Intra-respondent reliability and enumerator quality in field conjoint experiments*

Matthew K. Ribar†

January, 2025

Abstract

Conjoint experiments ask respondents to consider multiple 'treatments' simultaneously, leading respondents to make errors in their responses. Field conjoint experiments, which are often administered by teams of enumerators, introduce another error point because enumerators may be poor quality. However, calculating enumerator intra-respondent reliability (IRR) at the enumerator level provides researchers a tool to monitor enumerator performance specifically for the conjoint experiment. Other common proxies for enumerator quality do not correlate with IRR, which stabilizes at the enumerator level after as few as 15 completed surveys.

Word Count: 1,699

*This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1656518. The Senegalese data collection for this project was an output of the 'Structural Transformation and Economic Growth' (STEG), a programme funded by the Foreign, Commonwealth & Development Office (FCDO), contract reference STEG_LOA_1266_Ribar. Fieldwork in Cote d'Ivoire was supported This work was supported by the Stanford King Center on Global Development, the Stanford Institute for Economic Policy Research, the Institute for Humane Studies, and a SurveyCTO data collection research grant. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. The Stanford University IRB approved the human subjects portion of this research under protocols number IRB-68012 and IRB-72215.

†Ph.D. Candidate, Department of Political Science, Stanford University, mkribar@stanford.edu

Researchers increasingly use field conjoint experiments to reach populations in developing countries. Garbe et al. (2024) use a field conjoint experiment to study willingness to register for a biometric ID program in Kenya. Baldin, Kao, and Lust (2023) use a field conjoint across Kenya, Malawi, and Zambia to identify when constituents are willing to entertain requests by formal or informal leaders. Zhou (2024) explores public preferences over territorial disputes using a field conjoint in India which asks opinions about Kashmir. Auerbach and Thachil (2018) use a field conjoint to study clientelism and brokers in Indian slums.

In these contexts, enumerators must go to respondents to administer surveys, often using tablet computers.[1] In contrast to online conjoint experiments which use software like Qualtrics to engage directly with respondents, these tablet-based interviews require an enumerator to administer them. Adding an enumerator into the survey flow opens a space for enumeration errors to degrade survey quality (Adida et al. 2016; Di Maio and Fiala 2020).

Researchers implementing field conjoint experiments should include enumerator-specific average IRR in their regular quality monitoring to proactively diagnose problematic enumerators. I build on Clayton et al. (2023), who propose a statistical correction to account for switching error in conjoint experiments and introduce intra-respondent reliability (IRR) to calculate it. Conjoint experiments often have multiple rounds; by adding an additional round which repeats or inverts a previous round, researchers can calculate the likelihood that provide the same answer across identical conjoint rounds. Averaging this statistic at the enumerator level yields an enumerator-specific measure of conjoint reliability.

I illustrate the use of IRR as a tool to monitor enumerators through two related conjoint experiments I ran in Sénégal in 2023 and Côte d'Ivoire 2024. I presented respondents with two parties to a hypothetical land dispute and asked the respondent which profile was more likely to win. Profiles varied across the party's sex, the value of the party's land, the party's group (farmer or herder in Sénégal, autochthone or allochthone in Côte d'Ivoire), whether the party had given the traditional chief a gift, and whether

[1] Ferree et al. (2023) embed a conjoint experiment into a phone survey in Malawi, showing that enumerator considerations in conjoint experiments are not exclusive to tablet-based field surveys.

the part possessed a written title for their land.[2]

1 Intra-Respondent Reliability

The canonical conjoint experiment design present respondents with two profiles, side-by-side. Each profile has a set of attributes, which randomly vary between different levels. Respondents are then asked to choose one of the two profiles. Common questions include "for which of these two candidates would you vote" or "which of these two policies do you prefer?" Each pair of profiles comprises one conjoint round. Researchers often repeat rounds to increase statistical power. Bansak et al. 2018 show that even surprisingly large numbers of conjoint rounds do not degrade conjoint quality or induce survey satisficing.

