Round effects in economic experiments - A novel probabilistic programming approach to time-variant response behaviour

Leyens, Alexa[1]; **Feisthauer,** Philipp[2]**; Börner,** Jan[2]**, Hartmann,** Monika[2]**, Storm,** Hugo[1]

[1]Institute for Food and Resource Economics, Chair of Data science in agricultural economics, Rheinische Friedrich-Wilhelms-University Bonn, Meckenheimer Allee 174, 53115 Bonn, Germany

[2]Institute for Food and Resource Economics, Chair of Economics of Sustainable Land Use and Bioeconomy, Rheinische Friedrich-Wilhelms-University Bonn, Nussallee 19, 53115 Bonn, Germany

Prepared for the 98th annual conference of the Agricultural Economics Society at the University of Edinburgh, United Kingdom, 18th – 20th March 2024

Abstract

Round effects can arise in any kind of economic experiment, where participants have to make decisions over a course of various rounds. They are associated with changes of preferences and patterns of response variance, they may counteract or distort experimental treatment effects and ultimately result in the biased estimation of the “true” underlying preferences. To investigate how exactly round effects occur and can be captured statistically, we have designed and conducted a multi-round study with German agricultural students. We develop a novel Bayesian Probabilistic Programming approach to assess to which extent preference learning, institutional learning and fatigue effects influence the response behaviour of survey participants. We find strong evidence for fatigue effects, since the response behaviour of our participants became increasingly variant over the course of the experiment. We could also falsify the assumptions of institutional and preference learning for our participant group. Our results emphasize the importance of modelling round effects in economic experiments. The results and the developed modelling framework will be of interest to both, experimental agricultural economists and policy developers, who interpret and apply findings from business simulation games and similar multi-period experimental studies.

Keywords: Economic experiment, business simulation game, round effects, Bayesian Probabilistic Programming, smart farming technologies

JEL Codes: D91, B41, Q16, C11

1. Introduction

Experiments with human subjects are an important and frequently used tool in all fields of economic research, including agricultural economics. Oftentimes, these experiments feature repeated choice tasks or experimental settings which entail multiple game rounds in order to capture the effects of experimental treatments or analyse how respondents change their behaviour across different consecutive scenarios (Blasch et al., 2022; Feisthauer et al., 2024; Fleming et al., 2021; Musshoff & Hirschauer, 2014; Thomas et al., 2019; Thompson et al., 2019). However, multi-period experimental designs have been associated with distortion in behavioural patterns and preferences of respondents which can lead to erroneous estimation of subjects' preferences and over- or underestimation of experimental treatment effects. In the domain of consumer research, this phenomenon has received some attention, with a major focus on capturing how the order in which respondents are confronted with choice tasks could affect their response behaviour (Day et al., 2012). However, there has not been significant research of such effects in the experimental economics and universally agreed upon definitions or methodologies around round-dependent response behaviour changes have not been established.

In this paper, we focus on three specific types of effects which have, to the best of our knowledge, not been explicitly considered in experimental agricultural economics research. Specifically, we investigate institutional learning, preference learning and fatigue effects. Institutional learning describes the familiarisation of study participants with the experimental setting. In the first experimental round, respondents are said to be overwhelmed or unfamiliar with a given task, which limits their ability to respond according to their true preferences. In a second round, however, respondents tend to act closer to their true preferences, causing the experimenter to observe drastic changes in response behaviour, even if the task as such is not altered between rounds. Preference learning captures the consolidation of respondents' preferences as they move through the experiment. While initially being unsure, respondents develop their preferences as a consequence of in-game feedback mechanisms. When repeatedly asked to perform the same task across multiple rounds, the experimenter here observes a reduction in the adjustment of choices from one round to the next, i.e., the variance of the adjustments across rounds is smaller in later compared to earlier experimental rounds. Finally, fatigue arises when respondents become tired of or bored by the cognitive load of repetitive choice tasks and continuously lose vigour during participation. In reverse to preference learning, the experimenter observes seemingly more random behaviour represented by an increasing variance of the choice adjustments from round to round as the experiment progresses.

Our study is motivated by a lab-in-the-field experiment in Feisthauer et al. (2024), which we replicate with agricultural students using an adapted experimental design to enable the quantification and explicit accounting of the outlined effects. In the precursor study, evidence for the existence of such a temporal effect was discovered when aiming to analyse the influence of hypothetical policy treatments on German farmers' digital weeding technology preferences. Specifically, they found that the control group, albeit not being treated, changed their responses in a similar magnitude and direction when moving from the first to the second game round. In Feisthauer et al. (2024), this observation was coined 'round effect' and we henceforth adhere to this term in the present paper.

By defining and testing three theoretically and statistically clearly defined types of round effects, we can deliver a meaningful contribution to the debate around how to design, analyse and interpret the outcomes of multi-round studies. Firstly, this contribution entails that we empirically capture round effects by conducting a specifically-designed experiment. Secondly, we can use this experiment data as a base to evaluate which of the three hypothesized effects are likely to be present in the obtained data. Additionally, we can show that round effects can impact the estimation of treatment effects by using simulated data and our generative model. Building on this insight, we have deducted that the quantification and analysis of round effects is of high relevance to scientists and policy makers because they influence the conclusions drawn from multi-round studies. Lastly, we propose and apply a novel PP approach to explicitly capture round effects in experimental data. This approach further allows us to model and control for round effects in the estimation of treatment effects and thereby helps in producing more robust policy conclusions from complex experimental setups.

The remaining paper is structured as follows: In Section 2, we give an overview of the current state of research on effects arising from multi-period designs in experimental economics and we motivate our hypotheses to test institutional learning, preference learning and fatigue as specific candidates of round effects. In Section 3, we explain details of the adapted experimental design, the data acquisition process and sample. Section 4 describes the statistical and empirical model for our approach and further details on the implementation of the research hypotheses into the model using a Bayesian Probabilistic Programming framework. Here, we also give detailed insights into the definitions of our priors and show how we tested our model by generating conditioned dummy datasets. Section 5 presents the results and the discusses the implications to be drawn. Finally, we conclude the paper in section 6 with the limitations and by pointing out recommendations for future research and policy.

2. Theoretical framework

Previous literature presents several definitions, interpretations and approaches to analyse phenomena of temporal effects in multi-round experimental settings. In this section, we thus provide a summary of several explanations regarding behaviour changes induced by multi-round designs and discuss their respective implications for preference estimation in experimental studies in a broader context. For consistency and since several somewhat overlapping concepts exist, we henceforth subsume them under the umbrella of 'round effects'. We focus on three types of round effects that may have been pivotal causes of the observations made in the precursor study, which we therefore aim to capture and quantify by the novel methodological approach in the present study.

The effects pertaining to the ordering of repeated valuation tasks in consumer research have been studied most prominently. Here, a consumer is repeatedly confronted with consumption decisions and asked to evaluate goods, either in a stated or revealed preference format. Against this backdrop, Day et al. (2012) list and test several types of effects and categorise them as 'order effects'. The authors emphasise that the order in which respondents are confronted with choice task may be of particular relevance for the experimental findings. In the following, we delve deeper into specific phenomena and derive our research hypotheses.

2.1. Institutional learning

According to Day et al. (2012), one potential effect is 'institutional learning', which is an order effect causing a convergence in consumer response behaviour after a larger initial change. It is assumed that participants are initially confused or overwhelmed by the unfamiliar institutional setting, i.e. the experimental tasks, and hence respond differently in the first round compared to their actual preferences voiced in the following rounds. However, once they have understood the task, participants adjust their response behaviour in the subsequent rounds and thereby express their actual preferences. On a similar note, Chou et al. (2009) used the term 'failure of game form recognition' to describe their observation of participants who changed their responses in a social preference survey after the first round, presumably due to confusion about the setup of the survey. Generally speaking, the phenomenon of institutional learning manifests in experimental research data in a change in participants' response behaviour between the first and later rounds round independent of experimental treatments.

Institutional learning is more likely to occur in more complex experiments and is characterised by a switch in response behaviour when moving from the first to later rounds of a multi-round experimental setup. Arguably, this applies for our experimental approach, resulting to our first research hypothesis:

H1 (Institutional learning): Participants show a change in response behaviour when moving from the baseline to the first round.

