Response to the Reviews: "How to reduce Item Nonresponse in Face-to-Face Surveys? A Review and Evidence from the European Social Survey"

July 13, 2023

Dear editors and reviewers of MDA,

Thank you for your reviews and I am very grateful for the opportunity to improve upon my work. And I am sorry for the long time it took me to address your criticisms. I have organized my response in sections in roughly chronological order of the article. I hope I have addressed your criticisms to your satisfaction. When I disagree with a comment of the reviewers or am not able to incorporate a requested change, I provide my reasons here as well.

1 Language and Structure

> "All reviewers commented on the language and one very helpful reviewer even added an annotated file with suggestions for language changes in your rewrite. We have added this file. Also, it was suggested to ask a native speaker to go over the final manuscript. Several reviewers commented that the manuscript was unnecessarily complicated and long and suggested that you restructure it. Finally, it was suggested to incorporate the important Tables directly in the manuscript and not in a supplement. When rewriting the manuscript, we ask you to focus on this and take the detailed comments of the reviewers into account." (Editor letter, p. 1)

I have thoroughly rewritten all sections of the article and checked grammar and spelling throughout. I hope that the improvements in language are

to your satisfaction.

Regarding length, I am not sure where to cut it down. And the article is still within the word limits of a research note. As large parts of the article are literature reviews, it indeed might feel slow but that is inherent to literature reviews. If you still consider the manuscript too long, I would ask you to state what can be omitted and why.

Similarly, the general structure going from theory to strategies to empirical analysis is stringent to me so I do not know what you expect me to change here. If you still consider changes in structure necessary, please give me more details. However, I restructured the section on the tested strategies. This section only presents the selected strategies and why I expect them to reduce INR. Operationalization and the selection of controls have been moved to the data and methods section. And some reviewers have requested to connect the different parts more, especially when switching from review to analysis. I tried to smooth out the transitions between sections and make the article more coherent.

The regression table is now placed within the text, the exact design would be part of the layout though.

2 Conceptualisation of INR

In my article, I conceptualized four constructs related to the probability of item nonresponse, cognitive ability, task difficulty, motivation, and privacy concerns. Reviewer D writes:

> "I do not agree with this conceptualization, as it makes identifying the source difficult. Rather, it would be helpful to distinguish sources of item NR overall or within each construct the author proposes. For example, one can conceptualize respondents, questions, questionnaires, interview protocols/settings, and interviewers as sources of the item NR [...]. For example, motivation can be a product of all sources. Lumping these sources into the constructs is likely to make coming up with strategies for reducing item NR extremely challenging."

I disagree with this notion. Providing a nonresponse to an item is always an action of the respondent. And a specific respondent (with its characteristics) is asked to do a task in a specific situation. This task might be more or less difficult or boring etc. for this respondent in this situation. Survey design can influence the task and situation, but not directly item nonresponse, only indirectly. I think that it is necessary to conceptualize item nonresponse in

terms of the respondent. Questionnaire design etc. are not "sources" of INR but influence the sources. They are not acting entities and placing ontological meaning onto them misrepresents the survey response process.

Reviewer D again writes in a similar vein: "Privacy concern is hypothesized to come in in the editing process, but I argue it may also come during the comprehension stage.". Again, I have to disagree with this notion. Concerns about one's privacy are not a matter of comprehension. They are right that such a concern can become salient to the respondent before coming up with an answer. But Tourangeau et al. (2000) explicitly allow that although the steps of responding are typically in that order, they do not need to be. Editing may therefore kick in right after comprehension.

Additionally, this conception is based on the conceptions of Krosnick (1991) and Tourangeau, Rips, and Rasinski (2000) which are widely cited in this literature. Such objections are therefore going far beyond what this paper aims for and is capable of as they would require a different conceptualization of survey response entirely.

Since only one reviewer had such fundamental objections and the editors decided to allow me a revision even though such a fundamental part of the paper is objected to, I will not change this conceptualization. I have tried to make the position that the respondent is the acting entity clearer in section 2 though.

