lottery  $p$  could be written as follows:

$$P(\text{risky}_{ik}|p) = \mathcal{L}((p - p_{ik}^c) \cdot \sqrt{\tau_{ik}}) \text{ with } p \in \{.1, .2, \dots, 1\} \quad (5)$$

Here  $\mathcal{L}$  is the logistic function,  $p$  describes the probability of the good outcome in the Holt and Laury task,  $p_{ik}^c$  is be the critical value where participant  $ik$  is just indifferent between the two choices, and  $\sigma_{ik}(p_{ik}^c) = 1/\sqrt{\tau_{ik}}$  is the standard deviation of the latent utility of subject  $ik$  in the Holt and Laury task.

While  $\tau_{ik} = \infty$  and, thus,  $\sigma_{ik}(p_{ik}^c) = 0$  would guarantee us consistent choices, we will show below that a value for  $\sigma_{ik}(p_{ik}^c)$  between about 0.16 and 0.25 describes behaviour better. The larger  $\sigma_{ik}(p_{ik}^c)$ , the more frequent are inconsistent choices.

One possible reaction to inconsistent choices is 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. However, by enforcing consistent

choices for  $p_{ik}^c$  we lose information about the decision maker's precision  $\tau_{ik}$  of that choice. Below we will argue that information about this precision may be useful.

Before we do this, let us come to the contributions in the public good game.

### 3.2. Contributions to the public good

To test Hypothesis 1 we have to explain changes in the contribution to the public good  $\Delta c_{ikt}$  as a function of previous punishment  $\eta$  and risk aversion  $p_{ik}^c$ . We eventually want to estimate the following model:

$$\Delta c_{ikt} = \beta_0 + \beta_\eta \eta_{ik,t-1} + \beta_p p_{ik}^c + \beta_{\eta \times p} \eta_{ik,t-1} \cdot p_{ik}^c + \nu_k + \nu'_{ik} + \epsilon_{ikt} \quad (6)$$

$\Delta c_{ikt}$  is the change of contribution to the public good of individual  $i$  from matching group  $k$  at time  $t$ .  $\eta_{ik,t-1}$  is the punishment received by individual  $i$  from group  $k$  at time  $t-1$ , i.e. the punishment received in the previous period.  $p_{ik}^c$  is a measure for risk aversion of individual  $i$  from group  $k$ .  $\nu_k$  is a random effect for group  $k$ .  $\nu'_{ik}$  is a random effect for individual  $i$  from group  $k$ .  $\epsilon_{ikt}$  is the residual. In line with Hypothesis 1 we expect the interaction term  $\beta_{\eta \times p}$  to be positive.

In the example study, the aim is to explain reactions to punishment as a function of the attitude towards risk. The latter is described as a switching point  $p_{ik}^c$  in the Holt and Laury task. In Section 3.3 we briefly discuss four alternative approaches to deal with inconsistent choices. All four approaches can be used to estimate contributions in the public good task (Equation 6), but all assume that switching points  $p_{ik}^c$  could be measured with infinite precision. As a result all four approaches neglect the measurement error. In Section 3.4 we estimate the decision process determining switching points  $p_{ik}^c$  jointly with contributions (Equation 6). This approach offers a solution for the measurement error.

### 3.3. No correction for the measurement error

In this section we present three procedures (DROP, COUNT, and LOGIS) which all take estimated switching points as given and which all disregard the measurement error. In Section 3.4 below we will discuss two procedures (B-JOINT and B-JOINT-DROP) which integrate the measurement error in the estimated switching points.

#### 3.3.1. Drop inconsistent observations (DROP)

This procedure removes 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:<sup>2</sup>

$$\hat{p}_{it}^{c,D} = \frac{\max\{p | \text{choice}_{ik}(p) = \text{safe}\} + \min\{p | \text{choice}_{ik}(p) = \text{risky}\}}{2} \quad (7)$$

<sup>2</sup>Technically, the following expression could also be computed for subjects with multiple switching points.

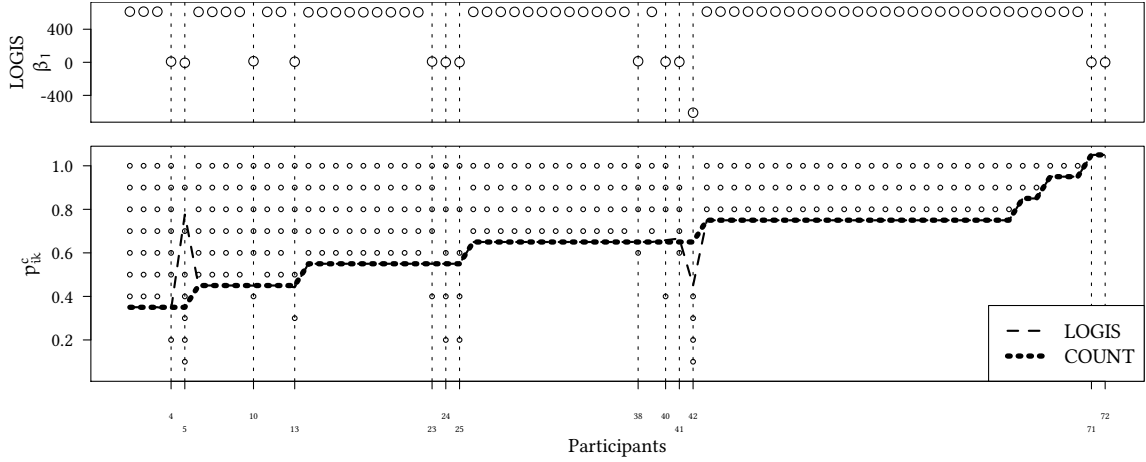


Figure 2: Choices and switching points  $p_{ik}^c$ .

The bottom panel shows for each participant the actual choices:  $\circ$  if the participant chose the more risky lottery. Participants are ordered by their risk attitudes, with the more risk seeking participants at the left. The dashed line shows the estimated switching points  $p_{it}^c$  from LOGIS. Vertical reference lines denote participants with inconsistent choices, i.e. with more than one switching point. The top panel shows the estimated value of  $\beta_1$  from the estimation of the switching point with LOGIS (Equation 9).

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  $p_{ik}^c$ , i.e. it does not help us to address the measurement error.

### 3.3.2. 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:

$$\hat{p}_{it}^{c,C} = \frac{1}{20} + \frac{1}{10} \sum_p [\text{choice}_{ik}(p) = \text{safe}] \quad (8)$$

Figure 2 shows the resulting estimates of risk preferences as a thick dotted line. This procedure avoids inducing the selection bias resulting from dropping inconsistent responders.

### 3.3.3. A logistic regression to estimate switching points (LOGIS)

In Equation (5) we have used the logistic model to describe choices in the risk task. We can rephrase this model as follows:

$$P(\text{risky}_{ik}|p) = \mathcal{L}(\beta_{0,ik} + \beta_{1,ik}p) \text{ where } p \in \{.1, .2, \dots, 1\} \quad (9)$$

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

$\hat{p}_{ik}^{c,L}$ . It is given by

$$\hat{p}_{ik}^{c,L} = -\hat{\beta}_{0,ik}/\hat{\beta}_{1,ik}. \quad (10)$$

The dashed line in the bottom part Figure 2 shows for each individual the critical value  $\hat{p}_{ik}^{c,L}$  obtained with this method.<sup>3</sup> The top panel in Figure 2 shows the value of  $\beta_1$  from the estimation of the switching point with LOGIS (Equation 9). These estimates for  $\beta_1$  are either close to positive or negative infinity, or close to zero, i.e. LOGIS implies either a very clear preference (when  $\beta_1$  is large in absolute terms) or indifference (when  $\beta_1$  is close to zero. As Figure 2 demonstrates, the results obtained with LOGIS are similar to COUNT, except for participants 5 and 42.<sup>4</sup> 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 the switching point (Equation 9) we should expect that for consistent choices  $\hat{\beta}_{1,ik} \rightarrow \infty$ . Since numerical precision is limited we find for consistent choices in our estimation  $605 \leq |\hat{\beta}_{1,ik}| \leq 613$  which is clearly smaller than  $+\infty$ , but already sufficiently large to make sure that the actual choices are made almost with certainty.<sup>5</sup> 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  $\hat{\beta}_{1,ik} \leq 13$ . These choices are clearly more noisy than the 82% consistent choices with  $\hat{\beta}_{1,ik} \geq 605$  but it is not obvious how to exploit this difference in precision in our estimate of contributions in the public good task (Equation 6).
2. The estimation of the switching point with LOGIS (Equation (9)) yields for two participants (5 and 42) negative values for  $\hat{\beta}_1$  ( $-6.1$  and  $-610$ ). 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 to interpret such, seemingly perverse, choices.