2.2. Preference learning

Similar to institutional learning, Braga and Starmer (2005) and Day et al. (2012) find that their survey participant pool shows a high variance in the responses of the first round(s) and decreasing variance in response behaviour in later rounds. Supposedly, in the beginning of the experiment, respondents are unaware or have no preferences towards the available choice options and only learn or discover them throughout the experiment. By becoming more familiar with the topic and the tasks at hand, receiving feedback from the experimental setup and learning about the consequences of their choices, the respondents become progressively more coherent in their responses and make choices which better reflect their true preferences. In a study by Brouwer et al. (2010), in which respondents were explicitly asked about their choice certainty after each round, preference learning could be confirmed and self-reported choice certainty increased in later rounds. Similarly, Carlsson and Martinsson (2001), Brown et al. (2008) and Reed Johnson and Bingham (2001) find evidence for preference learning in their studies, where respondents seemed to need 'warm-up rounds' to learn about the consequences and trade-offs associated with their decisions. This leads to a gradual decrease in the response variability.

One specific form of preference learning of scholarly emphasis is called the 'anchoring hypothesis' or 'anchoring effect' (e.g. Ladenburg & Olsen, 2008). Here, respondents with weak or uncertain preferences will use the options in the first round as reference points for choice options in subsequent rounds. In other settings, this is referred to as a 'starting point effect' (e.g. Carlsson et al., 2012; Ladenburg & Olsen, 2008) or the concept of 'coherent arbitrariness' (Ariely et al., 2003).

The preference learning effect is assumed to be stronger in experiments where participant's in-game decisions have real-life consequences in form of monetary pay-outs, and where they are also informed about their achieved pay-outs after each decision they took. Additionally, unfamiliar topics and/ or choice structures are also assumed to make preference learning more likely to happen in the course of multi-round experiments. Given these aspects, we have reason to assume that participants in our study did not go through the survey with stable choices but discovered their preferences over the course of the game. Accordingly, we formulate the second hypothesis:

H2 (Preference learning): Participants show a reduction of behaviour change variance when moving from the first to the second round.

2.3. Fatigue effects

As opposed to institutional and preference learning, Day et al. (2012) also discuss 'fatigue effects' where respondents show increasing randomness in their response behaviour over the course of the experiment. In line with Swait and Adamowicz (2001), Day et al. (2012) argue that participants will feel increasingly exhausted by the cognitive load of making adequate choices as the experiment progresses. Furthermore, respondents are assumed to develop a favourable attitude towards a status quo provision level (for public good related choice tasks) or one specific decision factor considering the good in order to minimise the mental strain of the decision task. By confirming that the complexity of choice tasks has a strong influence on choice consistency, DeShazo and Fermo (2002) also argue for the presence of fatigue effects and accumulation of mental load during repeated choice multi-period experiments.

Similar to the fatigue effect, the 'failing credibility' theory assumes increased randomness and favouring of status quo options due to a perceived lack of credibility and sincerity of the experimental setup (Carson & Groves, 2007).

Fatigue effects are assumed to be stronger for experiments with complex experimental setups and low variability in the given choice tasks, as this increases the mental load of decision-making. Hence, we assume that respondents in our experiment might encounter fatigue in the last round and we thus formulate the third hypothesis:

H3 (Fatigue): Participants show an increase of behaviour change variance when moving from the second to the last (fourth) round.

In line with the theory associated with hypotheses H1 to H3, we considered the respondents' choice changes between rounds instead of the absolute observed values to explicitly model changes in response variance. Details are outlined in Section 4 below.

3. Empirical strategy and data

3.1. Experimental design

Our experimental design is based on a recent lab-in-the-field experiment on German farmers' intentions to adopt smart weeding technologies (Feisthauer et al. 2024).[1] We adjusted the experiment for our purposes of studying round effects. Similar to Feisthauer et al. (2024), participants in our experiment played a business simulation in which they chose from three different technologies to conduct weed management on 50 hectares of hypothetical farmland. The weeding technologies were characterised by different private profits and environmental impacts, and could be allocated in any desired ratio. While in round one (baseline), all farmers received identical experimental conditions, in round two the sample was randomly assigned to a control and a treatment, i.e. a policy scenario, which was hypothesised to influence farmers toward choosing a higher share of more eco-friendly weeding technologies. In going beyond Feisthauer et al. (2024), however, we had half of the sample play two and four rounds, respectively, to be able to accommodate the hypothesised effects according to H2 and H3. Moreover and for parsimony, we introduced only one policy treatment after round one (baseline), i.e. we only retained a subsidy for eco-friendly smart weeding technologies identical to the precursor study. Both adaptations together yielded four groups to which participants were randomly assigned, i.e. two control and two treatment groups each of which one played either two and four game rounds, respectively (Table 1).

Following the business simulation game, participants were asked to answer several questions regarding their sociodemographic background, prior knowledge of and experience with smart farming technologies. Furthermore, a set of attitudinal measures was recorded, namely participants' levels of pro-environmental attitude, their personal innovativeness, the degree of trust in the security and privacy of farming data collected by smart and autonomous crop farming technologies and, lastly, their fatigue

[1] Details regarding the experimental design of Feisthauer et al. (2024) are available in the respective project repository on OSF (Link).

during the experiment. All items were measured via multiple questions operationalized on 7-point Likert scales which entered the final analysis as mean scores. The study, which was programmed in Qualtrics, made use of a dynamic incentivisation mechanism to model real life consequences of participants' in-game behaviour.[2]

Table 1. Flow of experiment and randomised assignment to treatment groups.

Treatment	**Round**			
	1	*2*	*3*	*4*
Subsidy = 0	Baseline	Group 1	-	
		Group 3	Group 3	
Subsidy = 1		Group 2	-	
		Group 4	Group 4	

3.2. Sample

Data were collected in November and December 2022. The online survey was distributed via email through multiple channels within the authors' networks to agricultural students from German universities, technical colleges and vocational schools (see Supplementary information for details). In total, 403 students clicked on the link to begin the survey of which only 277 completed it (dropouts n=126, dropout rate=31.3%). Furthermore, we excluded participants who did not consent to the use of their anonymized experimental responses (n=2), those who indicated they were no students (n=8), and those who answered a test question regarding the experimental instructions incorrectly (n=64) which yielded the final data set of 203 usable observations (Table 2). Among all usable survey submissions, the average participation duration across all groups was 37.1 minutes. Interestingly, there was no significant difference when looking at the two round cohort and the four round cohort separately who required 37.1 and 37.2 minutes, respectively. The average participant's age was 22.7 years and the sample was approximately composed of 55% male and 45% female student. While about half of the sample did not have any prior practical farming experience, 22.7% were accustomed to working on their family business, 14.8% had several years of experience and 14.3% had completed an internship. A large majority of participants were university students (83%), while the remainder were enrolled in technical colleges (13%) and vocational schools. Exactly half of the sample had prior experience or (and) knowledge of smart farming technologies (SFT) and participants did not perceive the experimental tasks as very fatiguing (mean=2.7). Lastly, regarding the attitudinal measures, participants had a high average level of pro-environmental attitude (mean=5.9) while expressing moderately high levels of personal innovativeness (mean=4.8) and trust in the security and privacy of farming data (mean=4.1).

[2] Please refer to the Supplementary Information for the details regarding the measurement and descriptive statistics for the attitudinal constructs, the mechanism and results of pay-outs of incentives.

Table 2. Sociodemographic sample characteristics.

Variable	Mean (SD), count (share)[b]
Age in years	22.7 (2.7)
Gender	111 (55.0%) male
	91 (45.0%) female
	1 (0.5%) diverse
Prior farming experience	98 (48.3%) no experience
	46 (22.7%) work experience on family farm
	30 (14.8%) several years
	29 (14.3%) completed an internship
Type of educational institution	169 (83.0%) university
	26 (13.0%) technical college
	8 (3.9%) vocational school
Experience/knowledge of SFT (yes/no)	102 (50.0%)
Fatigue[a]	2.7 (1.1)
Pro-environmental attitude[a]	5.9 (1.0)
Personal innovativeness[a]	4.8 (1.4)
Trust in security and privacy in farming data[a]	4.1 (1.2)

n=203

[a] 7-point Likert scale multi-item construct to evaluate participants' self-rated fatigue during the survey. See Supplementary information for details.

[b] Percentages are rounded to one decimal point and do not necessarily add up to 100%.

