Using Factorial Survey Experiments to Measure Attitudes, Social Norms, and Fairness Concerns in Developing Countries

Ulf Liebe[1,2], Ismaïl M. Moumouni[3], Christine Bigler[4,5], Chantal Ingabire[5], Sabin Bieri[2]

[1] Institute of Sociology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland. Email: ulf.liebe@soz.unibe.ch (corresponding author)

[2] Centre for Development and Environment (CDE), University of Bern, Hallerstrasse 10, 3012, Bern, Switzerland

[3] Département d'Economie et de Sociologie Rurales, Faculté d'Agronomie, Université de Parakou, BP: 123 Parakou, Bénin

[4] Interdisciplinary Centre for Gender Studies, University of Bern, Vereinsweg, 3012 Bern, Switzerland

[5] International Centre for Tropical Agriculture, Kigali, Rwanda

Abstract: Survey-based experimental methods are increasingly used in the social sciences to study, among others, attitudes, norms, and fairness judgments. One of these methods is the factorial survey experiment (FSE, or vignette experiment) in which respondents are confronted with various descriptions of situations that differ in a discrete number of attributes (or factors), and they are asked to evaluate those situations according to criteria such as agreement, approval, and fairness. Due to the systematic, experimental variation of the presented situations, an FSE can separate effects of single situational attributes, allowing the causal influence of relevant situational attributes to be determined. This is the key advantage over simple survey items. While most studies using FSEs are carried out in developed countries in which respondents are familiar with surveys, we add further evidence that this method can unfold its power also in a developing context. Building on previous applications of FSEs in Africa, we demonstrate the usefulness of this method in four novel studies on social norms regarding the physical punishment of children and the social approval of technology adoption in Benin as well as judgments of just earnings in Rwanda. We also test for the first time the applicability of multiple vignettes per respondents in a global South remote area context. The results of these studies are theoretically meaningful and the overwhelming majority of respondents discriminate between vignettes. This supports the validity of FSEs. However, conducting survey experiments in developing countries is different from similar experimental research in developed countries and, therefore, we also discuss some of these differences and corresponding challenges. Last but not least, our paper shows, provided a few precautions are heeded, that FSEs could be used as a vehicle to innovate social science research in a global South/remote area context.

Keywords: Attitudes, Experiment, Factorial Survey Experiment, Fairness, Social Norms

Introduction

Survey research in developed and developing countries has to deal with different forms of bias including socially desirable response behavior (e.g., Johnson and van de Vijver 2003; Kreuter et al. 2008; He et al. 2014). Respondents might (untruthfully) answer survey questions in line with social norms, political rules, and in a way to please the researchers. This problem is even more severe if the research question targets sensitive issues (Krumpal 2013). Further, while in developed countries the concept of a scientific interview and the different survey modes are mostly familiar to potential respondents, this is not always the case in a developing-country context (de Leeuw 2008). If respondents do not know how to behave in an interview situation, they might be generally more prone to agree with survey questions leading to a so-called acquaintance bias (Schaeffer and Presser 2003).

Against this background, and in response to the problem of detecting causal effects in survey research, researchers have started to use experimental methods within surveys (Mutz 2011). An experiment comprises at least two (experimental) groups, whereby each respondent is randomly assigned to one of these groups (Shadish et al. 2002). The questions and experimental tasks that these groups face vary by one factor. By looking at the difference in response behavior in both groups (treatment and control), researchers are able find out which effect this single factor has on the respondents' answers. This outcome – being able to single out the effect of one factor – is the major advantage of experiments compared with standard measurement instruments in survey research. Applications of survey-based experimental research in a developing-country context, more specifically in Africa, include studies on the value of cattle breeds in southern Ethiopia and northern Kenya (Zander and Drucker 2008); household definitions in Mali (Beaman and Dillon 2012); and ethnic voting in Uganda (Conroy-Krutz 2013).

The factorial survey experiment (FSE, also vignette experiment), which we employ in the present paper, is a multi-factorial survey method that is often applied in sociological research (Wallander 2009; Auspurg and Hinz 2015). The method was introduced by Rossi and Lazarsfeld in the 1950s (Rossi 1979) and since the 1970s it has become an important method for the study of justice concerns and social norms, among others (see Jasso and Rossi 1977; Jasso and Opp 1997). In FSEs respondents face one or more descriptions of a situation that differ in a discrete number of attributes (or factors). The respondents are then asked to evaluate each situation according to criteria such as support, agreement, and perceived fairness. Due to the systematic variation of the factors or situational attributes presented in the situation, an FSE is an experimental setup which can separate effects of single situational dimensions. Thus the

causal influence of relevant situational attributes can be determined. Further, FSEs measure beliefs, social norms, and judgments in an elegant way, because they do not measure the concepts directly via single survey items but indirectly, based on the relevance of corresponding situational variables. In multivariate regression analyses the evaluations are included as dependent variables and the factors/situational attributes as independent variables (e.g., Jasso 2006).

As in all empirical research, conducting an FSE includes several steps in which researchers have to make certain decisions (see Auspurg and Hinz 2015 for details and guidelines). First, researchers have to choose the number of attributes or characteristics of a situation they want to vary. These attributes should be relevant for the respondents and they can be selected using focus groups, for instance. Combining all possible attribute combinations gives the so-called "full factorial", the number of possible situations respondents can judge. If a FSE comprises many attributes, this number is often too large to present to all respondents. Therefore, second, if the full factorial is very large, an experimental design is used to reduce the number of vignettes that respondents face. At the same time, it should still be possible to separate the effects of single factors. Third, researchers have to decide which response scale they want to use to record respondents' judgments (e.g., five-point, seven-point, eleven-point response scales), and, fourth, there are different statistical models that can be used to analyze FSE data.

The aim of this paper is to test, encourage and discuss the application of FSEs in a developing-country context. We present four studies from Africa, two from Benin and two from Rwanda, which investigated novel topics: social norms regarding the physical punishment of children, the social acceptance of technology adoption, and judgments on just earnings. These studies deviate from the complex experiments that are typically applied in Europe and North America. To test and ensure the applicability of the method, and considering the remote area context in which the experiments were conducted, we use more "simple" experiments with a low number of factors. It has to be noted that we are not the first ones to use FSEs in Africa. Table 1 gives an overview of previous applications, which include topics such as fairness judgments regarding amnesty; formal and informal property rights norms; and the regulation of female behavior. Some observations from Table 1 are notable. First, all studies, as our own studies presented later, use the full factorial with a minimum of five and a maximum of sixteen vignettes. Second, every respondent answered one vignette that was randomly chosen from the full factorial. Third, there is a range of response scales varying from four-point scales to ten-point scales.