We will argue below that these 18% inconsistent participants can serve two purposes. First, although their observations are noisy, dropping them would lead to a selection bias. Second, and more importantly, the noise of these observations allows us to take into account the measurement error. If 18% of our participants clearly violate consistency, we should, perhaps, not expect that we can measure the remaining 82% with infinite precision. The inconsistent 18% will allow us to better assess the precision of the remaining 82% consistent observations.

<sup>3</sup>Figure 1 shows that we have 13 inconsistent participants in the dataset. We have no participants who, independent of  $p$ , always choose the risky lottery. These participants would correspond to  $\hat{p}_{ik}^{c,L} < 0$ . We have two participants who always choose the safe lottery. They correspond to  $\hat{p}_{ik}^{c,L} > 1$ .

<sup>4</sup>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  $p \leq .6$  and the more risky lottery for all choices  $p \geq .7$ , the logistic model will estimate a switching point just in the middle between .6 and .7 at almost exactly .65.

<sup>5</sup>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  $\beta_{1,ik} = 605$  we have  $\mathcal{L}(605/20) \approx 1 - 7.29 \times 10^{-14}$ ,  $\mathcal{L}(-605/20) \approx 7.29 \times 10^{-14}$ .



	DROP			COUNT			LOGIS		
	$\beta$	2.5%	97.5%	$\beta$	2.5%	97.5%	$\beta$	2.5%	97.5%
0	-1.267	-2.776	0.171	-0.809	-2.025	0.423	-0.732	-1.962	0.496
$\eta$	1.356	0.663	2.050	1.208	0.699	1.702	1.189	0.673	1.692
p	1.130	-0.944	3.291	0.457	-1.322	2.190	0.339	-1.427	2.085
$\eta \times p$	-1.172	-2.198	-0.145	-0.909	-1.632	-0.166	-0.876	-1.608	-0.126
	$\sigma^2$	$\sigma$	$1/\sigma^2$	$\sigma^2$	$\sigma$	$1/\sigma^2$	$\sigma^2$	$\sigma$	$1/\sigma^2$
$v'_{ik}$	0.000	0.000	Inf	0.000	0.000	Inf	0.000	0.000	Inf
$v_k$	0.287	0.536	3.483	0.375	0.612	2.666	0.379	0.616	2.638
$\epsilon_{ikt}$	10.381	3.222	0.096	10.296	3.209	0.097	10.301	3.210	0.097

Table 1: MLE Estimates of Contribution Changes Conditioning on Switching Points (Equation 6).

For each of the three procedures (DROP, COUNT, LOGIS) the top part of the tables provide the estimated effect  $\beta$  as well as a 95% confidence interval. The bottom part reports variance ( $\sigma^2$ ), standard deviation ( $\sigma$ ) and precision ( $1/\sigma^2$ ) for the individual specific random effect ( $v'_{ik}$ ), the group specific random effect ( $v_k$ ) and the residual ( $\epsilon_{ikt}$ ).

### 3.3.4. Estimation results for DROP, COUNT and LOGIS

Table 1 shows the estimation results for the impact of risk aversion on the reaction to past punishment. The table compares different ways to deal with inconsistent observations.<sup>6</sup> We see that, regardless which method we use here, the differences are not very large. We find  $\beta_\eta$  between 1.19 and 1.36,  $\beta_p$  is never significant and between 0.339 and 1.13, and  $\hat{\beta}_{\eta \times p}$  somewhere between -1.17 and -0.876.

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 ( $0.184 \leq \beta_\eta + \beta_{\eta \times p} \leq 0.313$  depending on the model).

## 3.4. Accounting for the measurement error – joint estimation of switching points and contributions

The previous three procedures (DROP, COUNT, and LOGIS) all take estimated switching points  $p_{ik}^c$  as given. Their estimate of the impact of  $p_{ik}^c$  on contributions  $\Delta c_{ikt}$  in Equation (6) does not take into account the measurement error of switching points  $p_{ik}^c$  in Equation (5). In this section we suggest that much can be gained if both problems are estimated together. We will use a Bayesian approach. We do not want to enter a discussion on the comparative

<sup>6</sup>For LOGIS we present estimation results which are based on a ML approach. We could also use a Bayesian approach to estimate contributions in the public good task (Equation 6) and we would find very similar results. Only when we estimate switching points (Equation 5) together with contributions (Equation 6) results change (see Section 3.4).

merits of the Bayesian versus the frequentist framework (Bayarri and Berger, 2004, or Kass, 2011 may provide a starting point for a discussion).

In the frequentist world the problem that we outline below could be described as a generalised multilevel structural equation (Rabe-Hesketh, Skrondal and Pickles, 2004). Neyman and Scott (1948) and Solari (1969) have pointed out deficiencies in the maximum likelihood approach to estimate models with measurement errors. Bayesian estimation has been shown to work well in the context of measurement errors for a long time and for a wide range of situations.<sup>7</sup> 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 particularly well to estimate models with measurement errors.

Here we employ the Bayesian approach, in particular since it facilitates a transparent description of the two processes we want to estimate jointly. To ease the comparison with the frequentist framework we base our estimations on vague priors.<sup>8</sup> We will present the joint model in Section 3.4.1.<sup>9</sup> We will then show in Section 3.4.2 that, once the measurement error is taken into account, results change substantially.

### 3.4.1. The joint model

**Likelihoods:** The likelihood of the Holt and Laury task is given by Equation (5). We rewrite Equation 6 for contributions in the public good task to obtain the likelihood as follows:

$$\Delta c_{ikt} \sim \mathcal{N}(\beta_0 + \beta_\eta \eta_{ik,t-1} + \beta_p p_{ik}^c + \beta_{\eta \times p} \eta_{ik,t-1} \cdot p_{ik}^c + \nu_k + \nu'_{ik}, 1/\sqrt{\tau_\epsilon}) \quad (11)$$

**Priors:** Estimating a Bayesian model requires fully specifying parameters and hyperparameters.<sup>11</sup> We deliberately make conservative choices and use the following vague priors:

For the coefficients from Equation (11):

$$\beta_l \sim \mathcal{N}(0, 100) \text{ with } l \in \{0, \eta, p, \eta \times p\} \quad (12)$$

<sup>7</sup>Arminger and Muthén (1998), Dellaportas and Stephens (1995), Florens, Mouchart and Richard (1974) and Polasek and Krause (1993).

<sup>8</sup>For a frequentist analysis the Stata package `gllamm` or the R package `lavaan` might be useful.

<sup>9</sup>We use R version 3.6.0 (2019-04-26) with JAGS 4.3.0 to estimate Bayesian models. Estimates are based on 4 chains with each 1000 samples for adaptation, 4000 samples for burnin, and then, for each of the four chains, 100000 actual samples per chain.<sup>10</sup> To estimate the mixed effects model we use `lme4` 1.1-19. Frequentist confidence intervals are based on normal bootstraps with 1000 samples.

<sup>11</sup>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  $\alpha$ , the second is rate  $\beta$ . There are a number of standard options to model a vague prior for the precision  $\tau$  of random effects and of residuals. Here we assume that  $\tau$  follows a Gamma distribution with mean and dispersion following a Gamma distribution, too. Alternatives would include a Gamma distribution with fixed shape and rate parameters or a truncated t distribution. In our experience the specific choice does not substantially affect the estimation results for coefficients. However, convergence (measured as `psrf`) with hyperparameters that follow a Gamma distribution themselves is much better than fixed hyperparameters. For the prior for the switching point of the risk task  $p_{ik}^c$  we use a Beta distribution, again with parameters  $\alpha$  and  $\beta$  following a Gamma distribution. Figure 6 in Appendix A should illustrate that the prior for  $p_{ik}^c$  is (almost) uniform..

