Revision memo

Manuscript ref: FJDS-2021-May-0027.R1

"Hiding, Shirking, or Pleasing: Spousal Disagreement among Ugandan Maize Farmers"

This memo provides a detailed response to the remaining concerns from the editor and reviewers on our manuscript "Hiding, Shirking, or Pleasing: Spousal Disagreement among Ugandan Maize Farmers". Text in **bold** *is the original text from the editor and reviewers. Our response is in italics.*

Response to comments from the Editor:

Thank you for your revised paper entitled "Hiding, Shirking, or Pleasing: Spousal Disagreement among Ugandan Maize Farmers". A further round is necessary before we can make a final decision as to its publication. Although the draft is improved, the referees raise a number of further points which we would like you to address. Copies of the comments are attached at the bottom of this letter.

I tend to agree with the second reviewer who takes issue with your use of the term "shirking"; not only can it have pejorative implications which are not warranted, but one person's shirking can be another's "every day form of resistance", or just tiredness, sickness, effort management, stress management, or what have you. I am tempted to say that "only economists shirk" (or those with a certain dominant type of economists mindset).

We agree that a more careful use of the term shirking is warranted. As you will see in our response to reviewer 2 below, the reviewer was right in observing that shirking was not a correct interpretation of time disagreement where one spouse reports having worked less than the other assumes (category 3). Category 3 disagreement ***could be an indication*** *that spouses engage in shirking, but that hypothesis is only tested afterwards (when we consider inaction effects of the treatment with an indicator for how easy spouses can monitor each other). And as the reviewer also notes, we do not find evidence of shirking. As such, after reading the comments of reviewer 2 and how the term is used in the manuscript, we do not think the pejorative undertone of the term was the main reason why reviewer 2 takes issue with the term.*

At the same time, we ultimately believe that "shirking" is the most appropriate term for the hypothesis we test in the context of this study. Reviewer 1 also seems to recognize the purpose and context behind our usage of the term, and offers a couple of minor revisions to improve clarity in its use.

The term is used widely and fits within the broader literature on intra-household bargaining and cooperation, where some studies suggest spouses may hide decisions, actions and resources from each other. In particular, we aim to identify instances where one spouse contributes less labor than what is agreed upon within the intra-household contract. We would thus argue that the negative undertone is intentional (just like hiding and pleasing are not entirely neutral).

We further prefer the term because it has traction in labour economics models such as the theory of efficiency wages and moral hazard problems that stem from the fact that different parties to a contract have some discretion due to monitoring problems (Shapiro and Stiglitz, 1984). Our study aligns with

this literature in that our strategy to test for the presence of shirking exploits potential heterogeneity in the ability to monitor different activities.[1]

In light of this, we decided to keep the term in the analysis, but also make the following changes to the text to accommodate the concerns raised:

First, we added the following endnote in the introduction where reference is made to the concept of shirking for the first time:

> *We realize that the term "shirking" may carry a pejorative connotation. That is not the intention here. Rather, shirking is defined in the context of this paper as putting in less effort than what is expected, consistent with its usage in labor economics and models of efficiency wages and moral hazard problems that stem from situations in which parties to a contract have some discretion over their labor effort due to monitoring problems (Shapiro & Stiglitz, 1984). Our study aligns with this literature in that our strategy to test for the presence of "shirking" exploits heterogeneity in the ability to monitor different activities.*

We have also made various changes in the part on labor provision to make sure that we do not simply equate Category 3 disagreement with shirking, but only use the term when the hypothesis that Category 3 disagreement is related to the ability to monitor activities is tested. We conclude that the evidence does not support the hypothesis that spouses engage in shirking behavior.

We hope that these revisions clear and satisfactory. If further concerns remain, we are happy to discuss them in greater detail.

When resubmitting your revision please include an indication from you of the ways in which the comments have been dealt with, whilst justifying those you choose not to address. Although you will understand that I am not able to guarantee that we shall publish the paper, I am hopeful that these changes will make that possible.

We would also ask you to do everything possible to make the paper as succinct and short (maximum 8,000 words with as few tables as feasible) as you can. A brief Appendix (about 1,000 words) may be included with papers to present important detail such as data sources and definitions and perhaps descriptive statistics. However, detailed material (such as on data construction where this is involved, additional results and sensitivity analysis) should be provided in a separate file as Supplementary Materials, which ideally would also include data and do files (where applicable and presented clearly with annotation to facilitate replication). Supplementary Materials can be provided in any file format and will be deposited in Figshare with a citable DOI linked to the article.

It is JDS policy that authors of accepted papers should make the data and code available (to bona fide researchers), either in a website or by committing to provide on request. We would be grateful if you could include a statement in the acknowledgements of how you will make it available.

[1] *Note that this testing strategy would also exclude the alternative explanations mentioned by the editor above: for instance if a spouse "works less as a form of resistance", we see no reason why this would happen more for activities that are more difficult to monitor (on the contrary, if you want to revolt you would probably want to do it in front of the person you are revolting to).*

Data and code for this paper is publicly available on github. We will add this as well as a link to the github repository in the acknowledgments.

The paper is not in housestyle. Please ensure that your revised version follows our housestyle guides which are available on the site under 'Instructions to Authors'.