Table 1: Overview of Factorial Survey Experiments in African Countries

Study	Country	Topic	Design	Respondents and Scale
Gibson 2002	South Africa	Judgments of fairness to victims regarding amnesty for a bomb attack	2x2x2x2=16 vignettes; attributes: victims' family got voice or no voice (procedural justice); offender was punished or not punished (retributive justice); victims' family received apology or not; victims' family got compensation or not	n=3,727; each respondent randomly assigned to one vignette; 10-point response scale
Duch and Palmer 2004	Benin	Property rights norms, the acceptance of property expropriation	2x2x2=8 vignettes, attributes: owner with valid title to land or not; improved land or not; state wants to use the land for water tower or headquarters of a party	n=1,513; each respondent randomly assigned to one vignette; several questions on the vignette such as agreement with the decision to expropriate on 4-point or 5-point response scales
Gibson 2008	South Africa	Judgments of fairness regarding the treatment of land squatters	2x2x2x2=16 vignettes; attributes: no other place to live or near to work (need); wait for government assistance or not eligible for government assistance (deservingness); land used by owner or not (owner's need); owner evicts squatter immediately or gives some time (rule of law)	n=2,054; each respondent randomly assigned to one vignette; 10-point response scale
Sundström 2012	South Africa	Willingness to comply with regulations	5 x 1 = 5 vignettes; attribute levels: officials are honest and four variants in which officials accept bribes	n=181; each respondent randomly assigned to one vignette; small scale fishermen; 7-point response scale on willingness to follow the regulations
Horne et al. 2013	Ghana	Bridewealth and norms regulating female behavior	3x2=6 vignettes where the man is described as beating the woman; attributes: completeness of bridewealth payment with the levels none vs. partial vs. full; domains with reproduction/using contraception vs. business/giving away money from her shop	n=276; each respondent randomly assigned to one vignette; 10-point response scale on expected social disapproval of the woman's behavior and man's violence

Our own studies do not only increase the scope of applications of FSEs in development research, they also contain characteristics such as providing respondents with more than one vignette which will be tested for the first time in a Subsaharan African remote-area context. Our research demonstrates how FSE can be fruitfully used in developing countries. All of the respondents in our studies live in remote rural areas. They are not familiar with surveys and research institutions. While it is difficult to evaluate the validity of FSE directly, our four studies

implemented in two surveys give us the possibility to compare FSE results within each survey and between the two surveys. Regarding construct validity we can investigate whether we obtain theoretically plausible results within each country context and across the country contexts. Even if the topics of the experiments differ between the two country samples, validity should be given when all four experiments provide meaningful results.

The two FSEs employed in Rwanda focus on similar topics, just wages and just incomes, and, therefore, consistency in the direction and significance of the effects of vignette attributes (e.g. wage/income levels and gender) across studies would be an indicator of validity. Further, in the two studies in Benin we present respondents, for the first time in a remote-area context, with more than one vignette, namely the full factorial. As an indicator of validity, we can analyze whether respondents can cope with this complexity. For example, constant response patterns across vignettes would suggest that the experiments did not work well and do not measure what is intended by the researchers. On the other hand, variation in responses per respondent could be interpreted in support of the method's validity.

As will be shown our results of the four studies are overall promising, but we also encountered some problems in conducting FSEs in a developing-country context which have to be considered.

In the following we present the four FSEs and their results nested within two survey studies, one in Benin and the other in Rwanda. The paper finishes with a conclusion and discussion of our findings and the applicability of FSEs in a developing-country context.

Two Applications of Factorial Survey Experiments (FSEs) in Benin (Study 1)

We implemented two FSEs in a face-to-face interview in Benin. The topics of the experiments were social norms regarding the physical punishment of children (experiment 1A) and the social acceptance of technology adoption (experiment 1B).

Study Areas, Sample, and Data Collection

The research presented in the present paper was part of a larger study on non-formal education carried out in the Fulani communities in Gogounou (North Benin), Baatonu communities in Banikoara (North Benin) and the Xwela communities in Come (South Benin). These districts were selected for a comparative analyses regarding farming, husbandry, and fishing activities; yet this distinction is not relevant for the FSE part of the survey.

We selected the villages in each district after careful discussions with leaders of farmer organizations and district agricultural service organizations. We considered criteria such as main activity in the village, geographic accessibility of the village, and population size. We selected three villages in Come (Zikpanou, Pedah-Comè, and Tossouhon); two in Banikoara (Somperekou and Kokey); and two in Gogounou (Gamarou and Katakpara). Within villages the households were randomly selected. We surveyed information on the household including all children within the household from the household head, and we interviewed one child per household for more detailed information. The children were selected based on quota criteria for sex and age. In this paper we use the household adult data and this sample comprises in total 491 adult respondents.

The interviews were conducted face-to-face from March to August 2012. All questions including the FSE were read out to the respondent in the local language. Three enumerators were involved in the data collection process. They received three days of training and were selected with respect to the local languages in the study region. From the household heads who were interviewed 77% were men and 23% women. Mean age was 45.26 years (median = 44 years) with a minimum of 22 and maximum of 99 years and a standard deviation of 12.32. All respondents have a very low formal education. In the study region, less than 10% attend the first years of the primary school (years one to four) and less than 5% complete primary school education. These 5 to 10% can be considered able to read and write in French. Further, about 10 to 15% can read and write in their local languages. Respondents would not be able to take a self-administered survey.

Experiment 1A: Social Norms Regarding Physical Punishment of Children

Social norms are behavioral proscriptions that are supported by positive and negative social sanctions such as social approval and disapproval (Hechter and Opp 2001). An FSE can be used to measure the extent to which norms and corresponding sanctions are prevalent in a social group and society (e.g., Jasso and Opp 1997). The physical punishment of children is widespread in many developed and developing countries (e.g., Straus 1991; Busmann et al. 2002; Gershoff 2002), including countries where this form of punishment is forbidden by law. We designed an FSE to find out to what extent physical punishment is socially accepted in rural Benin and how this varies according to social context. This allows us to obtain an idea of whether and to what extent there is an (un)conditional social norm regarding physical punishment.

Table 2 reports the attributes and attribute levels that were used in the vignettes on punishment of children. What is described in the vignettes is in line with theoretical considerations on second- and third-party punishment and social norms (Fehr and Fischbacher 2004). Social norms such as "you must not steal" are enforced by expected punishments when individuals, here children, do not comply. Experiments in the laboratory (Fehr and Fischbacher 2004; Diekmann and Przepiorka 2015) and field (Balafoutas et al. 2014; Diekmann et al. 2014) show that individuals expect punishment and are prepared to punish deviant behavior and enforce norms even if it is costly for them and regardless whether it is on their own behalf (second party) or on behalf of another person (third party). This can also be shown for non-student subjects in Africa, Asia, Oceania and South America (Henrich et al. 2006). Social preferences (Fehr et al. 2002: 18), direct as well as indirect reciprocity (Nowak and Sigmund 2005) and signaling theory (Przepiorka and Liebe 2015) are discussed, among others, as explanations for costly punishment. The laboratory and field experiments show for example that the extent of punishment is positively correlated with the extent of deviation from a social norm and that second-party punishment is stronger than third-party punishment (Fehr and Fischbacher 2004). Complementing this research and using the strength of FSE we do not investigate whether individuals do punish or not but how punishment is accepted in a social group or population depending on the severity of misbehavior (knocking over milk vs. stealing milk), type and severity of the sanction (scolding vs. beating) and type of punisher (a relative/uncle vs. stranger). The latter can be interpreted in terms of second-party punishment (family member) and third-party punishment (stranger). The situations in the vignettes referred to the respondents' own children. Does the acceptance of punishing behavior vary with the severity of the misbehavior, the type of sanction, and type of punisher? Is a more severe punishment

more accepted if the punisher is a peer/family member than a "stranger," as indicated by previous studies?

The attributes and attribute levels in Table 2 give a full factorial of eight vignettes (2x2x2) which we presented to each respondent who judged each vignette on a four-point response scale (not at all acceptable, not acceptable, acceptable, totally acceptable). An example of a vignette used in the survey can be seen in Figure 1.

Table 2: Attributes and Attribute Levels in Experiment 1A

Attribute	Levels
The matter	Knocked over milk, stole milk
Type of sanction	Scolding, beating
Sanctioning person	Relative/uncle, non-relative

Figure 1: Example of a Vignette in Experiment 1A

What do you think about the following situation?
A = Not at all acceptable B = Not acceptable C = Acceptable D = Totally acceptable

In the village …

Your child knocked over milk outside and was given a scolding by his uncle.