The group specific random effect in Equation (11):

$$v_k \sim \mathcal{N}(0, 1/\sqrt{\tau_v}); \text{ with } \tau_v \sim \Gamma(m_v^2/d_v^2, m_v/d_v^2); m_v \sim \Gamma(1, 1); d_v \sim \Gamma(1, 1) \quad (13)$$

The individual specific random effect in Equation (11):

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

For the switching point from the risk task, Equation (5)

$$p_{ik}^c \sim \mathcal{B}(\alpha_c, \beta_c) \text{ with } \alpha_c \sim \Gamma(2, 1/2); \beta_c \sim \Gamma(2, 1/2) \quad (15)$$

For the precision of the switching point from Equation (5):

$$\tau_{ik} \sim \Gamma(m^2/d^2, m/d^2); \text{ with } m \sim \Gamma(1, 1); d \sim \Gamma(10, 1) \quad (16)$$

The precision of the change in contribution in Equation (11):

$$\tau_\epsilon \sim \Gamma(m_\epsilon^2/d_\epsilon^2, m_\epsilon/d_\epsilon^2); \text{ with } m_\epsilon \sim \Gamma(1, 1); d_\epsilon \sim \Gamma(1, 1) \quad (17)$$

While this notation might look a bit intricate for the novice, priors and likelihoods can in a straightforward way be represented in JAGS or BUGS.<sup>12</sup> In Appendix B we provide the code for a stylised example.

### 3.4.2. B-JOINT

If we assume uninformative priors, we can of course replicate LOGIS with Bayesian methods.<sup>13</sup> The critical advantage of the Bayesian approach lies in the flexibility it provides in jointly estimating risk aversion, the precision of this measurement, and the treatment effects. We jointly estimate switching points (Equation 5) and contributions in the public good task (Equation 11). This approach automatically weighs the individual estimates of the risk attitude by their precision and, thus, takes into account the measurement error.

With this estimation strategy, we focus on what we are really interested in: how heterogeneous risk aversion affects choices in the main experiment. B-JOINT treats risk aversion and the precision of its measurement as a latent variable over which it integrates. The possibility to do so in a transparent and computationally feasible way is the main advantage of the Bayesian approach.<sup>14</sup> In Appendix B we provide the code for a stylised example. Priors are as given by Equations (12)-(17).

**Contributions in the public good task (Equation 11):** Estimation results (for the entire data set with 72 observations) are shown in the left part of Table 2. Figure 4 shows the highest posterior density (HPD) and confidence intervals for our estimates. The Figure illustrates the bias when not correcting for errors in the measurement of risk.

Figure 6 in Appendix A shows prior and posterior distributions for  $\alpha_\tau$ ,  $\beta_\tau$  and  $p_{ik}^c$ . Figure 7 in Appendix A illustrates the posterior precision  $\tau_{ik}$ . The figure confirms again that precision in any multiple price-list is limited. This limited precision can affect estimation results.

<sup>12</sup>See <http://www.kirchkamp.de/research/errorsInVar.html>

<sup>13</sup>Code and estimations are available from the authors upon request.

<sup>14</sup>For more information on the empirical distribution of precision in our example dataset, and further discussion of these estimates, please see Appendix A.

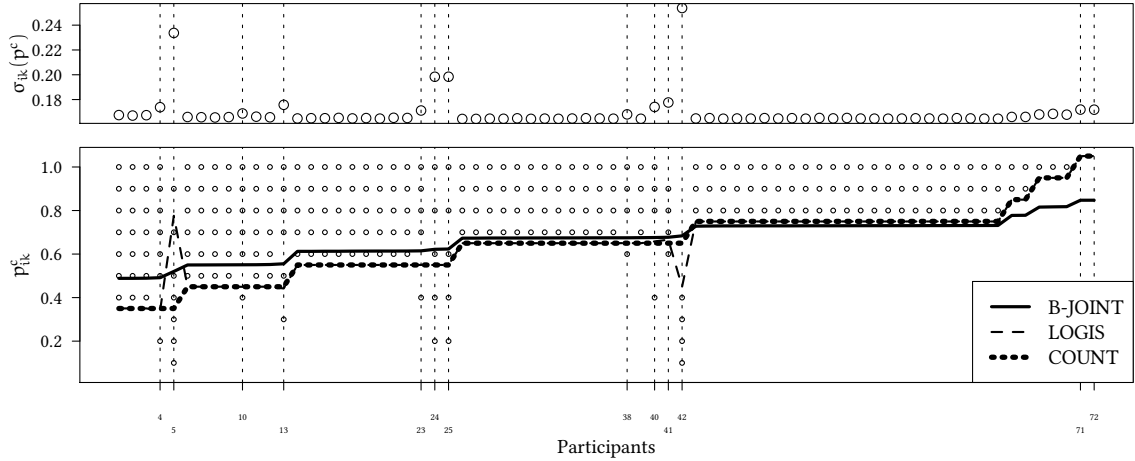


Figure 3: Choices, switching points  $p_{ik}^c$  and their standard deviation.

The bottom panel shows for each participant the actual choices:  $\circ$  if the participant chose the more risky lottery. Participants are ordered by their median switching points  $p_{ik}^c$  as estimated from the B-JOINT model. The solid line denotes the median estimated switching points  $p_{ik}^c$  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 top panel shows the estimated values of the standard deviation  $\sigma_{ik}(p_{ik}^c) = 1/\sqrt{\tau_{ik}}$ , from B-JOINT.

	B-JOINT					B-JOINT-DROP					
	Mean	2.5%	97.5%	SSeff	psrf	Mean	2.5%	97.5%	SSeff	psrf	
$\emptyset$	-1.765	-2.921	-0.513	3916	1.0002	$\emptyset$	-2.479	-4.044	-1.052	3843	1.0007
$\eta$	3.248	2.223	4.325	3891	0.9999	$\eta$	4.248	2.982	5.468	3913	1.0010
$p$	1.745	0.253	3.445	4000	1.0001	$p$	2.785	0.725	4.857	3905	1.0013
$\eta \times p$	-3.640	-5.099	-2.114	4334	0.9998	$\eta \times p$	-5.156	-6.883	-3.256	4105	1.0010
$\tau_v$	6.894	1.486	16.797	4411	1.0022	$\tau_v$	5.433	1.024	12.669	4041	1.0002
$\tau_v'$	2.330	0.456	5.053	4265	1.0004	$\tau_v'$	2.551	0.402	5.439	4257	1.0004
$\tau_\epsilon$	0.109	0.097	0.122	4000	0.9999	$\tau_\epsilon$	0.114	0.100	0.128	4000	1.0004

Table 2: Bayesian Estimates of Contribution Changes (Equation 11) Conditioning on switching points (Equation 5).

Each row in the table reports mean values, 95% credible intervals, and (to assess convergence) effective sample size (SSeff) and potential scale reduction factor (psrf). These values are reported for the coefficients of Equation 11 and the precision ( $\tau$ ) of random effect terms and residuals. The joint estimation corrects for the measurement error. The B-JOINT model uses all data (left table). B-JOINT-DROP uses only consistent participants (right table). We sample from 4 chains with 1000 samples each.

The joint estimate of B-JOINT finds (in absolute terms) a substantially larger effect size for  $\beta_{\eta \times p}$  (-3.64) than the estimates we got from the models where we did not control for the measurement error (between -1.17 and -0.876). In other words: Correcting for the measurement error (and thereby weighting the individual measure of risk attitude with its precision) changes the effect size by 210%.

**Switching points (Equation 5):** Figure 3 shows the predicted switching points  $p_{ik}^c$  as a solid line. The B-JOINT estimate for  $p_{ik}^c$  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 3 is not really in line with the distribution of the remaining values of  $p_{ik}^c$ . B-JOINT estimates, hence, a smaller precision  $\tau_{ik}$ , and, accordingly, adjusts the value of  $p_{ik}^c$  more towards the centre of the distribution.

For individuals 5 and 42 (those, who choose the safer lottery more frequently when the probability of the good outcome was larger) the estimation of the switching point with LOGIS (Equation 9) finds a negative slope  $\beta_1$  and, hence, a switching point which is economically meaningless. Without taking into account the measurement error, this switching point enters the estimation of contribution effects (Equation 6) with the same weight as any other observation. In contrast, the Bayesian model estimates for these two observations a precision  $\tau_{ik}$  very close to zero.