The paper was prepared in LaTeX, an open source typesetter. JDS author guideline notes that papers may be submitted in Word or LaTeX formats. Unfortunately, there is no LaTeX template available for JDS articles (while there is for Word). Nevertheless, we have tried as much as possible to follow housestyle.

Footnotes should appear as endnotes, after the text and before the references.

We now use endnotes that have been placed after the text and before the references.

Pay particular attention to requirements for bibliographic details in references (e.g. include place of publication for books and mimeos, page span for articles and chapters in books).

In addition, please provide the issue number only if each issue of the journal begins on page 1. In such cases it goes in parentheses: Journal, 8(1), pp–pp. Page numbers should always be provided. Please refer to the following link: http://www.tandf.co.uk/journals/authors/style/reference/tf_APA6.pdf

We use the apalike2 bibtex style file which follows APA6 guidelines.

Reviewer 1:

The authors have put a lot of effort into responding to the first round of comments, which I appreciate and believe makes the paper crisper and perhaps most importantly, easier for the reader to understand what is empirically supported (as opposed to being speculative). The null findings on gender norms are valuable to publish as the priors are reasonable and an insignificant result is still a finding that we likely see in print too infrequently. Researchers can work to explore this further, perhaps with larger sample sizes.

My major remaining concern lies with the T2 "decision hiding" hypothesis (shirking is fine), as it is unclear how "<u>decisions</u> are strategically hidden" and, I believe, an unnecessary (and untestable) layer.

We agree that decision making is a largely personal matter and it may seem strange to talk about "decision hiding". We were following the language used in Ambler et al (2021), who refer to decision hiding in several places in their article, eg:

> *"In the present context, information asymmetries could result in differing responses to survey questions if, for example, a woman indicates that she is involved in decision-making but her husband does not because she makes some decisions without her husband's knowledge."p 767.*

In the last two paragraphs of 780, they refer to "hiding of assets or decisions" several times, and they state that:

> *"The conceptual framework also indicates that disagreement will be higher for assets or decisions that are more likely to be hidden"*

However, we agree with the reviewer that decision hiding may be too strong, and that it will be hard to differentiate between decision hiding and different information sets as it is unclear how decisions can be monitored. Indeed, even Ambler et al (2021) note that:

> *"This behavior could be strategic or unintentional. She could be hiding her activities and thus the decisions concerning those activities (or the assets that she owns), or the husband simply may not be cognizant of the full range of decisions that she makes or assets that she owns."*

We therefore now refer to the T2 hypothesis as testing for "asymmetric information", and we remain agnostic about whether this is due to "strategic hiding" or simply due to the fact that "one spouse does not know the full range of decisions that the other takes". We have made changes throughout the manuscript to address this point, especially in instances where we have removed the term "intentional hiding of decisions" and used the more general term "intrahousehold information asymmetry". See also the next points.

On page 1, beginning around line 36, the authors write: We go one step further and, using a field experiment, formally test if this asymmetric information between husbands and wives <u>should be attributed to</u> spouses strategically hiding information from each other…Furthermore, we test for an alternative explanation: that discord is not the result of asymmetric information, but is instead <u>caused by</u> spouses answering survey questions in line with prevailing gender norms.

Given the absence of a baseline and control variables, I would advise against using causal language such as the text underlined above. The paper still makes a contribution by just framing the experiments as:

We go one step further and, using a field experiment, formally test if this asymmetric information between husbands and ~~wives should be attributed to spouses strategically hiding information from each other…Furthermore, we test for an alternative explanation: that discord is not the result of asymmetric information, but is instead caused by~~ spouses answering survey questions in line with prevailing gender norms <u>is associated with changes in the level of discord.</u>

We do attempt to identify a causal relationship, and believe this is what makes the paper an important extension to Amber et al (2021), who use observational data and build on a theoretical model to argue that disagreement is more than just measurement error. We believe our rigorous study design allows for the identification of a causal relationship.

Our field experiments randomly allocated households to a subset of the sample where only one co-head received information (control group) or to a subset of the sample where both spouses received information (treatment group). Because of the random allocation of the intervention, on average, both groups are statistically identical prior to intervention.

Using data from a baseline survey, we can confirm that the treatment and control groups are balanced on a series of baseline characteristics. The balance table below shows differences between treatment and control groups before the interventions on 10 characteristics for treatment 1 (top panel) and treatment 2 (bottom panel). We find that from the 20 comparison, only 2 are significant at the 10

percent level, which is what is to be expected due to pure chance alone. Furthermore, protocol adherence was high. Also, given the highly controlled one-to-one treatment through the enumerator-respondent interaction and the fact that all control units received a placebo treatment, we can exclude any systematic differential exposure to other influences.