Overall, we find a very high acceptance of sanctioning behavior described in the vignettes. The majority of respondents, 82%, find it totally acceptable, 14% acceptable, and the remaining 6% not acceptable or not at all acceptable. However, this figure varies between the vignettes. For example, in a vignette in which the respondent's child knocked over milk outside and was beaten by someone of the village who is not a relative, acceptance decreases to 64%. However, the overall figures on the acceptance of physical punishment remain high.

In experiment 1A five respondents (1%) out of 491 did not answer all vignettes. Three of these respondents refused to answer the third vignette and two respondents the eighth vignette. In the following we restrict our analyses to the respondents who answered all eight vignettes in experiment 1A. A closer look on the response behavior of these 486 respondents reveals that 65% did not make a difference between the eight vignettes presented in experiment 1A and from those who do not differentiate 92% always choose the response category "totally acceptable" and 8% the category "acceptable." This may have two reasons. First, the experiment might have simply not worked and the respondents did not understand the task well.

Second, this reflects the large extent to which the social norm regarding sanctioning behavior is unconditional. We have indications that the second reason is more plausible. In experiment 1B (see below) including another eight vignettes, 26% of the respondents always choose the same response category. This number is considerably lower than in experiment 1A.

Table 3 shows the results of multilevel ordered logit models (Snijders and Bosker 2012: 310) with acceptance as the dependent variable with values 1 (not at all acceptable), 2 (not acceptable), 3 (acceptable), and 4 (totally acceptable). A multilevel model is appropriate because respondents answered multiple vignettes and hence the data is structured hierarchically (see also Dülmer 2016: 308). Level-one variables are the vignette attributes and respondents constitute level two. All independent variables were dummy coded and the analyses were conducted using the meologit routine in Stata. Model A in Table 3 contains all main effects and model B in addition all interaction effects between the vignette attributes. Model C has the same specification as Model B but excludes all respondents with constant responses across the eight vignettes answered. In all models we tested stepwise for random intercepts and random slopes by comparing the model fit of model specifications with and without assuming respondent heterogeneity with respect to the intercept as well as with respect to individual slopes of each vignette variable. This approach resulted in a model with a random intercept and random slopes for the two vignette variables "stealing milk" and "beating." Therefore, heterogeneity regarding individual thresholds of acceptance judgments (intercept) and variables effects (slopes) is present in the data.

As can be seen in Model A in Table 3 all effects of the vignette variables are highly statistically significant. Sanctioning one's own child is associated with higher acceptance levels if the child has stolen rather than knocked over the milk. Irrespective of the other attributes in the vignette, sanctioning is less acceptable if the child is beaten rather than scolded, and it is more accepted if a relative, the uncle, is carrying out the sanctioning compared to a non-relative from the village.

Model B in Table 3 includes interaction effects which we interpret, as mentioned above, following theoretical plausibility and reasoning on the interplay of the severity of misbehavior, the type of sanction, and category of punisher. The interaction effect "beating x stealing milk" shows that the difference of the conditional effects of beating on acceptance when the child has stolen milk versus knocked over milk is positive and statistically significant. In other words: beating the child is more accepted when the child has stolen rather than knocked over the milk.

However, while the conditional effect of beating when the child has knocked over milk is negative and statistically significant (see the coefficient of the variable “beating” in Model B), the conditional effect of beating when the child has stolen the milk is slightly negative and statistically insignificant (i.e. combined effect of the variables “beating” and “beating x stealing milk”, coeff. = -0.075, SE = .525, p = .866).

Model B also reveals that, irrespective of the misbehavior, it is more accepted to beat a child if the uncle is the punisher compared with a non-relative (see the positive and statistically significant interaction effect “beating x uncle”). Yet, while the conditional effect of beating when a non-relative is the punisher is negative and statistically significant (see the coefficient of the variable “beating” in Model B), the conditional effect of beating when the punisher is the uncle is statistically insignificant (i.e. combined effect of the variables “beating” and “beating x uncle”, coeff. = -.091, SE = .437, p = .836).

The results for the negative interaction effect “stealing milk x relative” shows that the difference of the conditional effects of stealing on acceptance, when a relative is the punisher (effect of the variable “stealing milk” in Model B) versus a non-relative (combined effect of the variables “stealing milk” and “stealing milk x uncle”, coeff. = 3.977, SE = .595, p = .000), is statistically insignificant. Yet the negative and statistically significant three-way interaction effect in Model B can be interpreted in a way that beating the child for stealing is less likely to be judged as acceptable if the uncle is the sanctioning person compared with a non-relative. Model C, in which constant responses across vignettes per individual are excluded, reveals the same substantial results as Model B.

Table 3: Results of Multilevel Ordered Logit Models for Sanctioning Behavior towards Children

	Model A	Model B	Model C
Vignette variables			
Stealing milk (vs. knocking over milk)	4.784*** (.503)	4.504*** (.560)	4.733*** (.723)
Beating (vs. scolding)	-.742* (.359)	-1.659*** (.420)	-3.009*** (.404)
Relative/uncle (vs. non-relative)	1.015*** (.154)	.605 (.280)	.559 (.272)
Beating x Stealing milk		1.584*** (.478)	1.857** (.533)
Beating x Uncle		1.569** (.379)	1.743** (.383)
Stealing milk x Uncle		-.527 (.472)	-.439 (.516)
Beating x Stealing milk x Uncle		-1.407* (.638)	-1.547* (.705)
Cut point 1	-14.583*** (.812)	-15.370*** (.867)	-8.559*** (.625)
Cut point 2	-10.209*** (.618)	-10.778*** (.660)	-4.043*** (.348)
Cut point 3	-4.830*** (.470)	-5.191*** (.511)	-.230 (.262)
Random intercept			
Variance (Constant)	34.004*** (5.688)	35.917*** (6.136)	4.925*** (1.089)
Random slopes			
Variance (Stealing)	13.252*** (3.224)	13.182*** (3.290)	13.109*** (4.261)
Variance (Beating)	17.527*** (2.843)	19.333*** (3.134)	11.794*** (2.513)
LL	-1,374.406	-1,353.185	-905.313
N (respondents)	3,388 (486)	3,388 (486)	1,344 (168)
LR-Test versus ordered logit model	$\chi^2(3) = 1{,}668.64$, p = .000	$\chi^2(3) = 1{,}698.01$, p = .000	$\chi^2(3) = 416.76$, p = .000

Notes: $^{+}$ $p<0.10$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$; unstandardized coefficients, standard errors in parentheses.

Taken together experiment A1 shows that sanctioning of misbehavior of children is generally strongly socially accepted in the study regions. At first glance, the overall high acceptance rate might suggest that the severity of the misbehavior, the severity of the sanction, and the persons who sanction do not affect normative judgments and hence the social norm is unconditional.