The conjoint design allows respondents to evaluate profiles across multiple attributes, but it also violates several common principles of survey design. Respondents are asked to hold large amounts of information in their head, and all profiles combine multiple 'treatments' arms.[3] Because of these quirks of survey design, respondents in conjoint designs are particularly vulnerable to 'switching error:' respondents may unwittingly choose profiles that do not correspond to their true preferences (Clayton et al. 2023).[4] To overcome this problem, Clayton et al. (2023) introduce a statistical method to correct this source of bias for the two most common estimands for conjoint experiments: average marginal component effects (AMCEs) and marginal means.[5] These corrections rely on a a quantity they call intra-respondent reliability (IRR): the likelihood a respondent makes an identical selection when faced with identical tasks.

The preferred manner in which to capture IRR is to repeat the same conjoint task at the beginning of the set of conjoint experiments and at the end. In my running example, respondents were presented with six total paired conjoint tasks, the sixth being an

[2] See Ribar 2023 for a full description of the experiments.

[3] As a result, conjoint experiments are a sub-genre of factorial experiments (Hainmueller, Hopkins, and Yamamoto 2014.)

[4] Another concern is that respondents will anchor themselves to the first or last attribute, which is why researchers commonly randomize the order in which attributes are presented.

[5] See Clayton et al. (2023: 11–2) for the formulae, which I omit here for brevity.

inverted copy of the first. We can then calculate IRR as the fraction of respondents who agreed with themselves across the two tasks.[6] An IRR of 0.5 indicates that respondents agree with themselves 50 percent of the time—as randomly as if they flipped a coin. An IRR of one would indicate that respondents always provide identical answers to identical conjoint pairs. Online conjoint experiments average an IRR of about 0.77 (Clayton et al. 2023: 13); the in-person conjoint experiments in Sénégal and Côte d'Ivoire I explore below averaged an IRR of about 0.84.

2 Using Intra-Respondent Reliability to Monitor Enumerators

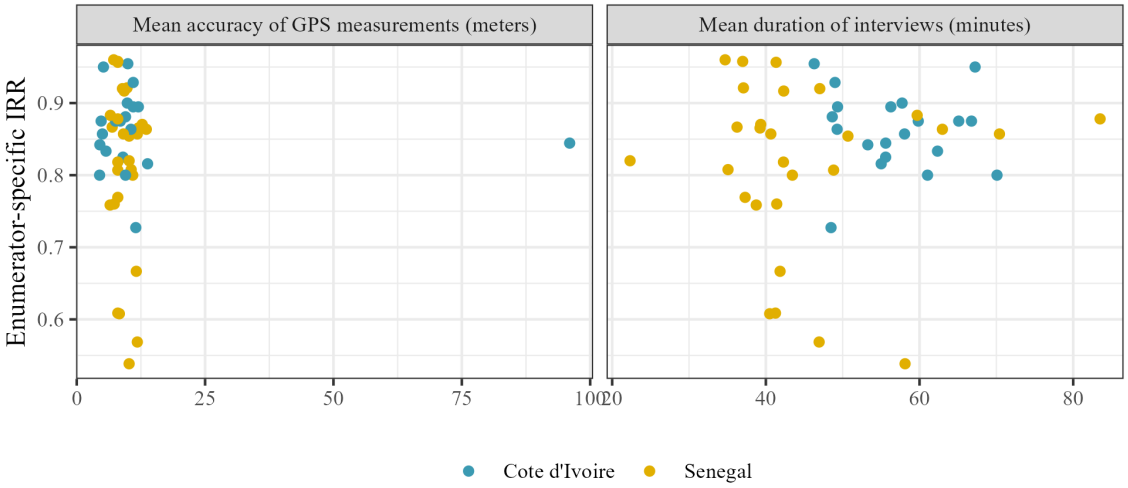
IRR is an excellent tool to monitor enumerator quality, beyond its role in correcting for measurement error. Monitoring the average IRR per enumerator allows survey researchers conducting high frequency data checks during data collection to identify which enumerators are effectively delivering the conjoint experiment's content.