Treatment T1				
	average	diff	se	P-val
Household size (nr)	7.77	0.002	0.266	0.993
Area of maize (acre)	1.62	0.010	0.080	0.901
Number of bags harvested	3.79	0.650+	0.350	0.064
Used improved seed variety (1=yes)	0.34	0.031	0.036	0.384
House has brick wall (1=yes)	0.69	0.020	0.034	0.566
At least on spouse has only primary education (1=yes)	0.91	0.002	0.022	0.921
Average age of spouses	38.72	-0.065	0.976	0.947
Distance to nearest agro input shop (kg)	5.44	0.324	0.405	0.425
Access to agricultural extension (1=yes)	0.09	0.033	0.023	0.163
Has non agricultural income source (1=yes)	0.40	0.046	0.037	0.218
Treatment T2				
	average	diff	se	P-val
Household size (nr)	7.49	0.284	0.209	0.174
Area of maize (acre)	1.51	0.120+	0.066	0.070
Number of bags harvested	3.97	0.472	0.305	0.122
Used improved seed variety (1=yes)	0.37	0.003	0.031	0.913
House has brick wall (1=yes)	0.70	0.003	0.029	0.920
At least on spouse has only primary education (1=yes)	0.89	0.019	0.019	0.318
Average age of spouses	40.26	-1.610	1.859	0.387
Distance to nearest agro input shop (kg)	5.34	0.424	0.337	0.209
Access to agricultural extension (1=yes)	0.11	0.017	0.020	0.406
Has non agricultural income source (1=yes)	0.45	-0.005	0.032	0.872

Note: + denotes significance at the 10 percent level. Average denotes sample averages. Diff is the difference in averages between treatment and control (Average Treatment Effect).

For the first treatment, through the intervention where both spouses receive information in the treatment group, any information asymmetry between spouses is expected to be reduced as now both spouses know all the steps involved in maize farming. As this is the only difference between the treatment group and the control group post-treatment, differences in discord between the two groups can be attributed to the intervention. Similar arguments hold for the other treatment, where a random sample of households is exposed to an intervention that is designed to challenge prevailing gender norms and stereotypes by showing a male and female actor farming together as a couple in a video (i.e. projecting a cooperative way of farming as a household, thereby challenging the local consensus that maize farming is a male activity) (treatment group). The extent of discord is then compared to the control group of respondents who are shown a male actor farming, and any difference post-treatment can be attributed to the intervention challenging prevailing gender norms and stereotypes.

In principle, control variables are not needed for identification in case of RCT designs. They are sometimes included to adjust for imbalances (not applicable here) and they could increase precision but only if well chosen (https://www.povertyactionlab.org/resource/data-analysis).

However, in line with the previous point, we agree with the reviewer that the decision-hiding hypothesis is too strong and our intervention does not demonstrate that spouses necessarily engage in strategic hiding. In other words, we now follow the recommendation the reviewer makes below, to "…just test the effect of reducing information asymmetry between spouses (T2) on discord."

In this revised version of our manuscript, the paragraph the reviewer refers to above has been rewritten to avoid causal claims related to "strategic hiding", as follows:

> *In this paper, we investigate spousal disagreement in survey responses from monogamous smallholder maize-farming households in eastern Uganda. Drawing on Ambler et al (2021), we explore the same three types of explanation for discord—random measurement error, bias caused by men and women interpreting questions differently, and asymmetric information where men and women have only partly overlapping information sets. However, while Ambler et al (2021) use variation in the probability of overall disagreement by asset and activity to draw conclusions about asymmetric information, we designed a field experiment to formally test the asymmetric information hypothesis. This is done by randomly assigning a sample of households to a video-based intervention designed to reduce information advantages of a single spouse, and then comparing spousal disagreement to a control group. Furthermore, we test for an alternative explanation: that discord is not the result of asymmetric information, but is instead caused by spouses answering survey questions in line with prevailing gender norms. Inspired by recent research on the use of role models to change attitudes and behaviour in traditionally male dominated sectors, we test the impact of a video-based intervention that aims to promote a mental image of maize farming as a cooperative or joint venture in which both spouses play an equal role (Porter and Serra, (2020), Bernard et al (2015)).*

Why not just test the effect of information provision (T2) on discord? That alleviates the need to:

This point is much appreciated, and this is exactly what we do now. See above.

i. explain why it is first proposed that "For disagreement about the female co-head's role in decision-making, the focus is on hiding," (p.13, line 9) but not for males (although the results on page 18 do find that "that male co-heads may also hide decision-making from the female co-head.");

We look at disagreement about decision making for both spouses. The text defines this for the female co-head as:

> *"the likelihood that the female co-head reports she was involved in decision making but the male co-head says she was not."*

and for the male co-head in the next sentence as:

> *"the likelihood that the male co-head reports he was involved in the decision but the female co-head says he was not. "*

ii. explain what "hiding" a decision means (and apologies pages may refer to the manuscript or pdf); p.21.line29. The authors write: Finally, if disagreement is due to strategic hiding, we expect that facilitating mutual monitoring would be relatively more effective for decisions that are easier

to hide… We then estimate conditional average treatment effects by interacting the treatment indicator with the measure of how easy it is to hide decision or activity.

The part that is referred to here comes from the section that discusses the hypotheses and tests. We wanted to keep this part general and refer to "decisions, actions and assets". However, later in the text where we present results for decision making, we acknowledge that concepts such as "decision monitoring" or "determining if one decision is easier to hide than another decision" may be difficult to operationalize. This is why, for decision making, we did not ask experts how easy it is to monitor decisions, and hence do not include an interaction between this variable and the treatment in Table 2.

In this revised version of our manuscript, we have rewritten the paragraph to now only discuss activities (and removed all references to decision making). We also added a footnote that this part of the analysis is only done for disagreement related to labor time allocations and not decision making, as it is hard to operationalize H_d as an indicator denoting how easy it is to monitor particular decisions.