4. Analysis

We employ a novel Bayesian Probabilistic Programming (PP) approach that allows to 1) identify if round effects are presents, 2) to quantify them in our specific setting and 3) to explicitly model and control for them in a general way when estimating treatment effects from multi-round experiments (Storm et al, in progress). To motivate our approach, we outline key advantages compared to frequentist approaches mostly used in previous studies (Brouwer et al., 2010, Carlsson et al., 2012).

First, the Bayesian approach allows us to explicitly model the variances in respondents' behaviour changes. Hence, the variance is an explicit, clearly defined part of the model. Achieving something similar in a frequentist setting seems to be less straightforward and a possible approach might resort to indirectly derive the variance from multiple, separately estimated, models. Deriving the correct statistical properties for estimates obtained in this way seems far from obvious. Having the variance as an explicit part of the (Bayesian) model also implies that we obtain a proper posterior distribution for the variance in each round. This allows a much clearer and more intuitive interpretation, compared to frequentist confidence intervals, even if we were able to derive them properly. Not having to estimate the variance from separate models seems particularly useful in our setting with varying sample sizes across rounds.

Second, PP requires the modeler to be very specific about the assumed data generation process (DGP) and the underlying assumption made for the empirical model. While this would in principle also be possible in a frequentist setting, it is not regularly done. PP also ensures that the empirical estimation procedure exactly matches the assumed DGP, thereby achieving consistency between the theoretical assumptions and the empirical model (see Storm et al, in progress for an in-depth discussion of this point).

Third, having a fully defined DGP enables us to inspect and test our code, model formulation and inference step extensively using generated dummy datasets before considering the actual data. The DGP allows us to generate data that follows our hypotheses precisely, which then enables us to test whether our modelling procedure is able to identify the properties we aim to estimate in a setting where the "true" properties of the data are known. Below, we generate data for specific sets of coefficients, motivated by our theoretically derived hypotheses. Using this conditioning, we can inspect how outcomes and estimated effects would look like if the hypotheses are indeed present. This step serves as a useful quality assurance and contributes to building trust in the obtained empirical results.

In the following paragraphs, we first describe a general Data Generating Process motivated by our theoretically derived hypotheses. Secondly, we describe how we translate this DGP to a specific statistical model. In the final part of this section, we present how we use the DGP to test the entire estimation approach using synthetic data. Additionally, we explore how round effects can impact the estimation of potential treatment effects.

4.1. Theoretical model

Building on our hypotheses H1 to H3, we developed a simple theoretical model reflecting participants behaviour when moving from round to round. In each round, participants decide which of the following three weeding technologies they wanted to allocate to each of their 50 hectares of hypothetical farm land: Broadcast application (BC), Spot Spraying (SS) and Weeding Robot (WR). Hence, the output deducted from the experiment can be described as hectare shares per round (r) and technology (tech) for each participant.

$$y_{i,r} = \left(bc_{i,r}\ ss_{i,r}\ wr_{i,r} \right) = \left(BC_{i,r}/50\ SS_{i,r}/50\ WR_{i,r}/50 \right)$$

The behavioural changes of participants when moving between rounds are then given by the differences of $y_{i,r}$ between rounds:

$$\Delta\, y_{i,r} = y_{i,r} - y_{i,r-1}$$

Given the experiment design, we observe $\Delta\, y_{i,r}$ either one or three times per participant, depending on the number of rounds played.

Considering our hypotheses, we assume that $\Delta\, y_{i,r}$ is a function of a constant that varies across rounds (roundeffect $\theta_{r,tech}$), a treatment effect (β_{tech}), and an additional term error term ($\epsilon_{i,r}$):

$$\Delta\, y_{i,r} = \theta_{r,tech} + \beta_{tech} * treat + \epsilon_{i,r}$$

Here, *treat* is a binary variable indicating being assigned to the treatment (=1) or control group (=0) and the β_{tech} parameter models the strength and direction of a treatment effect per technology. The $\theta_{r,tech}$

parameter covers how participants choice behaviour changes across rounds, for each technology. It thus allows us to represent H1 (institutional learning), reflecting how individuals (on average) adjust their responses across rounds. Under H1, we would expect that $\theta_{0,tec} \neq 0$ for all technologies. The noise parameter $\epsilon_{i,r}$ describes additional factors, beyond round effect and treatment effect, that determine behaviour changes per round and participant. Specifically, considering the standard deviation (σ_r) of $\epsilon_{i,r}$ across rounds enables to us to reflect hypotheses H2 (preference learning) and H3 (fatigue). Preference learning would lead to a decrease in the standard deviation σ_r across rounds ($\sigma_1 > \sigma_2 > \sigma_3$), while fatigue we lead to the opposite effect ($\sigma_1 < \sigma_2 < \sigma_3$). Note that, as our main variable of interest are changes in response behaviour ($\Delta\, y_{i,r}$), we do not explicitly consider how any personal characteristics might influence the absolute responses of participants.

4.2. Data Generating Process

In order to define a full Data Generating Process (DGP) based on our previously defined theoretical model, we need to make further assumptions about functional form relationships and prior distributions. They are given by:

$$\Delta y_{i,r} \sim Truncated\ normal\left(\widehat{\Delta y_{i,r}}, \sigma_r, low = -1, high = 1\right)$$

$$\widehat{\Delta y_{i,r}} = \theta_{r,tech} + \beta_{tech} * treat$$

$$\theta_{r,tech} \sim Normal(0, 0.3)$$

$$\beta_{tech} \sim Normal(0, 0.3)$$

$$\sigma_r \sim exponential(1)$$

Considering that $\Delta y_{i,r}$ gives changes in shares, which naturally are bounded between -1 and 1, we assume that $\Delta y_{i,r}$ is a truncated normal distribution with limited values between -1 and 1, a mean of 0 and a standard deviation σ_r. As the prior distribution for this round-wise standard deviation σ_r, we choose the exponential distribution around 1. We further assume a normal distribution with mean zero and standard deviation 0.3 for the round and treatment effect coefficients. The dimensions of $\Delta y_{i,r}$ are a 2x1 vector with one value between -1 and 1 for each row (technology). Accordingly, the dimensions of the round effect coefficient $\theta_{r,tech}$ are a 2x3 matrix with each technology in one row and each round played in one column. We only estimate two of the three output variables within the model, since the third one can be derived by summing to zero: As the participants can and must distribute exactly 50 hectares every round, their relative changes in the hectare distribution per technology have to sum up to zero for each round.

A key advantage of the PP approach is that it is sufficient to define the DGP in a PP library. The empirical inference model is then automatically derived. This ensures that the empirical model is consistent with the assumed DGP process (see Storm et al, in progress). It is then possible to condition the DGP on observed data to perform Bayesian inference, i.e., updating the prior belief based on the information in the data. The inference step of approximating the posterior distribution is conducted using Markov-Chain-Monte-Carlo (MCMC) sampling. Specifically, we use the "No U-turn sampler" (NUTS), a variation of a Hamiltonian Monte Carlo sampler that requires minimal tuning. The entire analysis was implemented in the Python PP library NumPyro. Details of the implementation are provided in the supplementary code available online, sources for this are included in the appendix.

4.3. Statistical implication of the research hypotheses

In order to link our research hypotheses with the statistical model, we now express each hypothesis in model formulation.

H1 (Institutional learning): Participants show a change in response behaviour when moving from the baseline to the first round.

- The absolute value of the round effect parameter is non-zero when moving from round zero to round one: $\theta_{0,tech} \neq 0$ for all technologies.

H2 (Preference learning): Participants show a reduction of behaviour change variance when moving from the first to the second round.

- The variance of the changes in behaviour is smaller for round 2 than for round 1: $\sigma_1 > \sigma_2$

H3 (Fatigue): Participants show an increase of behaviour change variance when moving from the second to the third round.