3 Extension of the Satisficing Model

Reviewer D claims that I fail to deliver the extension of the satisficing model promised in the abstract. This is somewhat contrasted by Reviewer E though: "Your novelty is the introduction of 'privacy concerns' in the model developed by Krosnick, so you should give more visibility to it. Therefore, my first suggestion is to make more evident the novelty you are bringing to the field."

So, Reviewer E thinks that I do provide a meaningful extension of Krosnick's satisficing model. Maybe the reason Reviewer D believes that I did not is indeed due to the lacking visibility.

My aim with this extension (and the formalization) is to synthesize and summarize results from the literature on INR. I believe that without synthesis and theoretical models, we can only achieve a piece-meal understanding of such processes and result in ad-hoc explanations or survey designs.

This holds for the formula as well. It is not intended to be an actual formula providing a point estimate of the probability of item nonresponse given some measured inputs (although it would of course be better). In terms of the aim of the formula, I again follow Krosnick. It serves as a

summary of what determines INR and how these concepts interrelate. The inputs, therefore, need not be measurable on a proper scale (as commented on by Reviewers D and F). I made this fact explicit in the article and changed the formula so that it reads "the probability of harmful item nonresponse is a function of..." instead of claiming to have a full formula of the probability.

Reviewer D also asked why I chose a max function in the formula. The reason is that I conceptualize privacy concerns as being independent of the other concepts. The probability will then either be driven by the relations of task difficulty, cognitive ability, and motivation OR privacy concerns depending on which of those exceeds the other. Reviewer E also requested to better explain the aim and meaning of the formula so I tried to make it clearer in section 2.

Reviewers D and F also point out that the formula is not picked up in the empirical analysis. This is a result of a shift from literature review and theoretical synthesis first to practical applications in survey design following from that model later. Reviewer F mentions that the link between those two parts is indeed weak. I tried to point out the link between the theoretical model and empirical analysis accordingly (introduction to section 4 and section 5.3).

4 Literature Review

> Lines 193-195: by 'changes in response categories' do you mean 'changes in response scales'? Routing and filtering are associated with higher non-response in PAPI, using CAPI mode may reduce routing and filtering problems to zero. (Reviewer E)

I indeed mean changes in response scales. I changed the wording. And I added that this result is likely different between modes.

> " Regarding ability, there is much more to say. The environment of the interview is important, but ability can (and must) be assessed before going into the field by a pilot study, in order to test the comprehension of the questions, the rates of item non-response, the impact of the order of the questions (contamination), among other aspect. (Reviewer E)

I would categorize your comments not under ability but task difficulty as they relate to survey design and not respondent characteristics or the interview situation. But you are right that review rounds and pilots are important tools to ensure survey quality also in this regard and I added a paragraph on them in section 3.1.

> "Matching education has no effect (Vercruyssen, Wuyts and Loosveldt, 2017;..." not quite...what the authors say is that "the effects of education level (mis)matching could unfortunately not be tested." What had no effect was the education level of the interviewer." (Reviewer E)

Thank you for the correction, I removed the citation (Vercruyssen, Wuyts and Loosveldt, 2017) but Silber et al. do test it and find no effect.

5 Don't Know as a substantive Answer

> "the author acknowledges that 'don't know answers' for some respondents actually be the most applicable answers (i.e. "because they do not know about the content of the question or are unable to remember an event"), initially this is not mentioned, and also when discussing the four constructs related to the probability of item nonresponse, this issue is ignored. I think the author could put more emphasis on the difference between satisficing and privacy-originated sources of item nonresponse on the one hand, and substantive don't know answers on the other. [...] I do not agree with the author's argumentation that in cases where DKs are genuine, this is "a product of the respondents' ability", since especially in attitude items, it may often happen that before the survey respondents have never thought about the issue being questioned." (Reviewer A)

Reviewer A makes a good point here. The reason why I do not consider DK as a genuine answer is simply because genuine answers are unproblematic. This paper aims to improve survey quality and therefore seeks to reduce *problematic* INR. However, I have clarified why I do not consider genuine DKs in detail and removed the argument that they are due to respondents' ability. Accordingly, the definition of item-nonresponse as ingenuine answers has been stressed.