I appreciate the effort to relate this to the spousal "hiding" literature, and I think that connection works well with the labor shirking arguments. But it seems a stretch to associate changes in discord by decision as related to a strategic hiding choice? In addition to trying to conceptualize what a hidden decision means, all of the activities resulting from those decisions can be monitored (though granted perhaps to varying degrees re the exercise to set thresholds for H discussed below). It is also unclear how viewing the video changes monitoring capacity. Instead, why not simply report these results as an experiment as to whether providing joint information is associated with changes in discord? The paper could be framed something like "Exploring Spousal Disagreement among Ugandan Maize Farmers".

In line with the suggestions of the reviewer, we have rewritten the part on decision making as testing how reducing information asymmetry between spouses though our video intervention affected disagreement.

iii. explain what the variation in decision visibility is (per priors or the expert's scoring).The authors construct two indicators based on expert opinion for how easy an activity is to monitor (H1) and the degree to which an activity is in the male domain (H2). The expert's continuous scores are translated to a 0/1 variable based on thresholds of 40 and 55. It would be helpful to see a distribution of the responses to understand the logic behind the chosen cut-offs, and also to understand the variation across activities by visibility (and gender role). If there is little variation, this is another reason for not adding on the "hiding" frame for decisions (with T2) and may also help to explain the results for T1.

- The indicator H2 takes a value of one for an activity of the average expert score was larger than 40
- The indicator H1 takes a value of one for an activity or decision if the average expert score is larger than 55, situating it in the male domain; and zero otherwise

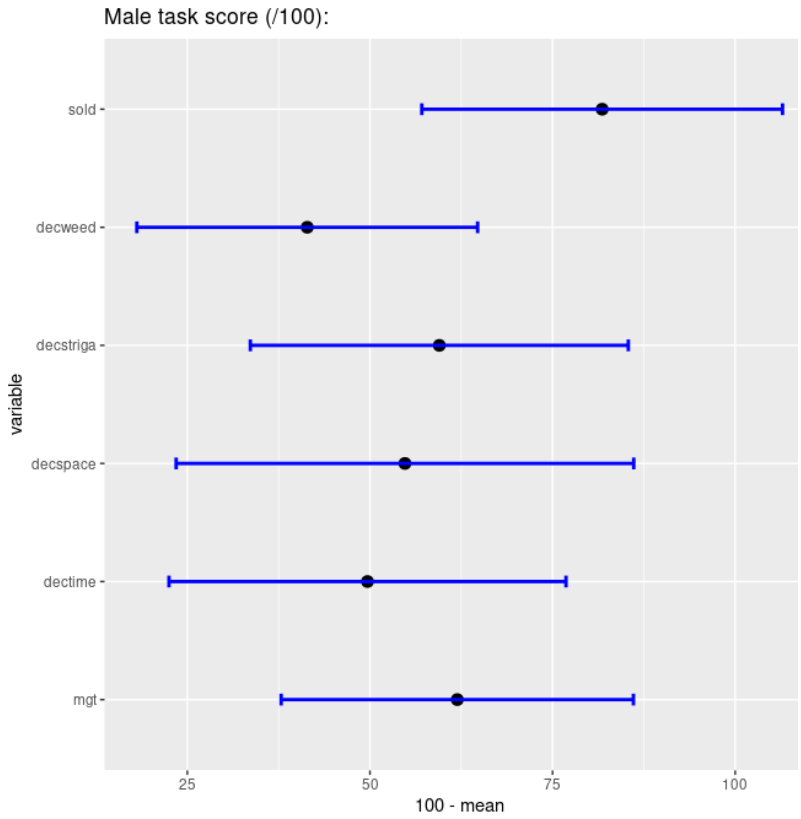
Note that we define H2 as the ease of monitoring indicator and it applies only to actions and activities, and not decision visibility. For H1, which is a proxy for the gender domain of the decision or activity, we consider both decisions and activities.

In this revised version of our manuscript, we have adapted the manuscript at different points to make it clearer that for disagreement related to decision making, we can not make statements related to "decision hiding" as it is hard to operationalize the concept of "monitoring of decisions".

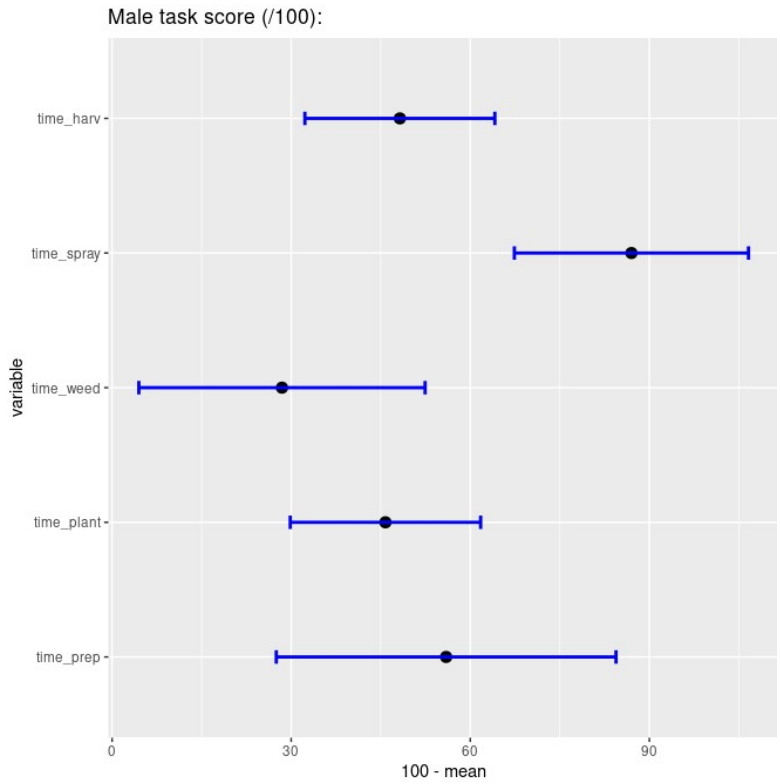
The reason for the choice of the cut-offs for defining the binary indicator based on expert opinion for how easy an activity is to monitor (H1) and the degree to which an activity is in the male domain (H2) are explained in the following note in the text:

> *We convert the indicator to a dummy to keep the analysis simple. The threshold was chosen to maximize variability, that is, we chose a cut-off to make sure we have approximately 50 percent of 1s and 50 percent of 0s. The results have proven not to be sensitive to conversion into a binary indicator nor the choice of the threshold for defining the binary indicator. The same is true for H1 below.*

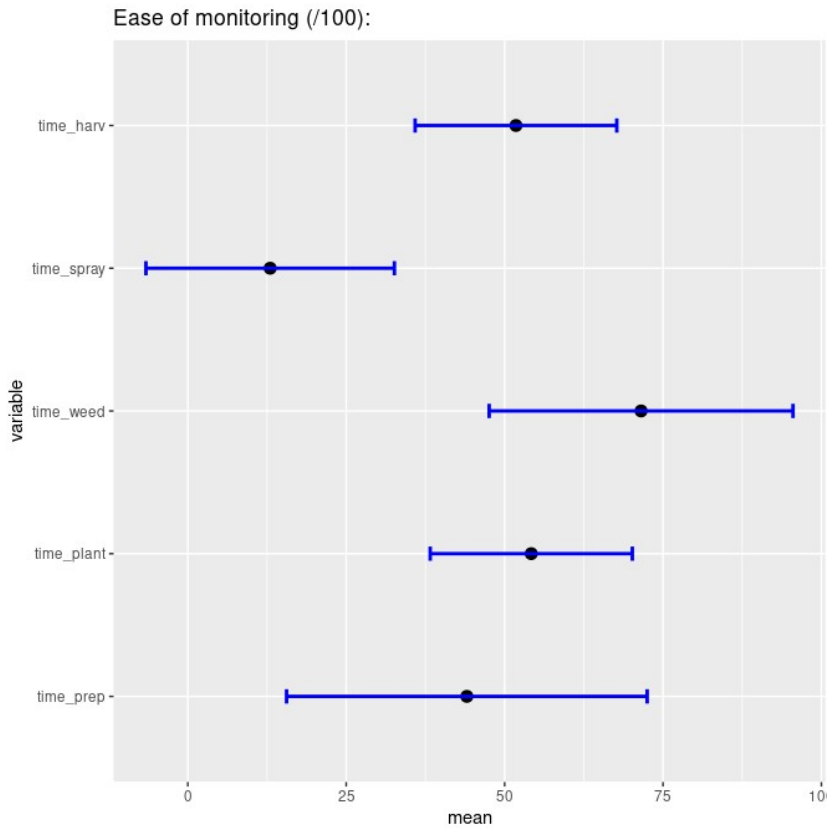
To get an idea of the distribution, one can have a look at bottom rows in table 1 in the text for the male task scores (H1) of the decisions. This is based on averages of the scores by the different experts. Here is a graphical representation that also has standard deviations.



At the bottom of table 3, we also report averages for the male task scores (H1) but now for the activities. Here is a graphical representation with standard deviations:



Finally, at the bottom of table 3, we report averages for the ease of monitoring score (H2) for the activities. Here is a graphical representation:



The “hiding” hypothesis could be related to labor shirking, though again, I still don’t see the value of extending this to “strategically hiding labour contributions” since we have no idea if behavior is strategic as in: Next, we conduct a series of regressions to test if disagreement is due to strategically hiding labour contributions within the household...(p.20) Why not just use the phrase “shirking?”

Agreed. In the revised manuscript, we have deleted the term strategically throughout the text. We now just use shirking, and only for instances where we consider interactions of T2 with H2 (see the point made by reviewer 2 on our inappropriate use of the term “shirking”).

Minor comments

p.3.line 29 Finally, we conclude that the video-based intervention does not affect estimates of shirking. Consider adding the underlined word “estimates” as shirking itself is not measured.

Done

p.5.line 25 spouses respond in line with what society expects from them

Corrected

p.15.line 20. However, we feel that differential disagreement on different decisions, asset categories, or activities need not always reflect the act of one co-head intentionally deceiving another. Consider slightly altered language to that underlined above such as: “it is plausible”

The comment is much appreciated. We have removed the phrase, “we feel that”.

p.16.line 37: which we refer to as T2, randomly varies whether the video is viewed jointly or not and the gender composition of the co-head(s) of the household. Consider adding the underlined clause in to the sentence above for greater clarity.