- The variance of the changes in behaviour is larger for round 3 than for round 2: $\sigma_3 > \sigma_2$

4.4. Model Inspection

An important advantage compared to classical econometrics is that PP puts a stronger focus on a clearly defined DGP and that this DGP can also be used to generate data (see Storm et al, in progress). This synthetic (dummy) data can then be inspected and used to test the functionality of the code and the inference model. This step can serve as an important quality assurance and to build trust in the modelling process. Additionally, one can condition this data generation process on specific parameter values, such that the synthetic data follows known properties. In our setting, we hence generate synthetic data assuming that our three hypotheses are individually or jointly present. This is ensured by defining sets of coefficient values that reflect the hypotheses and conditioning the DPG on the respective coefficient values. We thus test the inference step of our model: If we are able to recover the initially defined coefficient values by running the sampler, we know that the model and sampler function correctly. This also allows us to prepare the illustration of model results in a synthetic setting before moving to the actual data. Moreover, this process enables us to assess how our hypotheses about round effects affect the estimation of treatment effects in a situation where we know the "true" treatment effect present in the generated data. This delivers important insights into the potential biases which could be introduced by round effects.

After having ascertained the general functionality of our model in the present case, we generated four dummy data sets, which we conditioned on the following assumptions.

First, we assumed that *H1 is true,* which entailed the conditioning of the following parameter values:

$$\sigma \;=\; \begin{pmatrix}0.1 & 0.1 & 0.1\end{pmatrix}, \qquad \theta = \begin{bmatrix}-0.1 & 0 & 0\\ -0.1 & 0 & 0\end{bmatrix}, \qquad \beta = \begin{pmatrix}0.2\\ 0.2\end{pmatrix}$$

Where each row indicates one of the two modelled weeding technologies (weeding robot and spot spraying), and each column indicates one of the three $\Delta y_{i,r}$ for the θ matrix. Visualized, the distribution of the allocation changes for weeding robot and broadcast application resulting from this dummy dataset is given in Figure 1. Inspecting the descriptive visualisation of this dummy dataset gives an idea of how the mass in the density functions is pushed to the right when moving from the baseline to round one, which captures the “switching” behaviour introduced by the non-zero round effect parameter. This reflects institutional learning. In order to test the functionality of our sampler, we ran this dummy dataset with our NUTS sampler. The obtained posterior density distributions are also given in Figure 1.

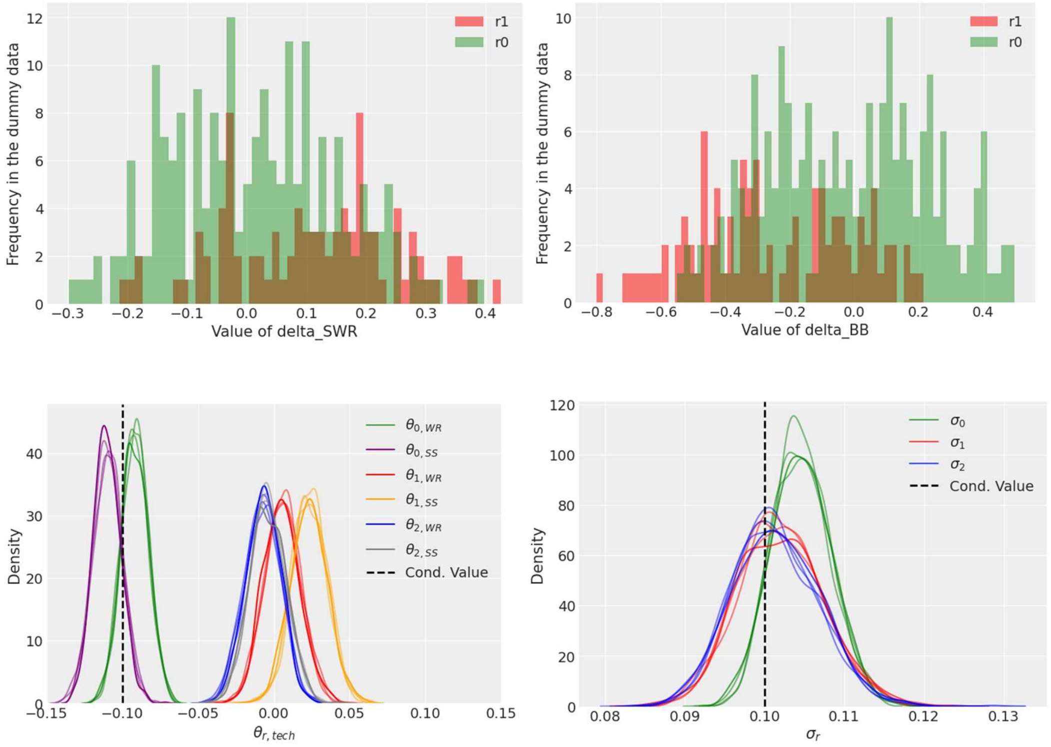

Fig. 1: First row: Distributions of changes in hectares allocated to the weeding robot and broadcast application when moving from round 0 to round 1 in the dummy data generated by prior sampling conditioned on fulfilling H1. Second row: Posterior Distribution obtained by sampling values for the round effect parameter and σ_r using the dataset conditioned on H1, per chain.

Second, we conditioned the model reflecting that *H2 is true*, which entailed the following parameter values:

$$\sigma \ = \ (0.5 \quad 0.2 \quad 0.2\,), \quad \theta = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \beta = \begin{pmatrix} 0.2 \\ 0.2 \end{pmatrix}$$

Here, we compare the changes of hectare shares distributed to weeding robot over the course of the rounds and we can clearly observe a reduction of the variance in round one compared to round zero, as depicted in Figure 2. Running this dataset with the sampler, we obtain posterior distributions for the variance term that closely resemble our conditioned parameters, as also shown in Figure 2.

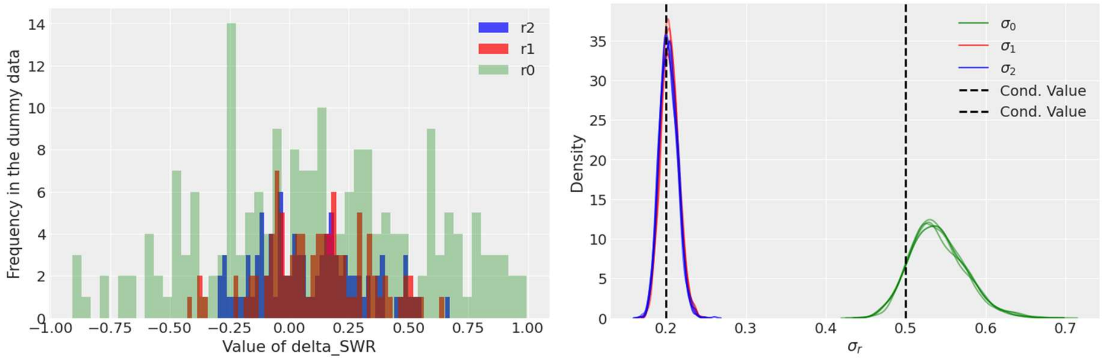

Fig. 2: Left: Distributions of changes in hectares allocated to the weeding robot for each game round in the dummy data generated by prior sampling conditioned on fulfilling H2. Right: Posterior Distribution of σ_r obtained by NUTS sampling using the dataset conditioned on H2, per chain.

Next, we conditioned that *H3 is true*, which entailed the following parameter values:

$$\sigma \ = (0.2 \quad 0.2 \quad 0.7\,), \qquad \theta = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \beta = \begin{pmatrix} 0.2 \\ 0.2 \end{pmatrix}$$

As above, we observe the distributions of response behaviour changes across rounds to check whether the conditioning functioned correctly. As the variance is more widely spread for the third round than for the second, we can confirm the generative model's functionality here as well.

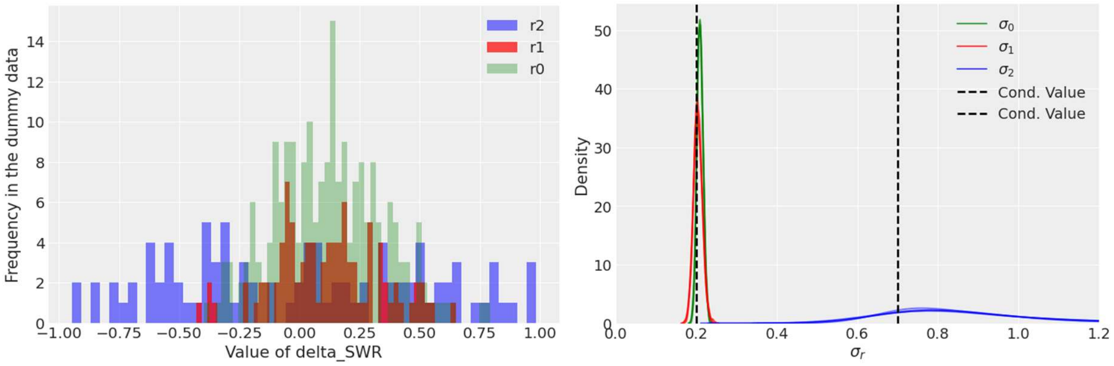

Fig. 3: Left: Distributions of changes in hectares allocated to the weeding robot for each game round in the dummy data generated by prior sampling conditioned on fulfilling H3. Right: Posterior Distribution of σ_r obtained by NUTS sampling using the dataset conditioned on H3, per chain.