However, our experiment was still able to find remarkable variations and important differences in the judgments of sanctioning situations. Social situations are more acceptable if the child's misbehavior is more severe and the sanctioning person is a relative of the respondent. This seems to be in line with basic results on actual punishment behavior in laboratory and field experiments (Fehr and Fischbacher 2004; Henrich et al. 2006). Further, beating children is judged less positively than a non-physical punishment but is more accepted if the misbehavior is more severe and the punisher is a family member. Again, this mirrors behavior in laboratory experiments that shows a positive association between the extent of deviant behavior and the extent of punishment (Fehr and Fischbacher 2004; Henrich et al. 2006). Yet using a physical punishment for a more severe misdoing is less accepted when the punisher is a family member. This is an interesting result which can be followed up in future research. The content of the social norm on the punishment of children depends, therefore, to a considerable extent on the social context. With this experiment and (to the best of our knowledge) for the first time we were able to isolate these social-context effects and reveal the conditionality of the social norm on the punishment of children in a developing-country context. Complementing previous research on norm enforcement, the FSE is also less "artificial" compared with the laboratory and field studies based on games from behavioral economics (Fehr and Fischbacher 2004; Henrich et al. 2006), and it captures more complex interactions of social context variables (i.e. interactions between vignette attributes).

Experiment 1B: The Social Acceptance of Technology Adoption

The diffusion and acceptance of innovations is one of the most crucial topics in development research (e.g., Feder et al. 1985). In this context, especially the role of social networks and extension agents are discussed. At the latest since the classic work by Rogers (1962) it is well known that the diffusion rate increases when an innovation spreads via network contacts (Valente 1996 for a discussion of diffusion patterns). The role of extension services is, in contrast, seen as controversial (Aker 2011). Further relevant factors include characteristics of the decision maker (e.g., comprehension) – the knowledge component (Rogers 1962: 170). Against this background, we designed an FSE to find out whether the social acceptance of innovation adoption depends on the involvement of an extension agent, the adopter's level of knowledge about the innovation, or the adoption rate in the village (only a few vs. many). Tables 4 gives an overview on the attributes and attribute levels and Figure 2 shows an example of a vignette used in the survey. The full factorial comprises eight vignettes (2x2x2) and each

respondent answered the full set of vignettes. We would like to note that agricultural extension services are common in villiages in Benin and, hence, respondents are familiar with their work, albeit the service leaves room for improvement (Moumouni 2006; Moumouni et al. 2011). Therefore, in the FSE respondents had a common frame of reference when talking about a new agricultural technology/technique (e.g., in farming communities: cotton-pest management based on insecticides, chemical fertilization; in husbandry communities: prophylactic management of cattle diseases, cattle diseases treatement ragrding; in fishing communities: automatic fishing techniques).

Table 4: Attributes and Attribute Levels in Experiment 1B

Attribute	Levels
Testified by an extension agent	No, yes
Own knowledge	Not well, well
Adoption in the village	Only a few, many

Figure 2: Example of a Vignette in Experiment 1B

To which extent do you agree with someone who adopts a new agricultural technique?
A = Agree not at all B = Not agree C = Agree D = Totally agree

… whose advantage has not been testified to by anybody, that the person understands well, and which has been adopted by only few people in the village**.**

In experiment 1B 17 (3.46%) out of 491 respondents did not answer all vignettes. Yet there is no clear pattern in non-response in the sense that some vignettes are answered considerably less frequently than others. Similar to experiment 1A we consider the 474 respondents who answered all eight vignettes; 126 (26%) of these respondents choose the same response category in all eight vignettes. This share is considerably lower than in experiment 1A. We do not know whether this is due to the content of the vignette task in the FSE or high task complexity.

However, while 19% of those respondents who do not differentiate between the vignettes in experiment 1A do also not differentiate in experiment 1B, 81% do differentiate in experiment 1B. Yet there is a positive association between non-differentiation in experiment 1A and experiment 1B which is statistically significant at the 10% level ($\chi^2(1) = 2.974$, $p = .085$).

To shed more light on the reasons why overall 19% of the respondents show a constant response pattern across all vignettes in experiment 1A *and* 1B, we estimated a binary logit model (n = 468, McFadden $R^2 = .017$, robust SE, Huber-White estimator) with the dependent variable

taking the value 1 if the respondents choose the same categories in both experiments and 0 otherwise. As explanatory variables we included gender, age as well as a variable from follow-up questions on the interview answered by the enumerator. The enumerator rated the confidence of the respondent's answers on a scale ranging from low (= 0) to very high (= 5). The mean value of this variable is 2.16 with a SD of .88. The logit model shows that gender (coefficient = .270, SE = .313, p = .390) and age (coefficient = -.017, SE = .014, p = .214) do not have a statistically significant effect on the likelihood of providing constant response patterns in both experiment 1A and experiment 1B. Yet, higher enumerator confidentiality in the respondent's answers (coefficient = -.265, SE = .134, p = .048) significantly decreases the likelihood of constant response patterns.

Clearly, non-variation is present in the sample but it does not apply across the two experiments for the majority of the respondents and rather seems to be task specific which indicates validity of the experiments. On the other hand, there is a share of respondents who might not have provided confident and valid responses.

Overall and irrespective of the vignette attributes we find a high variance in the judgments of the vignettes in experiment 1B: 25% find the adoption of the innovation totally acceptable, 35% acceptable, 17% rather unacceptable, and 23% totally unacceptable.

Table 5 shows the results of multilevel ordered logit models regarding the acceptance of the behavior described in the vignettes. The vignette variables constitute the first level and respondents the second level. The dependent variable is the original response scale as shown in Figure 2. All independent variables were dummy coded. For all models we tested stepwise for random intercepts and random slopes by comparing the model fit of model specifications with and without assuming respondent heterogeneity with respect to the intercept as well as with respect to individual slopes of each vignette variable. This approach resulted in a model with a random intercept and random slopes for the two vignette variables "own knowledge" and "adopted by many." Therefore, similar to experiment 1A, heterogeneity regarding individual thresholds of acceptance judgments (intercept) and variables effects (slopes) is present in the data.

Looking at Model A in Table 5 we see that two out of three effects of the vignette attributes are highly statistically significant. The adoption of an innovation is more likely to be perceived as more acceptable if the adopter understands the innovation well and if many others in the village

have also adopted it. However, it does not play a role for the respondents' judgments whether an extension agent has testified to the advantages of the innovation. The non-relevance of extension services does also not change if we take possible interactions between vignette attributes into account (Models B and C, Table 5).

We find only one statistically significant interaction effect which is presented in Model B in Table 5. We see that the difference in the conditional effects of "own knowledge" on acceptance if many in the village have adopted the innovation and the conditional effect if a few have adopted the innovation is negative and statistically significant (interaction effect "own knowledge" x "many adopted"). In other words: The adoption rate in the village – how many others have adopted the innovation – is less relevant for the acceptance judgments if the adopter understands the innovation well. Further, the conditional effect of "own knowledge" if the innovation has been adopted by few in the village is positive and statistically significant as shown by the coefficient for "own knowledge" in Model B in Table 5. The conditional effect of "own knowledge" if the innovation has been adopted by many in the village is positive and statistically significant at the 1% level (combined effect of the variables "own knowledge" and "own knowledge x many adopted", coeff. = 1.733, SE = .248, p = .000).

However, from a theoretical point of view, an interpretation of the interaction effect in a way that the adopter's knowledge is less relevant when others have adopted the innovation is also plausible. Corresponding analyses of the conditional effects indicate that the conditional effect of "adopted by many" on acceptance if the adopter has a good knowledge of the innovation is positive and statistically at the 1% level (combined effect of the variables "adopted by many" and "own knowledge x adopted by many," coeff. = 2.574, SE = .263, p = .000). The coefficient for "adopted by many" in Model B represents the positive and statistically significant conditional effect of peers on acceptance if the adopter does not have good knowledge. Considering both theoretically plausible interpretations the interaction effect regarding "own knowledge" and "adopted by many" might indicate that the knowledge effect weakens the importance of peer effects and vice versa.