We have change this to “The second intervention, which we refer to as T2, randomly varies whether the video is viewed by the spouses jointly or by one of the spouses only.”

p.19.line 39. The authors write: Ambler et al. (2021) test for asymmetric measurement error. They define asymmetric measurement error as ...measurement error that occurs systematically and leads to different patterns of responses for men and women (p. 770)…On the following page 20, line 28 they write: Therefore, instead of referring to this as asymmetric measurement error, we see this more as bias caused by factors that emanate from the individual level.

I agree with the authors. The authors might consider moving their paragraph on page 10 up - acknowledge the Ambler phrasing but note that they are going to refer to this as cognitive bias. Either way they could simplify the phrasing and cut out a few words, but if the current ordering works better for them that is fine too.

Thank you for this suggestion. We have moved this up and included this as a footnote in the part where we discuss the Ambler at al (2021) strategy the first time.

p.29 ~~As for decision-making~~, we consider disagreement related to labour provided by the female co-head and disagreement related to labour provided by the male co-head separate

Corrected.

p.32.line44 are specific to a very specific population. Consider, are specific to a particular population….

Done.

Do any regressions run at the plot level cluster standard errors (at the household level)?

Yes, all plot level regressions cluster at the household level (as this is the experimental unit) using the following R function:

vcovCR(model, cluster=hh_id, type = "CR2")

See https://www.rdocumentation.org/packages/clubSandwich/versions/0.5.5/topics/vcovCR for more information and the the small sample correction used ("CR2")

In the revised manuscript, we have added notes to the tables indicating that standard errors are clustered at the household level.

Referee: 2

Comments to the Author

The authors did a lot to improve the paper, however there is more to address. It is an interesting paper on decision-making, which could fit nicely in the other decision-making papers within the gender and development literature. However, additional aspects of gender have to be taken into consideration in order for it to fit.

1. I know these decision-making papers well and some of the interpretations of are incorrect or poorly stated.

Thank you for pointing this out. We have carefully reviewed the relevant sections and re-read the papers to make sure we did not misrepresent what is in there.

Additionally, these are some changes we made in relation to the more specific comments of the reviewer related to the interpretation of the different papers we build upon:

- **Page 5 Careful to the wording of sentence right after in depth discussion of Ambler et al. (2021). The authors also do not believe disagreement "always reflect(s) the act of one co-head intentionally deceiving the othe**r."

It is true that they state on page 771 that this does not always needs to be intentional.

> *"This behavior could be strategic or unintentional. She could be hiding her activities and thus the decisions concerning those activities (or the assets that she owns), or the husband simply may not be cognizant of the full range of decisions that she makes or assets that she owns."*

This is repeated on page 780

> *"Prediction 3 indicates that the hiding of assets or decisions (whether intentional or unintentional) would result in variation of the probability of overall disagreement by asset or activity."*

In our manuscript in section 4, we already recognized that disagreement can also be unintentional:

> *"Ambler et al (2021) argue that rejecting both types of measurement error indicates that disagreement also partly reflects a real difference in what spouses know from each other, a situation they refer to as asymmetric information within the household. While they admit that some of this disagreement may be unintentional, they also note that the fact that disagreement seems to be highest among assets that are more easily hidden (small livestock, poultry, and small durables) suggests that at least part of this hiding is strategic."*

Related to this, we also added following endnote:

> *In the literature, this type of disagreement is interpreted in different ways. For instance, Ambler et al (2021) note this is consistent with women (intentionally or unintentionally) hiding decisions from their husband, while Annan et al (2021) consider this proof of women taking power. We follow Ambler et al (2021) in interpreting this as instances where the female co-head could be hiding her decisions or the husband simply is not cognizant of the full range of decisions that she makes.*

- "**Page 5 The disagreement based on social/cultural norms that is discussed of Acosta et al. (2020) in the last paragraph of section 2, is used to partly explain asymmetric measurement error in Ambler et al. (2021), which you explain in another part of the manuscript, but it is blurred here."**

We agree with this helpful comment and note that this was also mentioned by the first reviewer. The revised manuscript now addresses this issue in full by bringing forward (into Section 2) the text where we note the link between asymmetric measurement error in Ambler (2021) and differential interpretation of questions (due to e.g., norms).

- **"Page 9 Ambler et al. does not state they test for just cognitive bias. Consider revising"**

Ambler et al (2021) test for measurement error, asymmetric measurement error, and asymmetric information. We argue that this asymmetric measurement error can be considered some type of bias as this is related to how individuals interpret questions.

We noted this in the following part (that is now brought forward and included in section 2 as an endnote):

It is important to note the implicit assumption that asymmetric measurement error has its origin at the level of the individuals within the household (for example the female co-head has lower education leading her to systematically underestimate effort), and so should not lead to heterogeneity within the household. Therefore, instead of referring to this as asymmetric measurement error, we see this more as bias caused by factors that emanate from the individual level. Generally, this individual level variation will manifest itself in differing cognitive capacities of co-heads and differences in how questions are interpreted by male and female co-heads (Ghuman et al 2006). In this paper, we will refer to this type of disagreement as cognitive bias, rather than categorizing this as a type of measurement error.