Next, we assume that *H1, H2 and H3 are all true,* which could be reflected by the below parameter values:

$$\sigma \;=\; \begin{pmatrix} 0.5 & 0.2 & 0.7 \end{pmatrix}, \qquad \theta = \begin{bmatrix} -0.1 & 0 & 0 \\ -0.1 & 0 & 0 \end{bmatrix}, \qquad \beta = \begin{pmatrix} 0.2 \\ 0.2 \end{pmatrix}$$

In case all three hypotheses were to be confirmed in the real data set, the distributions of the output variables would have to look similar to the plots presented in Figure 3 below. The obtained parameter values when using our NUTS sampler are depicted using the posterior distributions as also visualised in Figure 3 below.

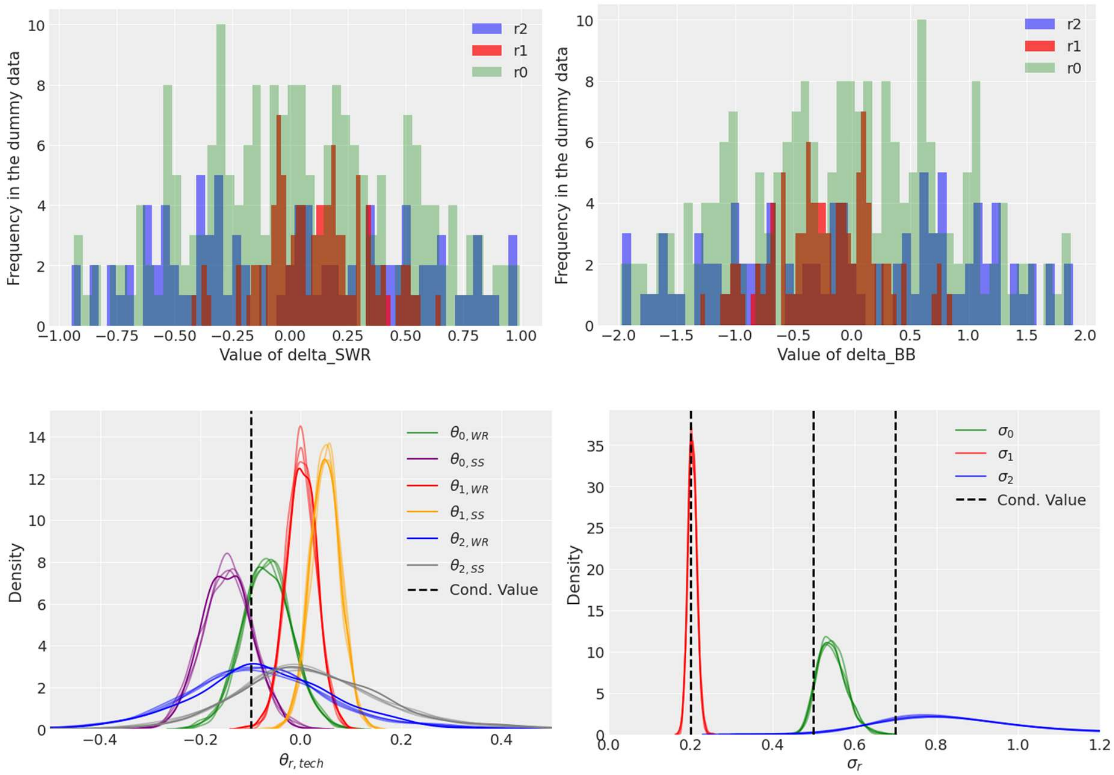

Fig. 3: First row: Distributions of changes in hectares allocated to the weeding robot and broadcast application for each game round in the dummy data generated by prior sampling conditioned on fulfilling all three hypotheses. Second row: Posterior Distribution of the $roundeffect$ parameter and σ_r obtained by NUTS sampling using the dataset conditioned on fulfilling all three hypotheses, per chain.

For all cases shown above, our MCMC parameter estimations were sufficiently close to the original conditioned values thereby confirming the functionality of our model.

Lastly, we assumed that *round effect modelling is relevant for treatment effect quantification.* One advantage of generating dummy data conditioned on a fixed set of parameter values is that it enables us to test alternative model specifications. Here, we aim to explore how round effects can potentially impact the estimation of treatment effects if they are not considered in the model. For this, we consider a model that does not explicitly capture changes in response behaviour or developments in the variance of this response behaviour. Specifically, we consider a model given by:

$$\Delta y_{i,r} \sim TruncatedNormal\left(\underline{\Delta y_{i,r}}, \sigma, low = -1, high = 1\right)$$

$$\underline{\Delta y_{i,r}} = \beta_{tech} * treat$$

$$\beta_{tech} \sim Normal(0, 0.3)$$

$$\sigma \sim exponential(1)$$

Using this alternative model specification, we investigate how ignoring round effects can impact the estimation of treatment effects. We use the model to estimate treatment effects using data generated with the model described in section 4.2 conditioned on the coefficient values of the ‘all hypotheses true’ specification above.

The Model runs without problems, but the resulting posterior distributions now differ strongly from the true treatment effect values that we used for generating the data, as is depicted in Figure 4.

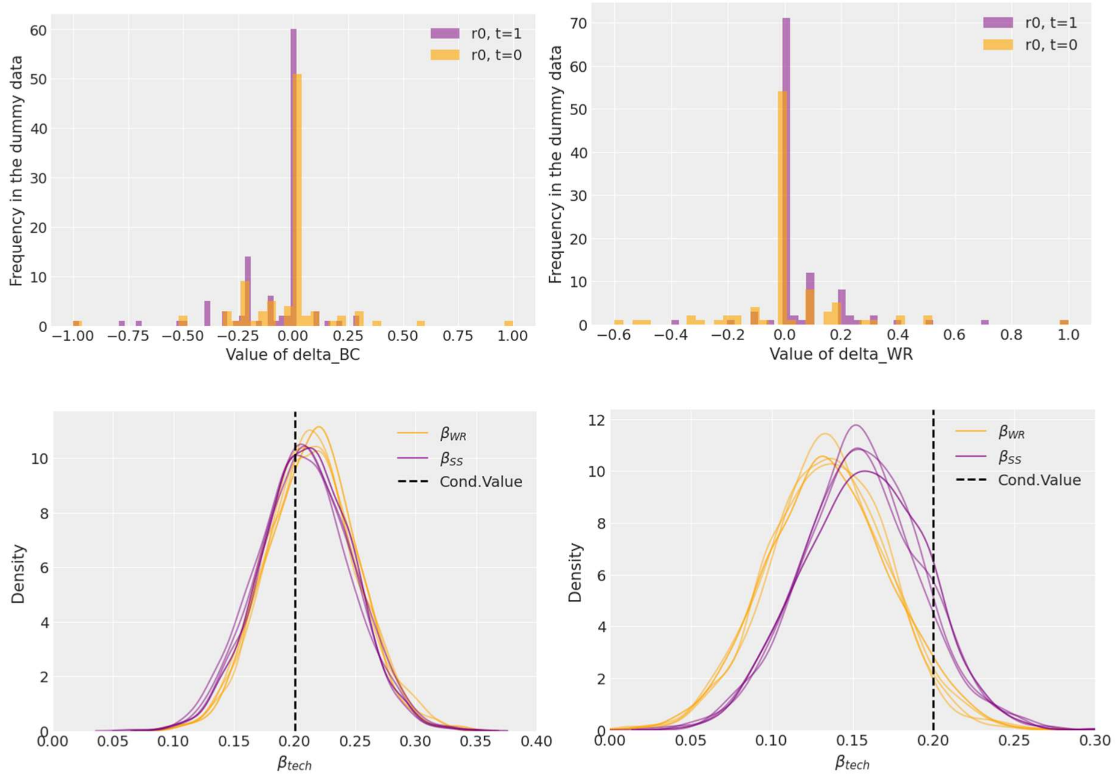

Fig. 4: First row: Distributions of changes in hectares allocated to the weeding robot and broadcast application when moving from r0 to r1, separated per treatment group, in the dummy data generated by prior sampling conditioned on fulfilling all three hypotheses. Second row: Left: Posterior Distribution of the β_{tech}parameter obtained by NUTS sampling using the dataset conditioned on fulfilling all three hypotheses and the original model including round effects, per chain. Right: Posterior Distribution of the β_{tech} parameter obtained by NUTS sampling using the dataset conditioned on fulfilling all three hypotheses and the simplified model excluding any $\theta_{r,tech}$ estimation, per chain.