Table 5: Results of Multilevel Ordered Logit Models for the Acceptance of Technology Adoption

	Model A	Model B	Model C
Vignette variables			
Testified by extension agent (vs. not)	-.091 (.083)	.014 (.157)	.017 (.163)
Own knowledge well (vs. not well)	2.185*** (.194)	2.645*** (.241)	2.715*** (.242)
Adopted by many (vs. only few)	3.044*** (.219)	3.487*** (.266)	3.604*** (.266)
Testified x own knowledge		-.221 (.223)	-.230 (.231)
Testified x many adopted		-.132 (-.233)	-.135 (.239)
Own knowledge x many adopted		-.912*** (.245)	-.876*** (.250)
Testified x knowledge x many adopted		.284 (.334)	.290 (.343)
Cut point 1	-1.593*** (.244)	-1.385*** (.259)	-.611** (.184)
Cut point 2	1.016*** (.243)	1.254*** (.260)	1.889*** (.194)
Cut point 3	5.990*** (.283)	6.230*** (.299)	5.924*** (.251)
Random Intercept			
Variance(Constant)	23.142*** (2.636)	24.000*** (2.753)	6.290*** (0.803)
Random Slopes			
Variance(Own knowledge)	10.857*** (1.307)	10.830*** (1.303)	8.388*** (1.133)
Variance(Adopted by many)	13.721*** (1.561)	13.666*** (1.553)	10.105*** (1.300)
LL	-3,513.215	-3,503.414	-2,947.566
N (respondents)	3,792 (474)	3,792 (474)	2,784 (348)
LR-Test versus ordered logit model	$\chi^2(3)$ = 2,886.40, p = 0.000	$\chi^2(3)$ = 2,895.60, p = 0.000	$\chi^2(3)$ = 1,105.15, p = 0.000

Notes: $^+$ $p<0.10$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$; unstandardized coefficients, standard errors in parentheses.

Taken together, with this experiment, based on a novel research design, we can replicate some ideas and findings in the adoption of innovation literature such as the importance of social networks in the adoption process (e.g., Rogers 1962; Valente 1996). A finding that is certainly unexpected, given the policy support and investments in the region, is the insignificant effect of extension services. This does not mean, of course, that agricultural extension services are not

relevant at all. It just suggests that as with a technology itself, the existence of extension services is not enough to gain social acceptance of a novel technology and promote its adoption. It might further indicate that the current practices of extension services have to be reconsidered.

Two Applications of Factorial Survey Experiments (FSEs) in Rwanda (Study 2)

The just gender wage gap is well documented for Western and other countries (Weichselbaumer and Ebmer 2005 for a meta-analysis). Besides labor market data, researchers use FSEs to uncover differences in the perceived fairness of wages of men and women (Jasso and Rossi 1977; Jasso and Webster 1997). Often there seems to be a double standard in the sense that male wages are judged differently from female wages, which are devaluated. Thus, the same pay for the same work is more likely to be judged as unfairly too low if the worker is a man compared to a woman. Next to this gender wage gap the perceived fairness of wages might also depend on other characteristics such as the type of job, income level, and number of children which can be theoretically related to justice principles such as the need principle (Auspurg et al. 2017).

We implemented two FSEs on just earnings by way of face-to-face interviews in Rwanda. Given that gender remains a very important status characteristic in many developing countries and that there is a social hierarchy with women being in the lower status position, our primary aim was to test whether an FSE can also be used to study judgments of fair earnings in a developing-country context. Yet in the two studies in Rwanda we use one vignette per respondent because, compared to the physical punishment of children in Benin and the adoption of an innovation, gender (in)equality is a "sensitive topic" in Rwanda. The country officially follows a strong gender equality policy (Burnett 2011) and according to The World Economic Forum (2016), which annually prepares "The Global Gender Gap Index," Rwanda ranks fifth out of 144 countries, higher than many Western countries like Denmark, Germany, France, UK, and USA. But actual living conditions and citizens' attitudes are often at odds with official gender policy goals. This makes gender (in)equality a sensitive issue. We provide, therefore, similar to studies conducted in North America and Europe, one vignette per respondent. This design is adopted in order to lower the likelihood of socially desirable response behavior (Jann 2008; Beyer and Liebe 2015). This problem might be less severe if the FSE comprised many attributes; yet in our rather simple experiments we only employ two or three attributes since the relevant variations would be very obvious for the respondents.

Study Areas, Sample, and Data Collection

The FSEs were part of a larger project in which we were interested in surveying households in the transition from subsistence to a market-oriented agriculture production in Rwanda. With this, mostly state-driven agriculture transformation, new rural employment opportunities in and outside agriculture appear (Ministry of Agriculture and Animal Resources (MINAGRI), 2013). The three study districts (Burera district; Gakenke district; Musanze district) were selected on the criteria that farming is the predominant activity, and that there is sufficient market access. These selection criteria are relevant for the FSE in which we study judgments regarding wages of on-field casual workers and incomes of persons who are members of an agricultural cooperative.

The survey was carried out in October 2015, and the data were collected in face-to-face interviews. All questions including the FSE were read out to the respondent. A total of 14 enumerators and two supervisors (team leaders) were trained to understand the questionnaire in detail; in the use of tablets to download and fill in forms; and to send data to a server on a cloud for aggregation and retrieval. The data collection was conducted in two phases sequenced within two weeks. In the first phase, a tool that captured information on household demographic characteristics such as off-farm employment, cropping activities, and asset endowment was administered to representative household members who were knowledgeable on most aspects of the household production and employment activities. In the second phase the data collection was on an individual level. Here, we obtained data on well-being, gender relations, household decision-making, and the vignettes from 567 male and female respondents who live in 381 households; 42% of the respondents being men and 58% women. Mean age was 42.23 years (median = 40 years) with a minimum of 19 and maximum of 99 years and a standard deviation of 14.50. While 30% of the respondents have no formal education, 13% have a primary school education between one and three years, 38% between four and six years and 19% have more than six years of education.

In each of the statistical models regarding experiment 2A and 2B we adjusted the standard errors taking into account that some respondents live in the same household. Yet, additional analyses showed that there was little difference whether the respondents are from the same household or not.

Experiment 2A: Fairness of Wages Earned by Casual Workers

We used the FSE to find out whether double standards in just earnings (as termed by Jasso and Webster 1997) are also present in a developing-country context regarding wages of on-field casual workers. In this respect we employed a simple experiment with the two attributes gender and wage (see Table 6). The worker could be male or female and earn a low or higher salary. Originally we had a third income category, but errors were made in recording the answers to the corresponding vignettes, and we therefore cannot consider vignettes with this attribute level. Figure 3 presents an example of the vignette. We presented each respondent with only one vignette because we did not want the respondents to be aware of the attributes, as this might have led to biased responses (Jann 2008 for a similar approach in a vignette study in Switzerland). If the respondents are aware that discrimination is the research interest they might not reveal their true judgments.

Table 6: Attributes and Attribute Levels in Experiment 2A

Attribute	Levels
Gender	Men, women
Daily wage	RWF 500, RWF 1,000

Figure 3: Example of a Vignette in Experiment 2A

Mr. Nyirahabimana is an on-field casual worker. His daily salary is RWF 500. How do you judge the salary of the described person?

Unfair too low=-3; -2; -1; 0=fair; +1; +2; +3=unfair too high

Overall, we find variance in the judgments of the vignettes: 26% find the wage unfair too low using the endpoint of the scale (-3), 39% find it unfair too low (-2, -1), 25% judge the wage as fair (0), and 10% find the wage unfair too high (+1, +2, +3). This distribution is well in line with studies carried out in Western countries, especially the substantial share of respondents that uses the midpoint of the scale (e.g., the example in Auspurg and Hinz 2015: 94).