The top panel in Figure 3 shows the estimated standard deviations  $\sigma_{ik}(p_{ik}^c) = 1/\sqrt{\tau_{ik}}$  for switching points of each subject (Equation 5). Comparing with the top panel from Figure 2, we see that the B-JOINT estimates are more differentiated than LOGIT. While the LOGIT implies either a clear preference or clear indifference, the B-JOINT estimates for standard deviation  $\sigma_{ik}$  show a more detailed picture of deviation from utility maximising behaviour. For the consistent choices the estimated parameter for  $\sigma_{ik}$  is small with a median value of 0.165. For the inconsistent choices  $\sigma_{ik}$  covers a range from 0.168 to 0.254.

**Selection bias versus measurement error:** If experimental economists have participants with more than one switching point in their dataset, they often react by dropping these data-points from the analysis altogether. This approach creates an obvious selection problem. The inconsistent participants might exhibit reactions to treatment that systematically differ from the reactions of other participants with at most one switching point. As we have explained in Section 3.3, this potential bias can be addressed by using COUNT or LOGIS. Yet, for estimating the population reaction to treatment, COUNT and LOGIS put the same weight on the risk aversion estimates from all participants. While the approach avoids selection, it potentially overestimates the relevance of reactions from inconsistent participants in the sample for estimating the population effect. By contrast the Bayesian approach holds the balance. It does not ignore choices from inconsistent participants (and thereby avoids selection), but it weights information from the sample by the individual precision of the measurement. This cautious approach is particularly important if the best guess about the degree of risk aversion of inconsistent participants, or of their reaction to risk aversion when exposed to treatment, is at the limit of the support. Taking this information into account without downgrading

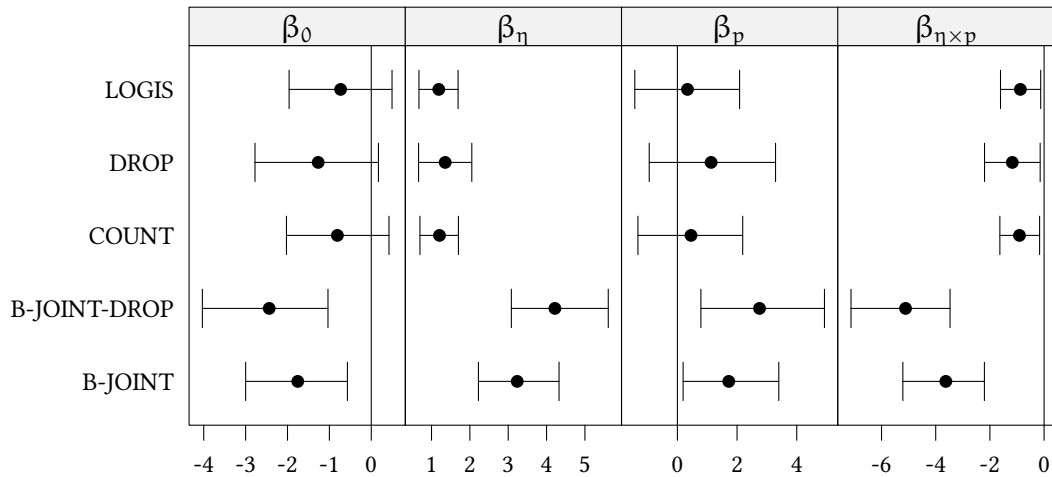


Figure 4: Highest posterior density and confidence intervals for contribution effects (Equation 6).

The figure shows 95% confidence intervals for the mixed effects model based on LOGIS, DROP and COUNT estimates for  $p_{ik}^c$ . The figure also shows 95% HPD intervals for two specifications of the Bayesian model: B-JOINT-DROP, the joint model based on only consistent choices, and B-JOINT, the joint model for all choices.

its importance could introduce bias (from outliers), rather than reducing bias. Actually this concern materializes in our example dataset. As Figure 2 shows, participants who accept high risk gambles are systematically more likely to be inconsistent.

Whether selection or measurement error is a more important source of bias of course depends on the particularities of the sample. In our example dataset selection happens to be a small problem, while measurement error is a large problem. Table 2 compares the two cases: B-JOINT is based on all observations in the sample. Results are shown in the left part of Table 2. B-JOINT-DROP is based on the same model, but shows results only for the 59 consistent observations. The comparison of these two models, B-JOINT and B-JOINT-DROP, allows us to decide whether our results are mainly driven by accounting for the measurement error or by avoiding a selection bias. Both models take into account the measurement error. Both models come to substantial effect sizes for  $\eta \times p$ : -5.16 for B-JOINT-DROP, and -3.64 for B-JOINT. Only B-JOINT avoids the selection bias. Above we have seen that accounting for measurement errors changes the coefficient  $\beta_{\eta \times p}$  by 210%. The selection bias, i.e. the difference between B-JOINT and B-JOINT-DROP, affects the effect size by only 30%.

## 4. Simulation

Should one take into account measurement errors? 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

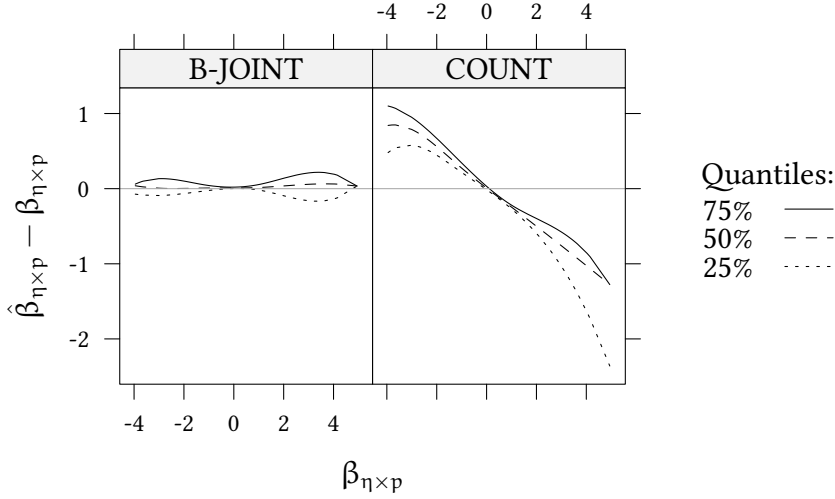


Figure 5: Simulation results: Bias

The figure shows 25%, 50%, and 75% quantiles of a B-spline (df=5) for the bias  $\hat{\beta}_{\eta \times p} - \beta_{\eta \times p}$  and for the different models.

which come in 25 groups.

Behaviour in the risk task follows the logistic model from Equation (5). Contributions in the public good game follow Equation (11). 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  $l \in \{0, \eta, p, \eta \times p\}$ . 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:  $p_{ik}^c \sim \mathcal{B}(6.98, 3.63)$ ,  $\tau_{ik} \sim \Gamma(0.847, 0.2)$ .

For each of the 100 simulations we obtain an estimate for the contribution effects (Equation 6). Here we are specifically interested in  $\beta_{\eta \times p}$ . Figure 5 shows for both methods COUNT and B-JOINT quartiles of the difference between the estimates and the true values,  $\hat{\beta}_{\eta \times p} - \beta_{\eta \times p}$ . We see that B-JOINT performs fairly well. The difference  $\hat{\beta}_{\eta \times p} - \beta_{\eta \times p}$  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 measurement errors are neglected.

## 5. Conclusion

We are clearly not alone in recommending that measurement errors should be taken into account in experimental research. Researchers might have the impression, however, that accounting for measurement errors with repeated measurements is tedious and expensive. Many researchers seem to assume the problem away.

The aim of this paper is to convince the experimental community that it makes sense to take measurement errors seriously, and that it is possible to correct for the resulting bias in a straightforward and simple way, without making the experimental design more complicated.