For the analysis, I do not expect that the presence of genuine DKs in the dependent variable would bias some of the estimates. At least, no such mechanism came to my mind. I, therefore, did not change the analysis based on this comment. Genuine DKs only increase variation in the dependent variables but do not bias the results, as mentioned in section 5.2.

6 Data and Variables

Reviewers C and D asked for the data to be introduced more thoroughly and especially to add information on protocols and contexts relevant to INR. I added some fundamental information about the ESS in section 5.1 but I still wanted to keep it short. I believe the ESS is a very widely known data set and I have seen in other research notes in MDA that data descriptions are short even with lesser-known data. I added a reference for further information though and more detailed information on the data has been added to the supplementary material.

The latter point of Reviewer D to provide more information on protocols and procedures relevant to the present study is very important though. I have added information about the policies regarding the use of showcards, the intended length of the interview, and interviewer behavior in dealing with respondents and especially item nonresponse in sections 5.1 to 5.3.

Reviewer D also asked why I restrict respondent characteristics to education and age. Since my analysis shifts from the theoretical model to strategies for reducing INR on the survey level, I only need to include respondent characteristics if they threaten to confound my analysis of strategies. I use both of them as proxies for cognitive ability but I do not need further respondent characteristics as controls. In section 5.3, I have stressed this point again. No other reviewer complained about the selection of controls so I have not changed them.

Reviewer E pointed out that matching respondent and interviewer characteristics in terms of ethnic/national origin could be important. That is true and I added this to the literature review of strategies. But since ethnic/national origin is not measured of interviewers, I cannot assess its effect in my analysis, unfortunately.

7 Statistical Analysis

> "it is important to note that attitude items often have higher levels of item non-response than factual items on the same topic, reflecting opinion censoring, which is actually higher among higher educated respondents than lower educated respondents. Therefore, I suggest that the type of item, attitudinal or factual, is also taken into account in the analysis of the ESS data." (Reviewer A)

Reviewer A makes an important point here and I have added this to the review. However, the empirical analysis aims to test strategies to reduce

item nonresponse at the survey level, not to understand influences on the item level. It might indeed be the case that these strategies would have a different effect depending on the type of question (attitudinal or factual). But to disentangle this, I would need to conduct separate analyses on different subsets of the items. Given that multiple reviewers already called the manuscript too long, I would not extend the analysis. This is an important shortcoming of the paper though and I added this to the discussion. If strategies differ in their effects depending on the type of question, then the external validity beyond the mix of question types as in the ESS could be limited.

> Unless reading the supplemental materials, readers will not know why negative binomial model is used. Also, why interviewer effects are analyzed as fixed? "Standard errors are clustered by interviewer." on p.15 does not make sense. (Reviewer D)

I added a short reason for the chosen model in section 5.4. I use interviewer fixed effects because I want to get rid of variation in levels of INR between interviewers since this part of the variation is not relevant to the tested strategies. Standard errors are typically clustered by the variable denoting the fixed effects to prevent heteroscedasticity within the clusters. As now pointed out in the supplementary material, this is common practice and even the default in the R package used. I compared the clustered standard errors to bootstrapped standard errors and found no substantive differences. I extended the discussion on the fixed effects and standard errors in the supplementary material and highlighted the supplementary material in the context of model selection as a compromise to the comment of Reviewer C to discuss model selection in the text and the need to keep the article brief.

Reviewer D asked about missing data in my analysis pointing out that I have a large amount of missing data. As mentioned in the supplementary material, there are two major sources of missing data in my analysis. The first one is the age of the interviewer is often missing and Romania used the wrong format for the interviewer's age altogether. I resultingly had to completely exclude Romania. But this would only be consequential if the nonresponse in the age of the interviewer is related to the interaction between respondent and interviewer since I control for interviewer effects using the fixed effects. The second source of reduced observations is a lack of variation within interviewers in the fixed effects regression. Since a within-estimator needs within-variance, interviewers without within-variance are excluded from the analysis. This is particularly often the case for the number of refusals since the number of refusals is relatively low. Although such a large loss of observations is indeed

unfortunate, they can be assumed to be random and therefore not bias the results. The levels of item nonresponse in the variables used for analysis is always reported with descriptive statistics in the supplement and are not particularly high. The largest part of the lost observations can therefore be assumed to be missing at random and does not bias the results. This discussion is extended in the supplementary material as well.