Table 7 shows the results of ordinary least square regression models. The dependent variable is the logarithm of the original response scale taking positive skewness into account. All independent variables were dummy coded. Model A shows that a higher wage is judged as fairer than a lower wage and this wage effect is highly statistically significant. Further, we find that wages by women and men are not judged differently. Model B indicates gender differences

because there is a tendency (statistically significant at the 10% level) that female respondents judge the higher wage as fairer than male respondents. This is no indication of discrimination but it implies that men and women have different reference levels for just wages. Such differences might well translate into labor market outcomes because justice perceptions can considerably affect bargaining over wages. This is the only gender difference that we found in the experiment.

Table 7: Results of Ordinary Least Square Regression Models for Just Wages of Casual Workers

	Model A	Model B
Vignette: wage of 1,000 (vs. 500)	.807*** (.048)	.702*** (.077)
Vignette: female worker (vs. male worker)	.062 (.046)	.059 (.046)
Respondent: female (vs. male)		-.106 (.085)
Vignette: wage of 1,000 x respondent: female		.176^{+} (.099)
Constant	.390*** (.038)	.460*** (.067)
R^2	.462	.467
N	407	407

Notes: $^{+}$ $p<0.10$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$; unstandardized coefficients, standard errors in parentheses, adjusted standard errors taking into account that some of the respondents live in the same household.

Taken together, this experiment finds a plausible variation in fairness judgments which is comparable to studies from Western countries (e.g., Auspurg and Hinz 2015: 94). Yet we do not find a gender gap in just earnings as is often presented in other studies. On the other hand, our findings point to different reference levels on just earnings by women and men and this might well affect bargaining processes in the labor market leading to higher income inequality.

Experiment 2B: Fairness of Incomes Earned by Cooperative Members

In the same survey as in Experiment 2A, we implemented another FSE on the gender wage gap. This time the vignette described an income earner who was a member of an agricultural cooperative. The attributes of the experiment and their corresponding levels are shown in Table 8. The income earner can be a man or a woman. Following theoretical ideas on justice related need principles (Auspurg et al. 2017) we also varied whether the income earner(s) is a single father, a single mother, or a married couple and whether he (she or they) has two, five, or eight

children. The attribute income ranges from RWF 150,000, over RWF 300,000 to RWF 600,000 per season (i.e. six months). This gives a full factorial of 3x3x3=27 vignettes of which Figure 4 shows an example. Each respondent judged one of these vignettes.

Table 8: Attributes and Attribute Levels of Experiment 2B

Attribute	Levels
Gender and marital status	Single Father, Single Mother, Married couple
Number of children	2, 5, 8
Income of sales per season	RWF 150,000, RWF 300,000, RWF 600,000

Figure 4: Example of a Vignette in Experiment 2B

Ms. Mugiraneza is an agriculture cooperative member. She is a single mother of five children. Her income from sales is RWF 150,000 per season. How do you classify the wage of this person?

unfair too low=-3; -2; -1; 0=fair; +1; +2; +3=unfair too high

Regarding the distribution of fairness judgments, 9% of the respondents find the income as unfair too low using the extreme response category (value -3), 31% find it unfair (values -2, -1), 31% perceive the income as fair (value 0), and another 29% as unfair too high (values +1, +2, +3). We use the original response scale as a dependent variable in multivariate regression analyses shown in Table 9. All independent variables were dummy coded.

Model A in Table 9 shows that, compared to single men, the respondents do not judge the income of single women and married couples differently in a significant way. This also implies that we do not find a gender wage gap regarding cooperative members. Yet we see two very strong effects of the other attributes in the vignettes. The more children the income earner has, the more unfair her/his income is judged. On the other hand, the more income the person described in the vignette earns, the more fair her/his income is judged. Apart from these main effects, we do not find remarkable interaction effects of the vignette attributes.

With respect to the heterogeneity of vignette judgments, Model B indicates a weakly significant (10% level) gender effect. Overall and independent of the vignette attributes, women tend to perceive the incomes as being fairer than men. Similar to study 2A, this difference might translate into labor market inequalities because women might negotiate wages and incomes

differently than men. They might, for example, accept lower wage offers as a result of different reservation levels of a fair wage/income.

Experiment 2A gives us the possibility to demonstrate a useful characteristic of FSE. Similar to research on life satisfaction and well-being which is based on a "happiness equation" (Frey and Stutzer 2002) Models A and B in Table 9 can be seen as fairness equations/regressions. Since we have included an income variable we are able to calculate the trade-off between income and children holding fairness perceptions constant. This can be done by dividing the coefficient for the number of children by the coefficient for income (i.e. -.225 / .003, but using the non-rounded value). This amounts to 69.29 x 1,000 = RWF 69,295 per additional child and season (the income coefficient in the regression models shows the value per RWF 1,000). Therefore, on average, a just earning premium per child and season would be RWF 69,295 (i.e. RWF 11,549 per month because a season lasts 6 months). Of course, it is a normative issue whether such monetary values should be used for policy purposes at all. If they are used, they should not be interpreted exactly to the cent but seen as another way of representing the importance of vignette attributes.

Table 9: Results of Ordinary Least Square Regressions Models for Just Incomes of Cooperative Members

	Model A	Model B
Vignette: women single (vs. men single)	-.069 (.155)	-.075 (.155)
Vignette: married couple (vs. men single)	-.154 (.151)	-.165 (.151)
Vignette: number of children (2,5,8)	-.225*** (.027)	-.224*** (.027)
Vignette: income level in RWF 1000	.003*** (.0003)	.003*** (.0003)
Respondent: women (vs. man)		.205^{+} (.123)
Constant	-.022 (.218)	-.145 (.226)
R^2	.253	.258
N	566	566

Notes: $^{+}$ $p<0.10$; * $p<0.05$; ** $p<0.01$; *** $p<0.001$; unstandardized coefficients, standard errors in parentheses, adjusted standard errors taking into account that some of the respondents live in the same household.

Taken together, like experiment 2A this experiment finds a plausible variation in fairness judgments regarding need principles (number of children) and income levels, which is

comparable to studies from developed countries (e.g. Auspurg et al. 2017). But we do not find a gender gap. Nevertheless, our findings strengthen the results of experiment 2A, which reveals different reference levels on just earnings by women and men.

Discussion and Conclusions

Survey-based experiments are used more and more often in social science research because they help to isolate effects of important (behavioral) determinants and to uncover causal relations. One of the most promising approaches is the multifactorial survey experiment including FSE (see Auspurg and Hinz 2015) as well as stated choice experiments (Louviere et al. 2000). Both methods are very similar in the sense that respondents evaluate situations which vary in their attributes. However, in stated choice experiments the respondents compare two alternatives and choose the alternative they prefer. In FSEs, which are discussed in the present paper, the respondents evaluate a single social situation on a response scale. FSEs are especially useful in the study of attitudes, judgments, and normative beliefs. In order to demonstrate the fruitfulness of FSEs for research in developing countries we presented four FSEs from two different African countries: Benin and Rwanda. These experiments covered different topics that have not been studied so far in Africa: social norms on child punishment, adoption of innovation, and just earnings. It can fairly be said that, in terms of construct validity, we have obtained meaningful and interesting results that are theoretically plausible and might stimulate future research on these issues in a developing-country context. Across the two experiments in Rwanda we find similar effects of wages and income levels as well as respondents' gender on fairness judgments. This consistency can be seen as an indicator of valid response behavior. Further, in line with some of the previous studies using FSEs in African countries, we found that useful results can already be gained based on a small sample of respondents. However, a step forward in analyzing the validity of FSE would be to replicate existing experiments in different country contexts. For example, the insignificant effect of the gender vignette variable in the studies on just wages and incomes in Rwanda might still be due to the country's strong policy on gender equality which citizens are highly aware of. Replicating the experiments in a similar rural employment context in another African country which does not have a strong gender policy would indicate to what extent justice judgments in Rwanda might be driven by social norms and official policy on gender equality.