Here we take advantage of the fact that a single Holt and Laury task is, actually, a multiple price list task, i.e. the task already contains multiple measures. 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: *Psychometrika* 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.
- Beauchamp, Jonathan P., David Cesarini and Magnus Johannesson (2017). “The psychometric and empirical properties of measures of risk preferences”. In: *Journal of Risk and Uncertainty* 54.3, pp. 203–237. DOI: 10.1007/s11166-017-9261-3.
- 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.
- 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.
- Friedman, Daniel, R. Mark Isaac, Duncan James and Shyam Sunder (2014). *Risky Curves: On the Empirical Failure of Expected Utility*. Santa Cruz Department of Economics, Working Paper Series qt87v8k86z. Department of Economics, UC Santa Cruz.
- Gillard, Jonathan (2010). “An overview of linear structural models in errors in variables regression”. In: *REVSTAT–Statistical Journal* 8.1, pp. 57–80.
- Gillen, Ben, Erik Snowberg and Leeat Yariv (2018). “Experimenting with Measurement Error: Techniques with Applications from the Caltech Cohort Study”. In: *Journal of Political Economy* Forthcoming.
- 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.



- Halevy, Yoram (2007). “Ellsberg Revisited: An Experimental Study”. In: *Econometrica* 75.2, pp. 503–536.
- Harrison, Glenn W., Eric Johnson, Melayne M. McInnes and E. Elisabet Rutström (2005). “Temporal stability of estimates of risk aversion”. In: *Applied Financial Economics Letters* 1.1, pp. 31–35.
- Hey, John D. and Chris Orme (1994). “Investigating Generalizations of Expected Utility Theory Using Experimental Data”. In: *Econometrica* 62.6, pp. 1291–1326.
- 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.
- Loomes, Graham (2005). “Modelling the Stochastic Component of Behaviour in Experiments: Some Issues for the Interpretation of Data”. In: *Experimental Economics* 8.4, pp. 301–323. DOI: 10.1007/s10683-005-5372-9.
- Neyman, Jerzy and Elizabeth L. Scott (1948). “Consistent Estimates Based on Partially Consistent Observations”. In: *Econometrica* 16.1, pp. 1–32.
- Niederle, Muriel and Lise Vesterlund (2007). “Do Women Shy Away from Competition? Do Men Compete Too Much?” In: *The Quarterly Journal of Economics* 122.3, pp. 1067–1101.
- 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 (2018). *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.
- Rabe-Hesketh, Sophia, Anders Skrondal and Andrew Pickles (2004). “Generalized multilevel structural equation modeling”. In: *Psychometrika* 69.2, pp. 167–190. DOI: 10.1007/BF02295939.
- Ross, L., R.E. Nisbett and M. Gladwell (2011). *The Person and the Situation: Perspectives of Social Psychology*. Pinter & Martin Limited.
- Sahm, Claudia R. (2012). “How Much Does Risk Tolerance Change?” In: *Quarterly Journal of Finance (QJF)* 2.04, pp. 1–38. DOI: 10.1142/S201013921250019X.
- 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.

## A. Posteriors for $\alpha_\tau$ , $\beta_\tau$ and $p_{ik}^c$

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  $\tau_{ik}$ . In Equation (15) we assume  $p_{ik}^c$  follows a Beta distribution with parameters  $\alpha_c$  and  $\beta_c$  following a Gamma distribution (so that a priori  $p_{ik}^c$  follows an almost uniform distribution). The median of the posterior parameters are  $\alpha_\tau = 7.8$  and  $\beta_\tau = 3.9$ , 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.

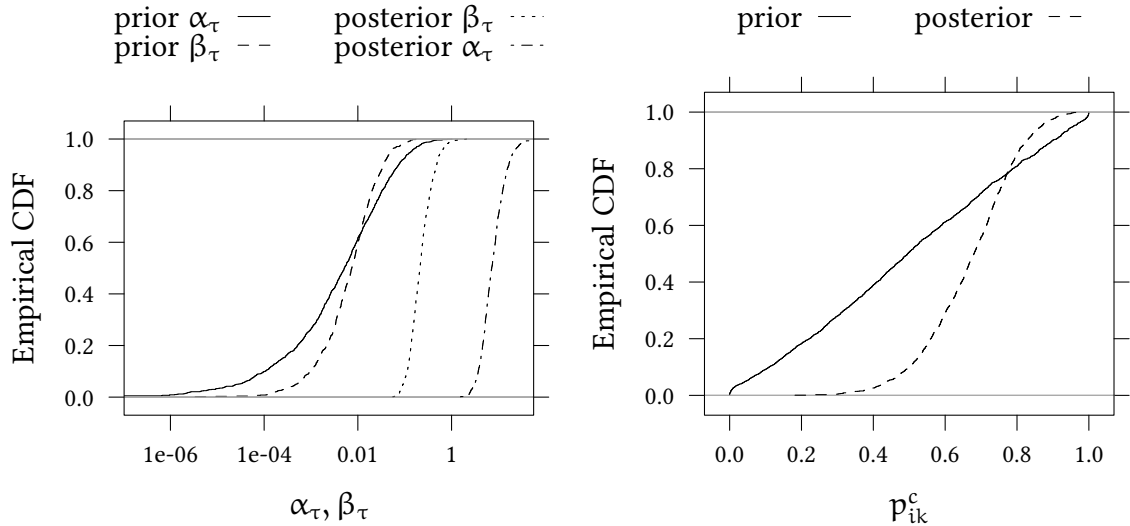


Figure 6: Posteriors for  $\alpha_\tau$ ,  $\beta_\tau$  and  $p_{ik}^c$

In Equation (16) we assume that the precision  $\tau_{ik} = 1/\sigma_{ik}^2$  is drawn from a Gamma distribution. The median of the posterior shape parameter of this distribution is  $\alpha = 7.14$  and the median of the posterior rate parameter is  $\beta = 0.209$ . Figure 7 shows the posterior distribution of  $\sigma_{ik}$  as well as the median values of  $\sigma_{ik}$  for the individual participants. Conceptually, this is not entirely trivial. Often we assume that “consistent” choices are infinitely precise, i.e.  $\sigma_{ik} = 0$ . However, if some choices, here 18% of all participants, are inconsistent, i.e. contain a substantial lack of precision ( $0.168 \leq \sigma_{ik} \leq 0.254$ ), 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 determine the left part of the distribution of  $\sigma_{ik}$ . The value of  $\sigma_{ik} = 0.165$  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.

## B. Example code

In this section we want to demonstrate how measurement errors can be taken into account in practice. We generate some random data. Then we explain how the parameters of the data generating process can be estimated with R (R Development Core Team, 2018) and JAGS (JAGS:43). We assume two types of measurements:

1. In a first stage of the experiment (stage A) a number of binary choices are used to measure a subject specific property, e.g. an attitude towards risk. This measurement contains some noise.

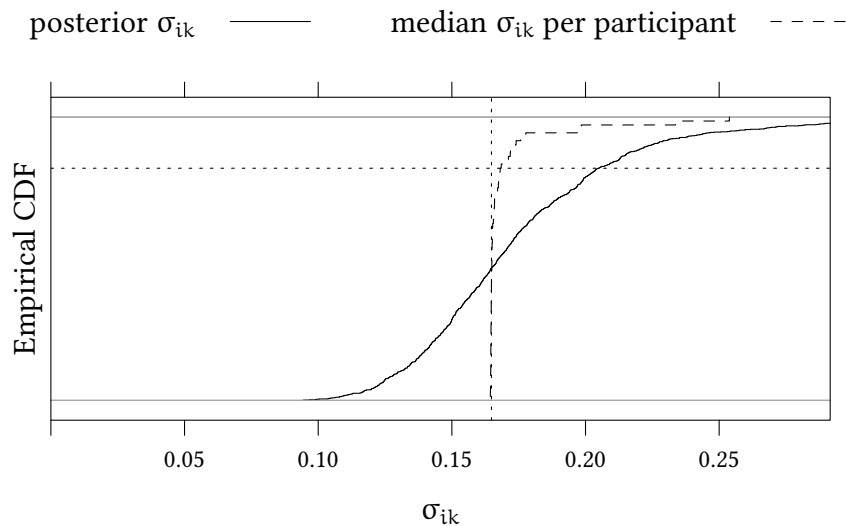


Figure 7: Standard deviation  $\sigma_{ik}$  of choices.

The solid line show the posterior distribution of  $\sigma_{ik} = 1/\tau_{ik}^2$  as in Equation (16). The dashed line shows the distribution of the median of  $\sigma_{ik}$  taken for each participant. The dotted horizontal line denotes the share of 82% consistent observations. The dotted vertical line denotes  $\sigma_{ik} = 0.165$  for the median consistent decision maker.