This confirms our assumption that not controlling for round-specific behavioural effects will bias the correct quantification of treatment effects. In the above case where we assume the presence of all three treatment effects, including no round effect parameter or a round-specific sigma term will result in the under-estimation of the effect of the policy treatment on hectares assigned to sustainable weeding technologies weeding robot and spot spraying.

5. Results and discussion

After thoroughly checking and validating of our model as described in the previous sections, we ran the model using an MCMC algorithm. We chose initialization to the median and a number of 1000 samples for warmup and sampling each, running four chains simultaneously. By inspecting the corresponding traceplots, we could confirm that our chains run 'healthily'. A detailed output summary table, as well as the mentioned traceplot, are included in the appendix. The following paragraphs will offer insights into our sample and more detailed information and interpretations for each parameter estimate, contextualised with the research hypotheses as defined in the previous sections.

5.1. Descriptive statistics

Table 3 below displays the absolute values of weeding decisions in all groups across the game rounds.[3] In the baseline, a clear preference of broadcast application over spot spraying over the weeding robot becomes clear across all groups. In going from the baseline to round one, all groups display similar allocations changes, i.e., they reduced the number of hectares allocated to broadcast application while the number of hectares allocated to spot spraying and the weeding robot increased. However, the allocation changes from broadcast application towards both SFT seem to be pronounced in groups two and four which received a hypothetical subsidy for each hectare dedicated to any SFT. In rounds two and three which was only played by one treatment and control group, respectively, the negative trend for broadcast application continued; however, the control group reduced more intensively, although admittedly being on a higher level to begin with. This pattern is inversely matched by the hectare allocation to spot spraying. Specifically, the control group allocated more hectares to spot spraying with larger increases from round to round compared to group four. Regarding the weeding robot, the control group reverted to allocating less hectares in rounds two and three while the treatment group allocated less (more) hectares to the weeding robot in round two (three). In summary, between the baseline and round one, both control groups display a more moderate change from broadcast application towards SFT, while the opposite is the case between rounds one and two, and between two and three, respectively, in which the remaining treated group displays more stable allocation behaviour.[4]

Table 3. Mean weeding decision by group, technology and game round.

Technology	Group[a]	*Round 0*	*Round 1*	*Round 2*	*Round 3*
		Mean (SD)			
Broadcast application	1	23,8 (19,1)	23,0 (20,0)	-	-
	2	22,1 (20,2)	17,7 (21,0)	-	-
	3	26,2 (18,7)	24,4 (19,4)	21,8 (20,9)	19,2 (20,9)
	4	23,5 (20,8)	19,1 (21,2)	18,8 (21,9)	17,0 (21,9)

[3] Test of successful randomization of treatment groups can be found in the Table S1 in the Supplementary Information.

[4] Average changes of weeding decisions per round and treatment group can be found in Table S3 in the Supplementary information.

Spot spraying	1	16,0 (15,3)	16,3 (15,4)	-	-
	2	20,1 (17,0)	22,1 (17,9)	-	-
	3	12,6 (14,3)	13,1 (13,5)	16,8 (16,3)	20,8 (19,6)
	4	16,9 (17,9)	18,2 (19,0)	19,6 (20,4)	20,1 (20,2)
Weeding robot	1	10,2 (12,6)	10,7 (13,9)	-	-
	2	7,8 (11,4)	10,2 (12,9)	-	-
	3	11,2 (14,4)	12,5 (14,4)	11,4 (13,8)	10,0 (14,5)
	4	9,6 (15,8)	12,7 (18,0)	11,7 (16,7)	12,8 (17,4)

[a] Group 1 – control, two rounds (n=51); Group 2 – subsidy, two rounds (n=57); Group 3 – control, four rounds (n=43); Group 4 – subsidy, four rounds (n=52)

Overall, the data shows a much higher concentration around zero than assumed by the priors in the generative model above. The distributions of the changes in hectare shares allocated to the weeding robot are visualized in figure 5.

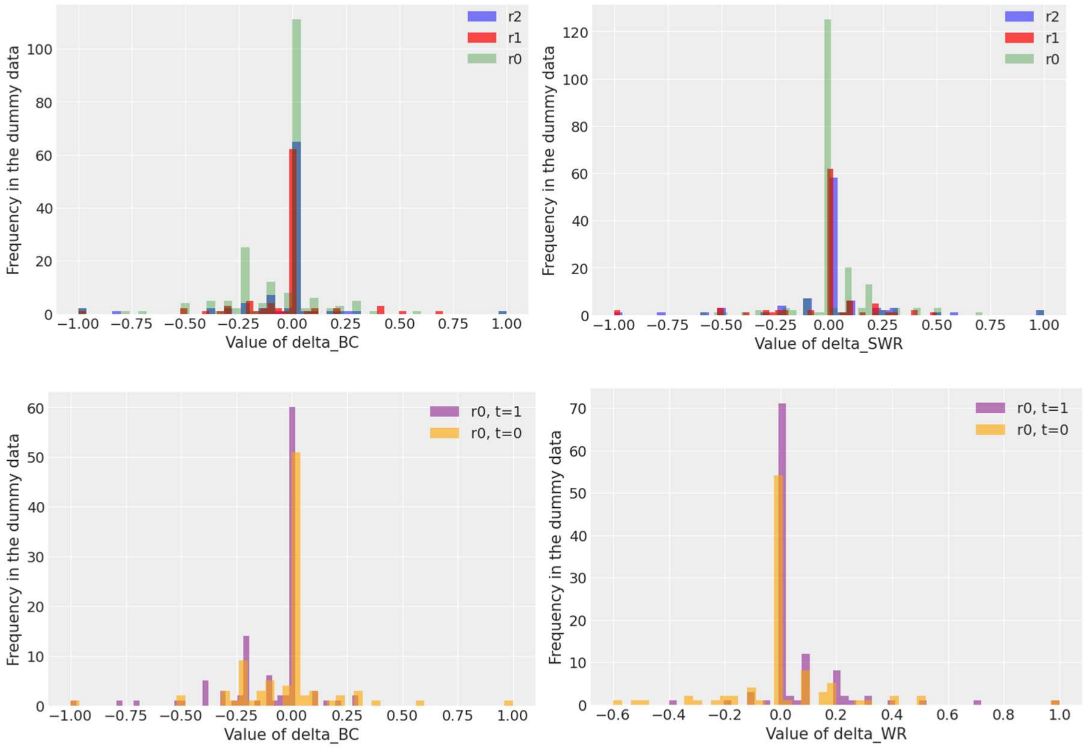

Fig. 5: First row: distributions of changes in hectares allocated to the weeding robot and broadcast application for each game round in the survey. Second row: Real data distributions of changes in hectares allocated to weeding robot and broadcast application when moving from r0 to r1, separated per treatment group.

5.2. Institutional learning (H1)

Institutional learning is defined as a switch of behaviour between the first and second round of an experiment, which implements in our study as a non-zero value of the round effect coefficient when moving from round zero to round 1. Hence, we will need to consider the parameter values for all technologies for moving from round zero to round 1 (r0).