Furthermore, face-to-face interviews are also more susceptible to interviewer effects (Loosveldt 2008) than self-administered surveys. The advantage of FSEs in the study of sensitive topics may therefore be limited by potential interviewer effects and, as was also our experience, it is

essential that interviewers are well trained to carry out an FSE. However, in any case FSEs are better than simple survey questions, which generally preclude the separation of relevant effects. Especially for sensitive issues and cultural norms it might be advisable to present respondents with only one vignette (see Jann 2008; Beyer and Liebe 2015). Otherwise, respondents could also judge multiple vignettes. This approach was used in our studies for the first time in Africa as well as remote area contexts and has proven useful in our two FSEs in Benin. At first glance, it seemed as if the high share of constant response behavior across vignettes per respondent in the first experiment in Benin indicated that the method did not work and provided invalid results. Yet we could show that the constant response patterns seem to be to a large extent task specific because we found it to a much lower extent in the second experiment. There seems to be a higher level of unconditionality regarding the acceptance of the (physical) punishment of children than the adoption of an innovation. Therefore, most respondents do not just "agree" or "disagree" with any vignette but discriminate between the different social situations presented. This supports the validity of FSEs in a developing-country and remote-area context. Yet there was also a share of respondents who did not differentiate between the vignettes in both FSEs in Benin and this was associated with low confidence in survey responses as evaluated by the enumerator. Further, we could not make sure that the order of the vignettes is randomized and hence order effects cannot be ruled out (Auspurg and Jäckle 2015). However, so far all studies that have been conducted in Africa (see Table 1) have used one vignette per respondent. We went one step further and considered multiple judgments per respondent which revealed new insights into the validity of response behavior. Future research might build upon our studies as well as the methodological research on vignette studies (response scales, complexity, etc.) that have been conducted in Western and Northern countries (Auspurg and Hinz 2015).

Similar to application of stated choice experiments there might be specific limits to the complexity of FSEs in a developing-country context (Bennett and Birol 2010). This brings us to the limitations and specific aspects of the method that have to be taken into account. Compared to Western countries in which FSEs are mostly conducted within a self-administered survey mode (mail or online survey), in developing countries FSEs mostly have to be conducted in face-to-face surveys. The reasons include the comparatively high illiteracy rate in these countries, the lower internet coverage, technical skills of research teams, and practical questions such as the availability of servers/clouds for immediate storage. The predominance of face-to-face interviews has some important implications. The situations described in the vignettes must be comprehensible for the respondents and therefore cannot be too long. Further, the length of the response scales might be limited. For example, in our study region in Rwanda most

respondents would not be able to understand the concept of a fairness judgment on an 11-point response scale. Yet for FSEs it might be advisable to use longer response scales so that respondents can express differences in their judgments (see Auspurg and Hinz 2015). This should be especially important when the vignette contains many attributes and hence a very large set of possible situations. In simple experiments, such as the one presented here, respondents can also express differences on a shorter response scale.

It is important to put into perspective the above-mentioned challenges and reservations to applying technically more sophisticated instruments in developing countries. For instance, mobile phones are widely spread in remote areas, and even people with low literacy quickly learn how to use them. As we have seen in other studies in a development context (albeit in urban areas), the motivational and training aspect of using technical devices for doing self-administered surveys can hardly be overestimated. In other words: the introduction of FSE and similar methods such as stated choice experiments in studies in developing countries should go along with facilitating self-administered or partly self-administered surveys (Tilley et al. 2013).

Without doubt experimental approaches within surveys are an important tool for social science research which is interested in the causal explanation of social phenomena. While randomized controlled trials (Banerjee and Duflo 2011) and behavioral (lab in the) field experiments (Cardenas and Carpenter 2008) are established methods in development research, there is much more room for survey-based experiments such as FSEs. It is the main point of this paper to demonstrate that FSEs can unfold its power also in a developing and, more important, remote area context. Furthermore, provided a few precautions have been taken into account, this method could be used as a vehicle to innovate social science research in a global South/remote area context; innovate it both in terms of the relationship between respondents and interviewer (i.e. assisted self-administered surveys), but also in terms of skills and capacities both for the respondent and the interviewer.

We hope that our four studies in two African countries and the ones that have been conducted previously by other researchers (see Table 1) encourage more applications in this direction. What is also needed (both in the global North and South) is more methodological research on FSEs regarding complexity of vignettes, attribute non-attendance, etc. The possibilities of combining methodological research with sociological topics are manifold. For example, FSEs can be used to study the foundation of decent and humane living conditions as well as key factors of poverty as perceived by the citizens themselves; this might be an important complement to conceptual, normative, and philosophical discussions on poverty reduction. The

empirical separation of (causal) determinants of individual normative judgments, fairness concerns, and agreement with political measures is relevant for basic social science research and political decision-making alike.

References

Aker, J. C. (2011). Dial "A" for agriculture: a review of information and communication technologies for agricultural extension in developing countries. Agricultural Economics, 42(6), 631–647.

Ali, D. A., Deininger, K., Goldstein, M. (2014). Environmental and gender impacts of land tenure regularization in Africa: Pilot evidence from Rwanda. Journal of Development Economics, 110, 262–275.

Auspurg, K., & Hinz T. (2014). Factorial survey experiments. Series: Quantitative Applications in the Social Sciences No. 175, Thousand Oaks, CA: SAGE Publications.

Auspurg, K., & Jäckle, A. (2015) First equals most important? Order effects in vignette-based measurement. Sociological Methods and Research, online first.

Auspurg, K., Hinz, T., Sauer, C. (2017). Why should women get less? Evidence on the gender pay gap from multifactorial survey experiments. American Sociological Review 82(1): 179-210.

Auspurg, K., Hinz, T., Sauer, C., Liebig, S. (2015). The factorial survey as a method for measuring sensitive issues. In U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, P. Sturgis (Eds.), Improving survey methods: Lessons from recent research (pp 137–152). London: Routledge, Talyor & Francis Group.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. Proceedings of the National Academy of Sciences of the United States of America, 111, 15924–15927.

Banerjee, A.V., & Duflo, E. (2011). Poor economics: A radical rethinking of the way to fight global poverty. New York: PublicAffairs.

Beaman, L., & Dillon, A. (2012.) Do household definitions matter in survey design? Results from a randomized survey experiment in Mali. Journal of Development Economics, 98(1), 124–135.

Bennett, J., & Birol, E. (Eds.) (2010). Choice experiments in developing countries: implementation, challenges and policy implications. Cheltenham: Edward Elgar.

Beyer, H., & Liebe, U. (2015). Three experimental approaches to measure the social context dependence of prejudice communication and discriminatory behavior. Social Science Research, 49, 343–355.

Burnet, J. E. (2011). Women Have Found Respect: Gender Quotas, Symbolic Representation, and Female Empowerment in Rwanda. Politics & Gender 7(03): 303–334.