2. In a second stage of the experiment (stage B) some other behaviour is measured. The noisy measurement from stage A of the experiment is used as an explanatory variable in the second experiment.

In the following we use R version 3.6.0 (2019-04-26). We use runjags 2.0.4-2 to interface with JAGS 4.3.0.

```
library(runjags)
```

We use plyr to generate data for the example.

```
library(plyr)
```

We initialise our random number generator and generate some data:

```
set.seed(1)
```

We simulate a data set with 40 participants.

```
N<-40
```

Task A is based on 10 binary choices:

```

choiceNum.A<-10                                     ## number of choices in A
sid.A<-1:N                                           ## subject IDs
thresh.A<-sort(choiceNum.A*rbeta(N,6.98,3.63))      ## (latent) threshold for
                                                    ##each subject; here we sort the individuals for convenience
tau.A<-rgamma(N,0.847,.2)                            ## (latent) precision of each individual for task A
data.A<-rbind.fill(lapply(1:N,function(sid) {
  choice.A<-rbinom(choiceNum.A,1,                    ## observable choices
                  plogis(((1:choiceNum.A)-thresh.A[sid])*sqrt(tau.A[sid])))
  data.frame(list(sid=sid,choice.A=choice.A,task.A=1:10)) })))

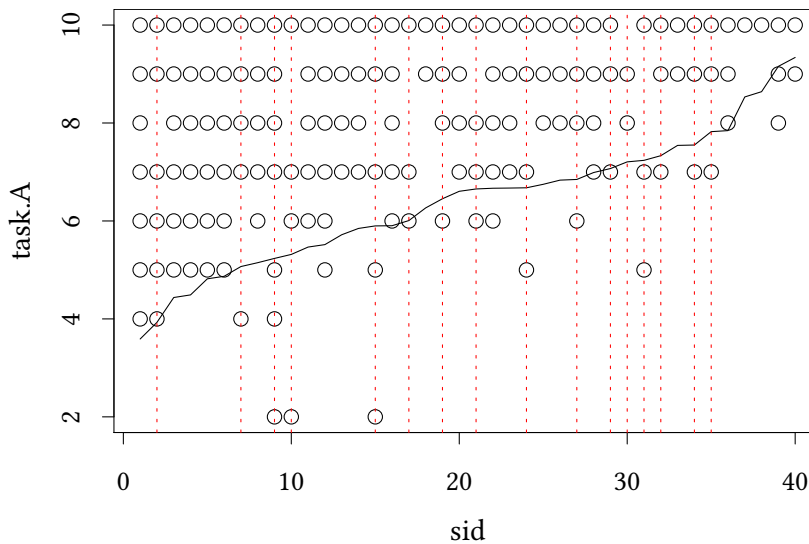
```

The following plot shows the 10 tasks on the vertical axis and the 40 individuals on the horizontal axis. Binary choices are shown as either  $\circ$  or blank. The solid line denotes the (latent) true preference thresh.A for each individual.

```

with(subset(data.A,choice.A==1),plot(sid,task.A))
lines(sid.A,thresh.A)

```



The 16 subjects whose choices are inconsistent are marked with a vertical red line.

## B.1. Independent observations for stage B

We start with the case of independent observations for stage B. Each subject generates one observation. Let us assume that, in addition to the measurement from stage A, we have two controls,  $x_1$  and  $x_2$ . The simulated data in stage B is generated as follows:

```

sid.B<-1:N
x1.B <- runif(N) ## control variable
x2.B <- runif(N) ## control variable

```

```

eps.B<- rnorm(N,0,sqrt(1/sqrt(.1))) ## residual
X.B  <- cbind(1,x1.B,x2.B,thresh.A) ## design matrix
beta <- rbind(-1,1,0,-1)          ## arbitrary coefficients
y.B  <- c(X.B %*% beta + eps.B)

```

The last two lines define  $\beta$  for our linear relationship and calculate the dependent variable  $y.B = X.B\beta + \text{eps}.B$ .

Now we prepare the data for JAGS:

```

sData<-as.list(data.A)
sData[["choiceNum.A"]] <- choiceNum.A
sData[["y.B"]]<-y.B
sData[["X.B"]]<-X.B      ## demeaning does not help convergence here
sData[["K"]]  <- ncol(X.B) ## easier to determine size of beta here
sData[["latentA"]] <- 4   ## latent variable is 4th column
sData[["X.B"]][,4]<-NA    ## we can't observe the latent variable

```

We use the following model in BUGS notation:

```

small.mod <- 'model {
  # binary choices in A:
  for (i in 1:length(choice.A)) {
    choice.A[i] ~ dbern(p[i])
    logit(p[i]) <- (task.A[i] - choiceNum.A*thresh.A[sid[i]]) *
                  pow(tau.A[sid[i]],.5)
  }
  # individual choices in B (here one choice per individual):
  for (i in 1:length(y.B)) {
    y.B[i] ~ dnorm(inprod(X.B[i,],beta),tau.B)
    X.B[i,latentA] <- choiceNum.A*thresh.A[i] # estimate for latent var.
  }
  # subject specific:
  for (s in 1:max(sid)) {
    thresh.A[s] ~ dbeta(a.A,b.A) # switching point
    tau.A[s]    ~ dgamma(s.A,r.A) # precision switching point
  }
  # parameters for switching point:
  a.A ~ dgamma(2,1/2) # switching point
  b.A ~ dgamma(2,1/2) # switching point
  s.A <- (m.A*m.A)/(d.A*d.A) # prec. switching point
  r.A <- m.A/(d.A*d.A) # prec. switching point
  m.A ~ dgamma(1,1) # prec. switching point
  d.A ~ dgamma(10,1) # prec. switching point
  ## residuals:
  tau.B ~ dgamma(mTB*mTB/(dTB*dTB),mTB/(dTB*dTB))
  mTB ~ dgamma(1,1)
  dTB ~ dgamma(1,1)
  ## coefficients:

```

```

for (k in 1:K) {
  beta[k] ~ dnorm (0,.0001) # prior fixed effect for beta
}
}'

```

Here the expression following `# binary choices in A:` corresponds to task A:

$$P(\text{"success"}|p) = \mathcal{L}((p - p_i^c \cdot \sqrt{\tau_i})$$

Different choices in the multiple price list task would be denoted by  $p$ . A “success” would be a specific type of a choice (e.g., the more risky choice in the Holt and Laury (2002) task).

The expression following `# individual choices in B:` corresponds to task B:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 p_i^c + \epsilon_i$$

Here we assume that  $\tau_i$  follows a Gamma distribution with hyperparameters  $s$  and  $r$  which are again chosen such that the distribution for  $\tau_i$  has a mean which follows a Gamma distribution and also a dispersion which follows a Gamma distribution. Choosing fixed hyperparameters does not change estimation results much, but seems to worsen convergence in our experience.

In this example we are using four different chains:

```

runjags.options(silent.jags=TRUE,silent.runjags=TRUE,method="parallel",
               modules="glm",predraw.plots=FALSE)
initJags<-list()
initJags[[1]]<-list(.RNG.seed=1,.RNG.name="base::Mersenne-Twister")
initJags[[2]]<-list(.RNG.seed=2,.RNG.name="base::Super-Duper")
initJags[[3]]<-list(.RNG.seed=3,.RNG.name="base::Wichmann-Hill")
initJags[[4]]<-list(.RNG.seed=4,.RNG.name="lecuyer::RngStream")