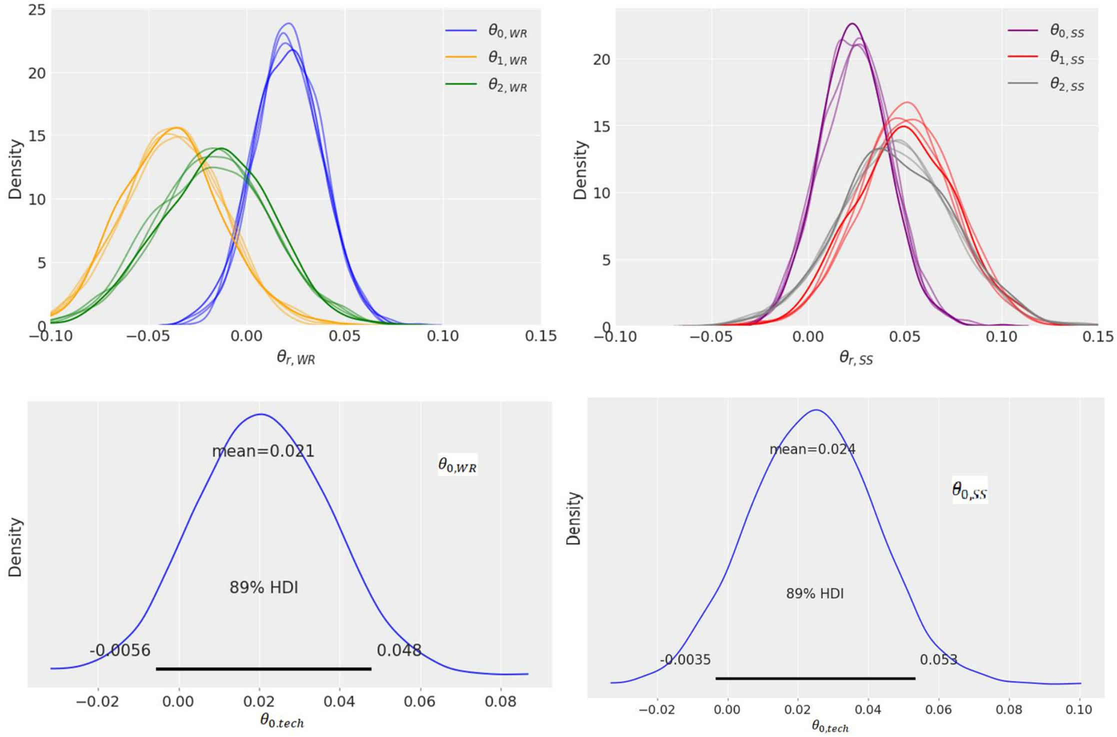

Fig. 7: First row: Posterior plots for the distribution of the round effect parameter of changes in relative hectare allocations to the smart weeding robot (left) and spot spraying (right) per round, second row: Posterior plots for the distribution of the round effect parameter of changes in relative hectare allocations to the smart weeding robot (left) and spot spraying (right), including a visual representation of the 89% highest density interval, for moving from round zero to round one.

As we find that the magnitude of changes in participants behaviour actually is smallest when moving from round zero to round 1 (figure 7), we conclude that, with our data and the model assumption, we do not find support for the hypotheses of institutional learning at the very beginning of the experiment. Considering the characteristics of the sample, one could reason that the predominantly young and educated majority of students among the respondents do not struggle with understanding the setting and framework of an online experiment. Here, a repetition of the same experiment and analysis with an older, less digitally native sample would be interesting to see whether the customisation to online experiments would make a difference for the effect of institutional learning.

5.3. Preference learning (H2) and Fatigue (H3)

As shown in figure 8, we find that the variance of participants changes in response behaviour, σ_r, progressively increases from round to round. This indicates that, considering our data and model assumption, we do not find evidence for an preference learning effect (H2). Instead, results are well in line with the assumption of fatigue (H3).   .

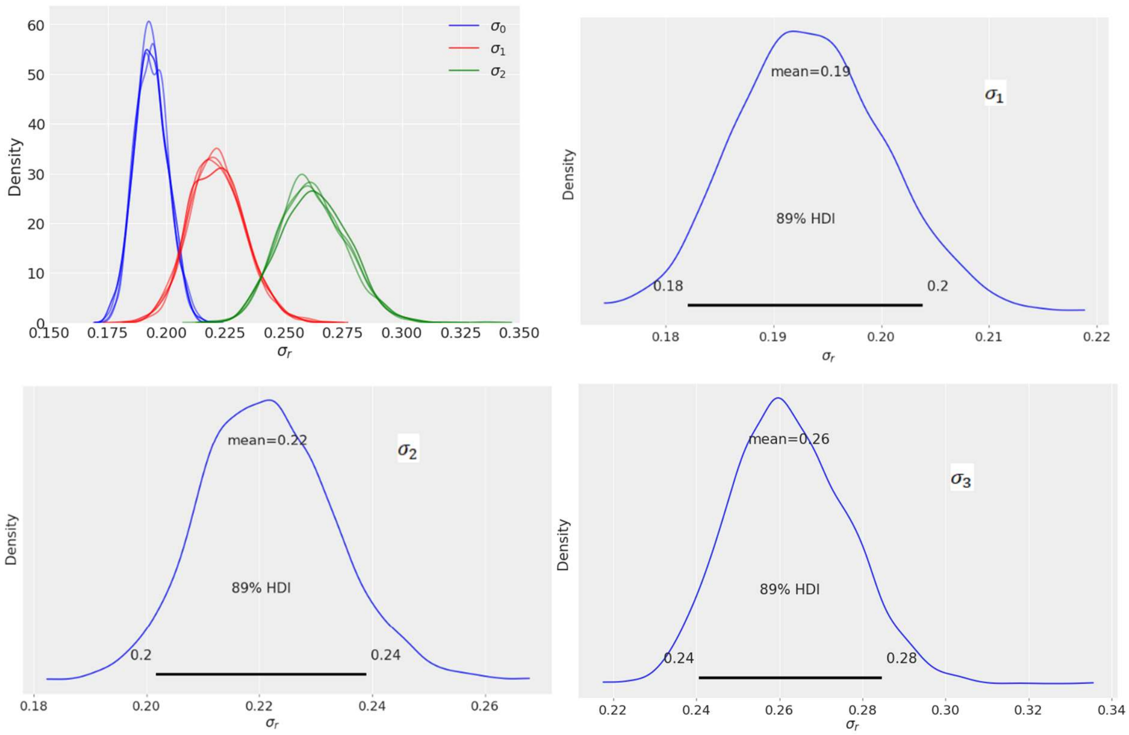

Fig. 9: Posterior plots for the distribution of the σ_r parameter of the variance in participants response behaviour, first plot: for all rounds, per chain, other plots: Per round, including a visual representation of the 89% highest density interval.

This lack of preference learning can be interpreted as a confirmation of strong preferences towards weeding technologies in the participant group, which they also seem to be aware and consistent about. Considering our sample, this interpretation can be supported by the high share of students from the agricultural field and by the fact that half of the respondents have indicated to have prior experience with smart farming technologies. Additionally, when asked about their attitude towards the environment, students on average indicated high environmental awareness. In combination with high trust in data security and high personal innovativeness, this seems to lead to strong preferences towards the weeding management and hence reduce the room for preference learning.

Although the parametric evidence we find shows a steady increase in respondents answering ‘randomness’, our participants did not indicate to feel tired. When we asked them about feeling a loss of concentration, a feeling of being overwhelmed by the information, or becoming increasingly annoyed, they expressed low overall agreement to these statements. This suggests the conclusion that participants were either getting more tired without being aware of it or that the increased variance in their response behaviour had a different cause. Another effect potentially causing such increases in random behaviour is the failing credibility theory (Carson & Groves, 2007), where respondents become wary of the experimental setup and choose to not respond with much consideration anymore. Further research including variations in the experimental framework and explicitly asking respondents about both their perceived fatigue and perception of credibility of the experiment could shed more insights into the distinction between causes for increasing response variance. Another aspect that could be important here is the online format of the experiment. In a study by Savage and Waldman (2008), online respondents were found to suffer from fatigue much more than mail respondents of the same survey. Here, the results from the mail respondents showed much higher data quality because the choice consistency was maintained over the course of rounds.

In light of the high evidence of fatigue effects and low evidence of the preference learning hypothesis, our results support the assumption by DeSarbo et al. (2004) that highly involved respondents might give their most valuable responses early on in the study before the onset of potential fatigue or failing credibility. This conclusion contradicts the finding of Carlsson et al (2012), which proposed to disregard the first round(s) of a choice experiment to give respondents room for institutional and preference learning, and strongly underlines the importance of round effect quantification and sample characterisation. Repeating the study with less involved or respondents less seasoned regarding online experiments might offer additional backing of this interpretation and help forming conclusions for future applications.