Bussmann, K.-D., Erthal, C., Schroth, A. (2002). The effect of banning corporal punishment in Europe: A Five-Nation Comparison. Manuscript. Halle: Martin-Luther-Universität Halle-Wittenberg .

Cardenas, J. C., & Carpenter, J. (2008). Behavioural development economics: Lessons from field labs in the Developing World. The Journal of Development Studies, 44(3), 311–338.

Conroy-Krutz, J. (2013). Information and ethnic politics in Africa. British Journal of Political Science, 43(2), 345–373.

de Leeuw, E. D. (2008). Choosing the method of data collection. In E. D. de Leeuw, J. J. Hox, D. A. Dillman (Eds.), International handbook of survey methodology (pp. 113-135). New York and London: Psychology Press.

Diekmann, A., Jann, B., Przepiorka,W., & Wehrli, S. (2014). Reputation formation and the evolution of cooperation in anonymous onlinemarkets. American Sociological Review, 79, 65–85.

Diekmann, A., & Przepiorka, W. (2015). Punitive preferences, monetary incentives and tacit coordination in the punishment of defectors promote cooperation in humans. Scientific Reports, 5, 10321.

Duch, R. M., & Palmer, H. D. (2004). It's not whether you win or lose, but how you play the game: Self-interest, social justice, and mass attitudes toward market transition. American Political Science Review, 98(03), 437–452.

Dülmer, H. (2016). The Factorial Survey: Design Selection and its Impact on Reliability and Internal Validity. Sociological Methods & Research 45(2): 304-347.

Feder, G., Just, R. E., Zilberman, D. (1985). Adoption of agricultural innovations in Developing Countries: A survey. Economic Development and Cultural Change, 33(2), 255–298.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. Evolution and Human Behavior 25, 63-87.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. Human Nature, 13, 1–25.

Frey, B. S., & Stutzer, A. (2000). Happiness, economy and institutions. Economic Journal, 110(466), 918–938.

Gershoff, E. T. (2002). Corporal punishment by parents and associated child behaviors and experiences: A meta-analytic and theoretical review. Psychological Bulletin, 128(4), 539–579.

Gibson, J. L. (2002). Truth, justice, and reconciliation: Judging the fairness of Amnesty in South Africa. American Journal of Political Science, 46(3), 540–556.

Gibson, J. L. (2008). Group identities and theories of justice: An experimental investigation into the justice and injustice of land squatting in South Africa. The Journal of Politics, 70(3), 700–716.

Government of Rwanda (2014). Government Annual Report: 2013/2014. Kigali.

He, J. et al. (2014) Socially desirable responding: Enhancement and denial in 20 countries. Cross-Cultural Research, 49(3), 227-249.

Hechter, M., & Opp, K.-D. (Eds.) (2001). Social norms. New York: Russell Sage Foundation.
Henrich, J. et al. (2006). Costly punishment across human societies. Science, 321(5781), 1767-1770.

Horne, C., F. Nii-Amoo Dodoo, Naa Dodua Dodoo (2013). The shadow of indebtedness: Bridewealth and norms constraining female reproductive autonomy. American Sociological Review, 78(3), 503–520.

Kreuter, F., Presser, S., Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. Public Opinion Quarterly, 72(5), 847–865.

Krumpal I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. Quality & Quantity: International Journal of Methodology, 47(4), 2025–2047.
Jann, B. (2008). Erwerbsarbeit, Einkommen und Geschlecht: Studien zum Schweizer Arbeitsmarkt. Wiesbaden: VS-Verlag.

Jasso, G. (2006). Factorial survey methods for studying beliefs and judgments. Sociological Methods Research, 34(3), 334–423.

Jasso, G., Opp, K.-D. (1997). Probing the character of norms: a factorial survey analysis of norms and political action. American Sociological Review, 62(6), 947–964.

Jasso, G., Rossi, P. H. (1977). Distributive justice and earned income. American Sociological Review, 42(4), 639–651.

Jasso, G., Webster, M. (1997). Double standards in just earnings for male and female workers. Social Psychology Quarterly, 60(1), 66–78.

Johnson, T. P., & van de Vijver, F. J. R. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. J. R. van de Vijver, P. Ph. Mohler, Cross-cultural survey methods (pp. 195–204). Hoboken, New Jersey: Wiley.

Loosveldt, G. (2008). Face-to-face interviews. In E. D. de Leeuw, J. J. Hox, D. A. Dillman (Eds.), International handbook of survey methodology (pp. 201–220). New York and London: Psychology Press.

Louviere, J. J., Hensher, D. A., Swait, J. D. (2000). Stated choice methods: Analysis and application. Cambridge: Cambridge University Press.

Ministry of Agriculture and Animal Resources (MINAGRI) (2013). Strategic plan for the transformation of agriculture in Rwanda Phase III.

Moumouni, I. (2006). Services Development Perspectives in new Rural Districts in Benin: Case Study of the Agricultural Extension in Banikoara. *Proceedings of APEN International Conference 6-8 March 2006 at Beechworth,* Victoria, Australia.

Moumouni, I., Nouatin, G.S., & Baco, M.N. (2011). Du système formation et visites au conseil à l'exploitation agricole familiale au Bénin : rupture ou continuité ? *Cahiers Agricultures* 20(5): 376-381.

Mutz, D. C. (2011). Population-based survey experiments. Princeton: Princeton University Press.

Nowak, M.A., & Sigmund, K. (2005). *Evolution of Indirect Reciprocity.* Nature, 437, 1291-1298.

Przepiorka, W., & Liebe, U. (2015). Generosity is a sign of trustworthiness – the punishment of selfishness is not. *Evolution and Human Behavior* 37(4): 255-262.

Raihani, N. J., & Bshary, R. (2015). The reputation of punishers. Trends in Ecology and Evolution, 30, 98–103.
Rogers, E. M. (2003). Diffusion of innovations. 5th edition. New York: Free Press.

Rossi, P. H. (1979). Vignette analysis: Uncovering the normative structure of complex judgments. In R. K. Merton, J. S. Coleman, P. H. Rossi (Eds.), Qualitative and Quantitative Social Research (pp. 176–186). New York: Free Press.

Schaeffer, N. C., Presser, S. (2003) The science of asking questions. Annual Review of Sociology, 29, 65-88.

Shadish, W. R., Cook, T. D., Campbell, D. T. (2002).Experimental and quasi-experimental designs for generalized causal inference. Boston, Mass.: Houghton Mifflin.

Snijders, T.A.B., Bosker, R.J. (2012). Multilevel Analysis: An introduction to basic and advanced multilevel modeling. 2nd Edition. Thousand Oaks: Sage.

Straus, M. A. (1991). Discipline and deviance: Physical punishment of children and violence and other crime in adulthood. Social Problems, 38(2), 133–154.

Tilley, E., Bieri, S., Kohler, P. (2013). Sanitation in developing countries. A review through a gender lens. Journal of Water, Sanitation and Hygiene for Development, 3(3), 298–314.

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. Social Networks, 18(1): 69–89.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. Social Science Research, 38(3), 505–520.

Weichselbaumer, D., & Winter-Ebmer, R. (2005). A meta-analysis of the international gender wage gap. Journal of Economic Surveys, 19(3), 479–511.

World Economic Forum. (2016). The Global Gender Gap Report 2016: Insight Report. Geneva: World Economic Forum.

Zander, K. K., & Drucker, A. G. (2008). Conserving what's important: Using choice model scenarios to value local cattle breeds in East Africa. Ecological Economics, 68(1), 34–45.