```

To confirm that we have a sufficiently vague prior for stage A we look at the distribution of `thresh.A`, i.e. the thresholds for the case of no data:

```

thresh.jags<-run.jags(model=small.mod,within(sData,{y.B<-NA;choice.A=NA}),
                    monitor=c("thresh.A"),n.chains=4,init=initJags,sample=100)
thresh.df<-data.frame(as.mcmc(thresh.jags))

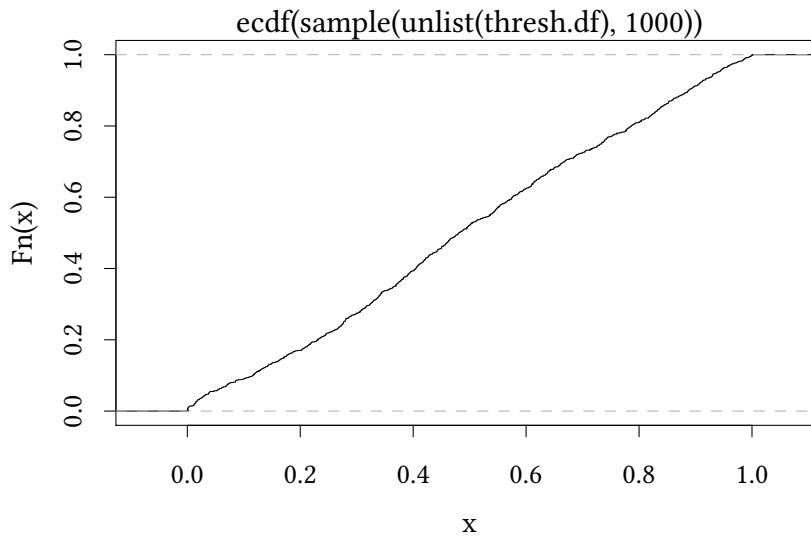
```

The following plot shows the cumulative distribution of `thresh.A`.

```

plot(ecdf(unlist(thresh.df)))

```



The distribution is almost a straight line, i.e. with our prior all threshold values are almost equally likely.

In a next step we use the same model, now with data for choices in stage A and B:

```
run.jags(model=small.mod,data=sData,monitor=c("beta"),n.chains=4,init=initJags)
```

JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

	Lower95	Median	Upper95	Mean	SD	Mode	MCerr	MC%ofSD
beta[1]	-4.8177	-0.97878	2.9369	-0.96509	1.9712	--	0.027153	1.4
beta[2]	-0.49231	1.555	3.5872	1.5472	1.0405	--	0.0067797	0.7
beta[3]	-2.3593	-0.38987	1.6219	-0.38618	1.0175	--	0.0069433	0.7
beta[4]	-1.499	-0.97826	-0.42573	-0.98024	0.27311	--	0.0041521	1.5

	SSeff	AC.10	psrf
beta[1]	5270	0.11439	1.0004
beta[2]	23553	0.0056848	1.0004
beta[3]	21474	0.0021126	1.0001
beta[4]	4326	0.14232	1.0005

Total time taken: 14.5 seconds

## B.2. Repeated observations for stage B

Let us next simulate data with repeated observations for stage B. Here we assume that we have matching groups of size 4. Furthermore, each participant makes not one, but 2 choices. We introduce one random effect  $\nu_k$  for the group and one random effect for the subject  $\nu'_{ik}$ .

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 p_{ik}^c + \nu_k + \nu'_{ik} + \epsilon_{ikt}$$

The data for stage B is simulated as follows:

```
groups.B      <-4
choiceNum.B  <-2
sid.B<-rep(1:N,each=choiceNum.B) ## subject ID
gid.B<-sid.B %% groups.B + 1     ## group ID
x1.B <- runif(N*choiceNum.B)
x2.B <- runif(N*choiceNum.B)
eps.B<- rnorm(N*choiceNum.B,0,sqrt(1/sqrt(.1))) ## residuals
nu1.B<- rnorm(groups.B,0,sqrt(1/5))[gid.B] ## random effect for each group
nu2.B<- rnorm(N,0,sqrt(1/3.5))[sid.B]     ## random effect for each subject
X.B  <- cbind(1,x1.B,x2.B,thresh.A)      ## design matrix
beta <- rbind(-1,1,0,-1)                ## arbitrary coefficients
y.B  <- c(X.B %*% beta + nu1.B + nu2.B + eps.B)
```

Now we prepare the data for JAGS:

```
lData<-as.list(data.A)
lData[["choiceNum.A"]] <- choiceNum.A
lData[["sid.B"]]<-sid.B
lData[["gid.B"]]<-gid.B
lData[["y.B"]]<-y.B
lData[["X.B"]]<-X.B ## demeaning X.B is not necessary here
lData[["K"]] <- ncol(X.B) ## easier to determine size of beta here
lData[["latentA"]] <- 4 ## latent variable is 4th column
lData[["X.B"]][,4]<-NA ## we can't observe the latent variable
```

We use the following model in BUGS notation:

```
large.mod <- 'model {
  # binary choices in A:
  for (i in 1:length(choice.A)) {
    choice.A[i] ~ dbern(p[i])
    logit(p[i]) <- (task.A[i] - choiceNum.A*thresh.A[sid[i]]) *
                  pow(tau.A[sid[i]],.5)
  }
  # individual choices in B:
  for (i in 1:length(y.B)) {
    y.B[i] ~ dnorm(inprod(X.B[i,],beta) + nu1[gid.B[i]] + nu2[sid.B[i]],tau[3])
    ## nu1, nu2 are the two random effects
    X.B[i,latentA] <- choiceNum.A*thresh.A[sid.B[i]] # estimate for latent var.
  }
  # subject specific:
  for (s in 1:max(sid)) {
```



```

    thresh.A[s] ~ dbeta(a.A,b.A)      # switching point for A
    tau.A[s]    ~ dgamma(s.A,r.A)     # precision switching point
    nu2[s]      ~ dnorm (0,tau[2])    # random effect subject for B
  }
  # group specific
  for (g in 1:max(gid.B)) {
    nu1[g] ~ dnorm (0,tau[1])        # random effect group for B
  }
  # parameters for switching point:
  a.A ~ dgamma(2,1/2)                # switching point
  b.A ~ dgamma(2,1/2)                # switching point
  s.A <- (m.A*m.A)/(d.A*d.A)        # prec. switching point
  r.A <- m.A/(d.A*d.A)               # prec. switching point
  m.A ~ dgamma(1,1)                  # prec. switching point
  d.A ~ dgamma(10,1)                 # prec. switching point
  ##
  for (k in 1:3) { ## priors for prec. of residual and random effects:
    tau[k] ~ dgamma(mTB[k]*mTB[k]/(dTB[k]*dTB[k]),mTB[k]/(dTB[k]*dTB[k]))
    mTB[k] ~ dgamma(1,1)
    dTB[k] ~ dgamma(1,1)
  }
  for (k in 1:K) {
    beta[k] ~ dnorm (0,.0001) # prior fixed effect for beta
  }
}'

```

```
run.jags(model=large.mod,data=lData,monitor=c("beta"),n.chains=4,init=initJags)
```

JAGS model summary statistics from 40000 samples (chains = 4; adapt+burnin = 5000):

	Lower95	Median	Upper95	Mean	SD	Mode	MCerr	MC%ofSD	SSEff
beta[1]	-4.9048	-1.1461	2.4637	-1.1345	1.8672	--	0.023343	1.3	6398
beta[2]	-2.293	-0.82914	0.69036	-0.82166	0.75646	--	0.0040438	0.5	34994
beta[3]	-2.4435	-1.0431	0.33221	-1.0418	0.70468	--	0.0038875	0.6	32859
beta[4]	-1.2578	-0.7344	-0.19558	-0.732	0.26924	--	0.0036276	1.3	5509

	AC.10	psrf
beta[1]	0.091985	1.0001
beta[2]	0.00398	1.0001
beta[3]	0.0095094	1
beta[4]	0.11116	1.0001

Total time taken: 16.6 seconds

## C. Instructions

**General Instructions** In the following experiment, you can earn a substantial amount of money, depending on your decisions. It is therefore very important that you read these instructions carefully.

**During the experiment, any communication whatsoever is forbidden.** If you have any questions, please ask us. Disobeying this rule will lead to exclusion from the experiment and from all payments.

You will in any case receive 4 € for taking part in this experiment. In the first two parts of the experiment, we do not speak not of €, but instead of Taler. Your entire income from these two parts of the experiment is hence initially calculated in Taler. The total number of Taler you earn during the experiment is converted into € at the end and paid to you in cash, at the rate of

$$1 \text{ Taler} = 4 \text{ Eurocent.}$$

The experiment consists of four parts. We will start by explaining the first part. You will receive separate instructions for the other parts.

**Part One of the Experiment** In the first part of the experiment, there are two roles: A and B. Four participants who have the role A form a group. One participant who has the role B is allocated to each group. The computer will randomly assign your role to you at the beginning of the experiment.