2. I am not convinced that shirking is the correct interpretation for the time use analysis. In agricultural households in Uganda, women saying they work more in particular activities on particular plots (weeding for example) than their spouses say they work does not likely mean wives are shirking. In this context, both men and women likely are working long hours. Women, in particular, are likely engaged in overlapping unpaid and market activities on and off the household farm and overburdened with work. The labor section concludes, "More surprisingly, we also find that shirking by the female co-head is lower for activities that are harder to monitor," which suggests to me, shirking is indeed the wrong way to be thinking about this in this context. There are also a number of time use specific limitations to this type of data, which is not cited anywhere in the paper. And these data limitations limit what can actually be inferred from the findings.

Thank you for this observation. We agree that we should not interpret Category 3 disagreement (i.e. the co-head reported having spent less time than the other co-head thinks they did) as shirking; Category 3 disagreement may be an indication that shirking is going on, but there may be other reasons. In fact, by looking at the impact of the treatment that provides both spouses with equal information and the interaction with ease of monitoring, we aim to test for shirking behavior, and indeed, we do not find evidence of shirking by women co-heads; on the contrary.

In the revised manuscript, we have rewritten the part on labor provision to make sure that we do not simply equate category 3 disagreement with shirking, but only use the term when the hypothesis is tested (that category 3 disagreement is related to the ability to monitor activities). We conclude that the evidence runs against the hypothesis that spouses engage in shirking behavior.

We included a short discussion of the limitations of our labor/time use data and potential implications for our findings in the text:

The way in which we collected data about spouses' labor time could have been subject to some limitations, that could have implications for symmetric measurement error as well as asymmetric measurement bias. Yet, as there is no reason to believe these challenges manifest differently in our respective treatment and control groups, there are no likely implications for the results of our experiments. The limitations relate to the stylized questions asking each of the co-heads about total labor days spent on specific maize farming related activities on specific maize plots over the previous maize growing season which may have been subject to recall bias, particularly for less salient or irregular activities (Arthi et al. 2018; Seymour, Malapit, and Quisumbing 2020). If such activities are gender specific, that could have contributed to observed difference between male and

female co-heads. There are also cognitive aspects to properly identifying the activity in question and remembering and aggregating all instances of labor spent on the particular activity, which is possibly variable, over the reference period (Seymour, Malapit, and Quisumbing 2020). Gender differences in education levels and numeracy skills could have induced gender specific challenges with regard to the latter. Lower education levels are also associated with higher likelihood of overestimation in recall (Arthi et al. 2018). Gender specificity of activities and differences regularity and variability in time spent on the activities could have influenced the ease of identifying the activity and aggregating total time spent on it. While the reference period (previous maize growing season) is relatively long, increasing the risk of recall and estimation based on established patterns of activity, it is likely to be clear and salient for these maize farmers, and not likely to be subject to gender differences in correctly identifying it (Seymour, Malapit, and Quisumbing 2020). Seasonality bias is not likely.

Seymour, G., Malapit, H., and Quisumbing, A. 2020. "Measuring Time Use in Developing Country Agriculture: Evidence from Bangladesh and Uganda." Feminist Economics 26 (3), 169–99.

Arthi, V., Beegle, K., De Weerdt, J., and Palacios-López, A. 2018. "Not Your Average: Job Measuring Farm Labor in Tanzania." Journal of Development Economics 130, 160–72.

3. For the labor analysis, the authors use a difference of one standard deviation to define disagreement without any explanation for this specific threshold. This would be a good place to build on the time use literature (methods and issues) which should have already been discussed.

The problem we aim to solve is reminiscent of the problem one encounters when trying to compare outcome variables that use different units (eg. progress in terms of learning outcomes (perhaps measured as a percentage) and productivity (perhaps defined as kg/acre)). In such cases, it is common to standardize the variables to put them on the same (unit-less) scale.

The choice to take a one standard deviation cut-off seemed reasonable to us (as opposed to eg. 1.04). But it is true that this is a choice that we as researchers made. At the same time, we do not immediately see why a narrower (eg 0.5) or a larger (eg 2) threshold could bias our estimates.

4. Spouses in the monogamous relationships may be part of an extended family. There are likely other family members that take part in maize production on the farm. I am not convinced by your response that it is not useful to include the extent to which decisions are made by one or both of the "co-heads" as compared to decisions that also include others. In the review response it states "…we expected most impact on individuals directly involved" in the interventions. However, if others outside the "co-heads" are involved in the decisions, this does, in fact impact the effectiveness of your interventions and the extent that your analyses actually could pick up change. If there are major differences by gender, this also could impact the interpretation of your findings. You have the data, I do not see why this cannot be done.

It is true that spouses in monogamous relationships may be part of an extended family and we do not deny that other family members may take part in maize production. For instance, in section 5, we explain how this was captured in the data collection. In this paper, we study spousal disagreement and therefore focus on male and female co-heads.

The field experiments we implement are randomly assigned to households. Hence, on average, households in the control group and the treatment group are affected by the extent to which others participate in decision making to the same degree. In other words, by design (i.e. because of the randomization), this potential confounder (and all other confounders one can think of like for instance the number of plots cultivated, access to agricultural extension, distance to agro-input shop, etc) are highly unlikely to affect the (average) difference in outcomes between the two groups.