8 Length of Interview as a Covariate

> "Since length of the interview may vary depending on the applicability of questions, I am not sure what the mechanism here is: respondents for whom more questions are applicable will have longer duration and are different from those who have fewer applicable questions due to filter questions [...]. Since interview length cannot be disentangled from respondent characteristics, it is completely unclear what is causing the positive effects of interview length on item nonresponse." (Reviewer A)

As discussed in Section 4, there are two reasons for longer interviews. The first reason is that the respondent takes longer to answer which is a result of the respondent's cognitive ability. The second reason is that a higher number of questions asked will lead to the respondent losing interest and getting tired. The strategy to be tested is that a shorter questionnaire prevents INR via the second mechanism. But to be able to assess this, I need to control for cognitive ability.

However, based on this comment by Reviewer A, I have realized that I have a better measure for questionnaire length than time duration, which is the number of items not applicable to the respondent. I have therefore exchanged the duration with the number of items. This variable should have a less direct connection to cognitive ability. The comments on interview duration by Reviewer E resultantly do not apply anymore.

The quoted comment by Reviewer A claims that the people with a different number of items might be substantially different from those with many items not applicable. But to bias the results, they would need to differ concerning motivation, cognitive ability, or privacy concerns following the presented theoretical model. I was not able to think of a mechanism for why that would be the case in general, therefore it would depend on the specific questions that would have been asked. I checked descriptively whether there is some form of a relation between the number of applicable items and education (as a proxy for ability) and found no strong bivariate relation (see the

supplementary material). Unfortunately, I do not have proxies for motivation and privacy concerns to explicitly check them but I think that a spurious correlation between ability and the number of applicable items would have been the most likely case. I would therefore argue that the number of applicable items is a reasonable test of the mechanism but caution is of course advised.

However, the results indicate that the number of applicable items has a negative effect on INR and this result is robust across model specifications that also include duration. I, therefore, need to retract my claim from the previous version that longer questionnaires increase the number of INR. Surprisingly, I find a negative association between INR and longer questionnaires and I do not know how to interpret this result myself. I added a short disclaimer to the discussion section. One might argue though that it is in fact the time they are required to concentrate that is more important than the number of questions. But this would again run into the problems outlined above. If you have any recommendations or ideas on this issue, please let me know.

9 Miscellaneous

> Be careful with contradictions. For instance, in lines 49-51 (and 368-3689) you write: "Matching respondents' and interviewers' gender and age seem not to be efficient tools to reduce item nonresponse."; but in lines 234-236 it is said that "Vercruyssen, Wuyts and Loosveldt (2017) find less item nonresponse when matching age of interviewers and respondents and matching gender reduces item nonresponse for males and increases it for females"; and in lines 276-277, "Previous research suggests that respondents might be more willing to answer to socially similar interviewers." (Reviewer E)

Those are not contradictions but differing results from different studies. Some scholars have hypothesized this link and could show it empirically but in my analysis, this did not seem to matter. The conflicting statements report an inconclusive state of research.

> "The motivation for this particular study is not clearly provided in Introduction. Rather, Introduction mixes in the findings from the study." (Reviewer D)

I tried to stress the contribution in the Introduction. However, summarising the general findings in the introduction is quite common in academic

papers. If requested, this can be removed but I do not see the general issue with that practice.

> "The discussion section would gain if the novelty of this study was stressed Finally: the tendency of many surveys (due to high costs and low response rates) is to change from face-to- face to self-administrated modes (PAPI and CAWI). Consider discussing a bit this incoming shift and how your final recommendations could be adapted." (Reviewer E)

Thank you for these suggestions, I have tried to satisfy your requests.