On the following pages, we will describe to you the exact procedure of this part of the experiment.

**Information on the Exact Procedure of the Experiment** This part of the experiment has two steps. In the first step, role A participants make a decision on contributions to a project. In the second step, the role B participant can reduce the role A participants' income. At the start, each role A participant receives 20 Taler, which we refer to in the following as the endowment. Role B participants receive 20 points at the start of step 2. We explain below how role B participants may use these points.

**Step 1:** In Step 1, only the four role A participants in a group make a decision. Each role A member's decision influences the income of all other role A players in the group. The income of player B is not affected by this decision. As a role A participant, you have to decide how many of the 20 Taler you wish to invest in a project and how many you wish to keep for yourself.

If you are a role A player, your income consists of two parts:

1. the Taler you have kept for yourself ("income retained from endowment")
2. the "income from the project". The income from the project is calculated as follows:  
Your income from the project = 0.4 times the total sum of contributions to the project

Your income is therefore calculated as follows:

$$(20 \text{ Taler} - \text{your contribution to the project}) + 0.4 * (\text{total sum of contributions to the project}).$$

The income from the project of all role A group members is calculated according to the same formula, i.e., each role A group member receives the same income from the project. If, for example, the sum of the contributions from all role A group members is 60 Taler, then you and all other role A group members receive an income from the project of  $0.4 * 60 = 24$  Taler. If the role A group members have contributed a total of 9 Taler to the project, then you and all other role A group members receive an income from the project of  $0.4 * 9 = 3.6$  Taler.

For every Taler that you keep for yourself, you earn an income of 1 Taler. If instead you contribute a Taler from your endowment to your group's project, the sum of the contributions to the project increases by 1 Taler and your income from the project increases by  $0.4 * 1 = 0.4$  Taler. However, this also means that the income of all other role A group members increases by 0.4 Taler, so that the total group income increases by  $0.4 * 4 = 1.6$  Taler. In other words, the other role A group members also profit from your own contributions to the project. In turn, you also benefit from the other group members' contributions to the project. For every Taler that another group member contributes to the project, you earn  $0.4 * 1 = 0.4$  Taler.

Please note that the role B participant cannot contribute to the project and does not earn any income from the project.

**Step 2:** In Step 2, only the role B participant makes decisions. As role B participant, you may reduce or maintain the income of every participant in Step 2 by distributing points.

At the beginning of Step 2, the four role A participants and the role B participant are told how much each of the role A participants has contributed to the project.

As a role B player, you now have to decide, for each of the four role A participants, whether you wish to distribute points to them and, if so, how many points you wish to distribute to them. You are obliged to enter a figure. If you do not wish to change the income of a particular role A participant, please enter 0. Should you choose a number greater than zero, you reduce the income of that particular participant. For each point that you allocate to a participant, the income of this participant is reduced by 3 Taler.

The total Taler income of a role A participant from both steps is hence calculated using the following formula:

$$\text{Income from Step 1} - 3 * (\text{sum of points received})$$

Please note that Taler income at the end of Step 2 can also be negative for role A participants. This can be the case if the income-subtraction from points received is larger than the income from Step 1. However, the role B participant can distribute a maximum of 20 points to all four role A members of the group. 20 points are the maximum limit. As a role B participant, you can also distribute fewer points. It is also possible not to distribute any points at all.

If you have role B, please state your reasons for your decision to distribute (or not to distribute) points, and why you distributed a particular number of points, if applicable. In doing this, please try to be factual. Please enter your statement in the corresponding space on your screen. You have 500 characters max. to do this. Please note that, in order to send your statement, you will have to press “Enter” once each time. As soon as you have done this, you will no longer be able to change what you have written.

The income of the role B participant does not depend on the income of the other role A participants, nor on the income from the project. For taking part in the first part of the experiment, he or she receives a fixed payment of

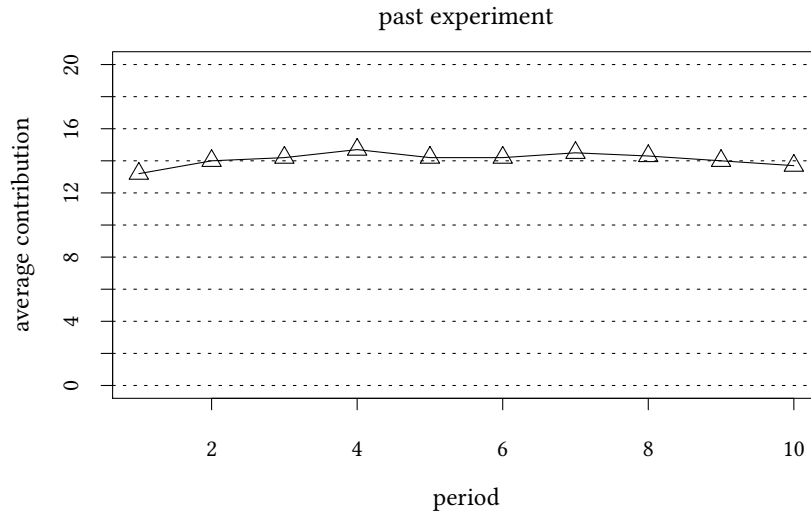
1 €.

In addition, the role B participant receives the sum of 0.01 € for each point that he or she did not distribute. Once all participants have made their decisions, your screen will show your income for the period and your total income so far.

After this, the first part of the experiment ends. You will then be told what your payment is for this part of the experiment. Hence, you will also know how many points you and all other participants have been given by player B.

**Experiences from an Earlier Experiment** For your information, we give you the following graph, which tells you the average contributions made in a very similar experiment that was conducted in this laboratory.

In this experiment, too, there were groups of 4 role A participants and one role B participant each. The role A participants’ income was calculated in exactly the same way. The experiment had 10 equal periods. The role B participant also had 20 points at his disposal in each period. At the end of each period, the role A participants were told how much each of the other participants had contributed and how the role B participant had reacted to this.



**Part Two of the Experiment** The second part of the experiment consists of 10 repetitions of the first part. Throughout the entire second part, all participants keep the role they had in the first part of the experiment. The computer randomly rematches the groups of four in every period. In each period, the computer randomly assigns a role B participant to each group.

As a reminder: In each period, each role A participant receives 20 Taler, which may be contributed to the project entirely, in part, or not at all. For each period, calculating the income from the project for the role A participants in a group happens in exactly the same way as it did in the first part of the experiment. In each period, each role B participant receives 20 points, which may be used to reduce the income of the players A in the group. For each point that a role A participant receives in a period, 3 Taler are subtracted. For each point that a role B participant does not use, he or she is given the sum of 0.01 €. In addition to the income from the points retained, each role B participant receives a flat fee of 10 € for participating in this second part of the experiment.

At the beginning of Step 2 of each period, the four role A participants and the role B participant are told how much each of the role A participants contributed to the project.

Please note that the groups are rematched anew in each period.

After each period, you are told about your individual payoff. You are therefore also informed how many points you and the other participants have been assigned by the role B participant.

**Part Three of the Experiment** We will now ask you to make some decisions. In order to do this, you will be randomly paired with another participant. In several distribution decisions, you will be able to allocate points to this other participant and to yourself by repeatedly choosing between two distributions, 'A' and 'B'. The points you allocate to yourself will be paid out to you at the end of the experiment at a rate of 500 points = 1 €. At the same time, you are also randomly assigned to another participant in the experiment, who is, in turn, also able to allocate points to you by choosing between distributions. This participant is not the same participant as the one to whom you have been allocating points. The points allocated to you are also credited to your account. The sum of all points you have allocated to yourself and those allocated to you by the other participant are paid out to you at the end of the experiment at a rate of 500 points = 1 €. Please note that the participants assigned to you in this part of the experiment are not the members of your group from the preceding part of the experiment. You will therefore be dealing with other participants.

The individual decision tasks will look like this:

Possibility A		Possibility B	
Your points:	The points of the participant of the experiment allocated to you:	Your points:	The points of the participant of the experiment allocated to you:
0	500	304	397

In this example: If you click 'A', you give yourself 0 points and 500 points to the participant allocated to you. If you click 'B', you give yourself 304 points and 397 points to the participant allocated to you.