At the same time, we agree that there may be heterogeneity in the treatment effect related to the potential confounder. For instance, it may be that households where more decisions are made by more people are more (or less) receptive of the intervention that attempts to challenge gender norms and customs. However, to investigate this kind of heterogeneity in treatment effects, one would need to interact the treatment indicator with an indicator of the degree of involvement of others, much like we investigate interactions with the "ability to monitor" and "gender domain" variables.

We believe that treatment heterogeneity related to the involvement to outsiders, although interesting, would take the focus of the paper in a different direction. This is in contrast to the treatment heterogeneity with respect to ease of monitoring and gender domain, which has a clear rationale in the context of the hypotheses we test.

Minor

• In the categories the define decision-making in the paper include "either alone or jointly with spouse"? Yet, decisions can also be jointly made with others are not with the spouse? How are these integrated into the decision-making involvement categories?

See the response above. While we did provide the possibility to survey participants to indicate if others were also involved in decision making, we focus on spousal disagreement in this article. As such, our indicators of agreement and disagreement are only based on decisions by spouses alone and/or jointly with the other spouse.

• The study focuses on the spouses on monogamous households in Uganda, yet you also state that your sample is representative. Were households with monogamous co-heads randomly assigned to the treatment and comparison groups? This is not what it states in the manuscript. Is it that there were non-couple households dropped from your sample for the analysis?

*Our sample is representative of *monogamous* households in Uganda. Households with co-heads in monogamous relationships were randomly assigned to the treatment and comparison groups at the design stage of the study.*

• Why does the intervention include arms with inequal samples: 261 compared to 240, and 261 compared to 270 (540/2)? It needs to be explained.

The unequal treatment allocation in our case originates from a 3x3 factorial design set-up, with three levels for the first factor (showing a video to the woman alone, man alone, or couple for the recipient factor) and three levels for the second factor (video featuring man alone, video featuring woman alone, video featuring a couple for the messenger factor). The 3x3 factorial design thus had 9 treatment cells. Each cell had about 250 households (although different predicted effect sizes used during power calculations for different comparisons resulted in slightly different samples for particular comparisons). In the present study, we consider only some comparisons.

In the first treatment (T1), the proportion in control group that gets to see a video featuring only a man is approximately equal to the proportion in the treatment group that gets to see a video featuring a couple; we do not use the subgroup that received videos that only featured a woman in this study (for the orthogonal factor we only used the level where both spouses got to see the video). As such, we simply compare two cells in the 3x3 cells design, resulting in approximately equal sample size in treatment and control.

In T2 in this study, we combine woman and man recipients into the control group (as these are all instances where only one spouse receives information) and compare to the treatment group where both spouses were shown the video (for the orthogonal factor we now use only the level where a couple features in the video). For the control group in this treatment we thus pool two cells of the 3x3 design and compare it to a single cell. As such, here the control group is approximately twice the size of the treatment group.

While statistical power is optimal in cases where there are approximately similar numbers of observations in treatment and control groups, more observations are always better. Furthermore, different numbers of observations between treatment and control groups does not lead to bias in the estimates.

In the text, we adapted footnote 7 which read:

> *We made sure that in about half of the cases the video was shown to the male co-head alone and in the other half of cases to the female co-head alone.*

To

> *The reason why we have about double the number of households in the control group is due to the design of the experiment. The experiment was set up as a 3x3 factorial design to also test various other hypotheses (unrelated to survey response disagreement). For the second treatment we effectively pool two treatment cells to construct the control group: one treatment cell where only the male co-head was shown the video and one treatment cell where only the female co-head was shown the video. While statistical power is optimal in approximately equal sized samples in treatment and control, more observations increase precision.*

• Last paragraph of conclusion is weak.

Agreed, we have removed it as it did not contribute anything.

Page 1 …"decision-making processes…" Very few of the papers cite actually dive into processes of decision-making (e.g., negotiation, avoidance, etc.) I suggest rewording this.

Agreed. We have changed this to the more general "intrahousehold dynamics".

Page 2 The first sentence in the in section 2 is problematic.

The first sentence in Section 2 reads:

> *Our review of prior work on spousal disagreement focuses mainly on the empirical literature on intrahousehold resource allocation and decision-making in (dual-headed)*

agricultural households in low- and middle-income countries, and draws primarily on analyses that use observational data from surveys in which both male and female co-heads are interviewed separately and are asked the same set of questions.

The purpose of this sentence is to delineate our focus in a vast literature. We do not immediately see why this sentence is problematic. Perhaps the reviewer was referring to another first sentence in this section, however, we also did not find any of the other sentences sound problematic.

Page 4 …focusingon…

Corrected

Page 5 …cause by…

Corrected

Page 9 …can not…

Corrected

Page 10 The error term needs to be defined.

Done

Page 10 extra tab after the equation

We are not sure what is meant by the reviewer with 'extra tab after the equation'… The formatting for this paper is done in LaTex. This issue will be addressed at the proofing stage.

Page 16 …f…

Corrected

Page 17 17.1 percent in the text. This should be 5.5 percent (based on the table).

The text says it reduced to 17.1 (from 22.6 percent) which is a reduction of 5.5 percentage points (that is in the table).

Page 21 …monitoring.In…

Corrected

Page 21 …resourcesoften…

Corrected

Page 22 …findings sheds…

Corrected