6. Conclusion

Round effects matter – in order to accurately identify potential treatment effects, to discern which rounds are most valuable for abstraction of policy implications, and overall to better understand experiment respondents' behaviour. In our study, we find strong evidence of fatigue effects and were able to falsify the assumptions of institutional and preference learning, which we ascribe to the high awareness about environmental topics, high knowledge around smart farming technology and assumed prior experience with online experiments in our sample. From this connection we deduct, that future studies should carefully investigate the round effects at play in their experiments and consider the characteristics of their sample in this analysis. Specifically, respondents' involvement with the topic at hand and their previous experience with the experimental setting should receive more attention.

From a policy perspective, the inclusion of quantification and controlling for round effects will be vital in any policy-informing experimental research settings, because the neglecting of round effects will lead to wrong estimation of treatment effects and misinterpretation of study outputs.

Further research should repeat multi-period experimental designs with several types of respondents as well as varying the designs and experimental procedures to gain more insights into how exactly those factors influence the degree to which round effects manifest. Another extension to future experiment designs would be to track and compare the response time per round in order to investigate fatigue preference learning through a different lens.

By using our novel Bayesian approach implemented via probabilistic programming, we were able to model round-specific changes in the response behaviour, as well as the development of the variance of behaviour changes over multiple rounds. These quantifications would not have been possible with most classic econometric approaches and yield much room for further procedural extensions and adjustments in other settings. One important advantage of the Bayesian approach is the possibility of extensive model testing and refinement before inspection of real data as well as the ability to obtain complete posterior densities instead of point estimates. Future applications of our model could include multi-period experiments in all fields of research and we are encouraging potential extensions or alternatives to our idea in other settings.

Funding

This work is partly funded by the German Research Foundation (DFG) under Germany's Excellence Strategy, EXC-2070 – 390732324 – PhenoRob. It is also partly funded by DFG under grant number 504991777

Supplementary information

The Supplementary information to this article can be found following this OSF link:
https://osf.io/mn6gc/?view_only=0c74af1114254f5ab8ea3c632cf9bdcc

Data availability

Data to this article can be found online following this OSF link:
https://osf.io/mn6gc/?view_only=0c74af1114254f5ab8ea3c632cf9bdcc

Compliance with ethical standards

This study involved human participants. Prior to data collection, ethical clearance was granted by the ethical board of the Centre for Development Research (ZEF) at University of Bonn. The approved ethical clearance form is available upon request with the corresponding author.

CRediT author statement

Alexa Leyens: Conceptualization, Methodology, Formal analysis, Resources, Investigation, Writing – original draft, Writing – review and editing, Visualization. **Philipp Feisthauer**: Conceptualization, Investigation, Resources, Data curation, Writing – original draft, Writing – review and editing, Visualization, Project administration. **Monika Hartmann**: Conceptualization. **Jan Börner**: Conceptualization, Supervision. **Hugo Storm**: Formal analysis, Writing – review and editing, Supervision.

References

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106. https://doi.org/10.1162/00335530360535153

Blasch, J., van der Kroon, B., van Beukering, P., Munster, R., Fabiani, S., Nino, P., & Vanino, S. (2022). Farmer preferences for adopting precision farming technologies: a case study from Italy. *European Review of Agricultural Economics*, *49*(1), 33–81. https://doi.org/10.1093/erae/jbaa031

Braga, J., & Starmer, C. (2005). Preference Anomalies, Preference Elicitation and the Discovered Preference Hypothesis. *Environmental and Resource Economics*, *32*(1), 55–89. https://doi.org/10.1007/s10640-005-6028-0

Brouwer, R., Dekker, T., Rolfe, J., & Windle, J. (2010). Choice Certainty and Consistency in Repeated Choice Experiments. *Environmental and Resource Economics*, *46*(1), 93–109. https://doi.org/10.1007/s10640-009-9337-x

Brown, T. C., Kingsley, D., Peterson, G. L., Flores, N. E., Clarke, A., & Birjulin, A. (2008). Reliability of individual valuations of public and private goods: Choice consistency, response time, and preference refinement. *Journal of Public Economics*, *92*(7), 1595–1606. https://doi.org/10.1016/j.jpubeco.2008.01.004

Carlsson, F., & Martinsson, P. (2001). Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments? *Journal of Environmental Economics and Management*, *41*(2), 179–192. https://doi.org/10.1006/jeem.2000.1138

Carlsson, F., Mørkbak, M. R., & Olsen, S. B. (2012). The first time is the hardest: A test of ordering effects in choice experiments. *Journal of Choice Modelling*, *5*(2), 19–37. https://doi.org/10.1016/S1755-5345(13)70051-4

Carson, R. T., & Groves, T. (2007). Incentive and informational properties of preference questions. *Environmental and Resource Economics*, *37*(1), 181–210. https://doi.org/10.1007/s10640-007-9124-5

Chou, E., McConnell, M., Nagel, R., & Plott, C. R. (2009). The control of game form recognition in experiments: understanding dominant strategy failures in a simple two person "guessing" game. *Experimental Economics*, *12*(2), 159–179. https://doi.org/10.1007/s10683-008-9206-4

Day, B., Bateman, I. J., Carson, R. T., Dupont, D., Louviere, J. J., Morimoto, S., Scarpa, R., & Wang, P. (2012). Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of Environmental Economics and Management*, *63*(1), 73–91. https://doi.org/10.1016/j.jeem.2011.09.001

DeSarbo, W. S., Lehmann, D. R., & Hollman, F. G. (2004). Modeling Dynamic Effects in Repeated-Measures Experiments Involving Preference/Choice: An Illustration Involving Stated Preference Analysis. *Applied Psychological Measurement*, *28*(3), 186–209. https://doi.org/10.1177/0146621604264150

DeShazo, J. R., & Fermo, G. (2002). Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. *Journal of Environmental Economics and Management*, *44*(1), 123–143. https://doi.org/10.1006/jeem.2001.1199

Feisthauer, P., Hartmann, M., & Börner, J. (2024). Adoption intentions of smart weeding technologies—A lab-in-the-field experiment with German crop farmers. *Q Open*, *4*(1), Article qoae002. https://doi.org/10.1093/qopen/qoae002

Fleming, P. M., Palm-Forster, L. H., & Kelley, L. E. (2021). The effect of legacy pollution information on landowner investments in water quality: lessons from economic experiments in the field and the lab. *Environmental Research Letters*, *16*(4), 45006. https://doi.org/10.1088/1748-9326/abea33

Ladenburg, J., & Olsen, S. B. (2008). Gender-specific starting point bias in choice experiments: Evidence from an empirical study. *Journal of Environmental Economics and Management*, *56*(3), 275–285. https://doi.org/10.1016/j.jeem.2008.01.004

Musshoff, O., & Hirschauer, N. (2014). Using business simulation games in regulatory impact analysis – the case of policies aimed at reducing nitrogen leaching. *Applied Economics*, *46*(25), 3049–3060. https://doi.org/10.1080/00036846.2014.920482

Reed Johnson, M., & Bingham, M. F. (2001). Evaluating the validity of stated-preference estimates of health values. *Swiss Journal of Economics and Statistics (SJES)*, *137*(I), 49–63.

Savage, S. J., & Waldman, D. M. (2008). Learning and fatigue during choice experiments: a comparison of online and mail survey modes. *Journal of Applied Econometrics*, *23*(3), 351–371. https://doi.org/10.1002/jae.984

Swait, J., & Adamowicz, W. (2001). The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching. *Journal of Consumer Research*, *28*(1), 135–148. https://doi.org/10.1086/321952

Thomas, F., Midler, E., Lefebvre, M., & Engel, S. (2019). Greening the common agricultural policy: a behavioural perspective and lab-in-the-field experiment in Germany. *European Review of Agricultural Economics*, *46*(3), 367–392. https://doi.org/10.1093/erae/jbz014

Thompson, N. M., Bir, C., Widmar, D. A., & Mintert, J. R. (2019). Farmer perception of precision agriculture technology benefits. *Journal of Agricultural and Applied Economics*, *51*(1), 142–163. https://doi.org/10.1017/aae.2018.27