Typically, researchers monitor enumerators and data quality through high-frequency data checks which produce a variety of diagnostic statistics to assess the data quality and identify problems. Many researchers use such data checks to detect enumerator shirking. To that end, most survey researchers monitor the duration of surveys (or the duration of specific modules within surveys) as part of these high-speed checks. Another common check is to use GPS coordinates to ensure enumerators are administering the surveys in the correct locations or following a sampling plan. While it is difficult to automate GPS based checks, many researchers also require a minimum accuracy threshold for any GPS measurements.

Unlike these tools to detect shirking, the IRR directly monitors the outcomes of the conjoint experiment. Figure 1 demonstrates that enumerators who do not rush through the survey and who take accurate GPS coordinates may nevertheless do a poor job administering the conjoint. Both mean duration and mean GPS accuracy are poorly correlated

[6] Across 1,965 unique respondents and 48 unique enumerators in the example below, zero identified that the first and sixth conjoint tasks were inverses of each other.

Figure 1. Average survey duration and enumerator-specific IRR are poorly correlated

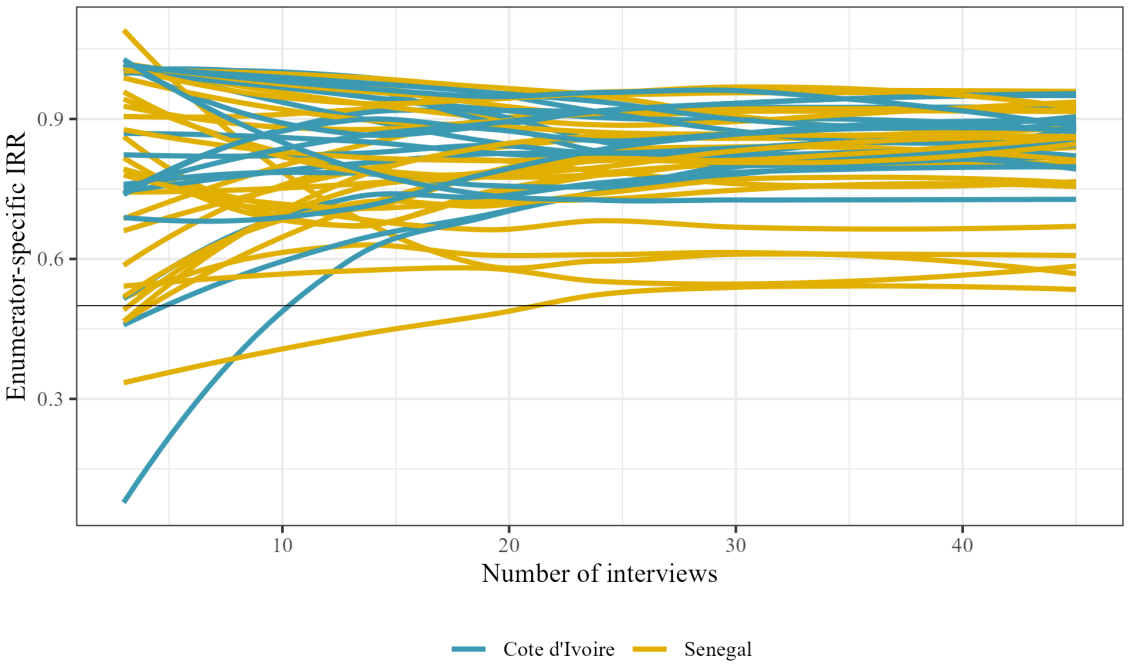


with IRR.

One concern with using IRR for monitoring enumerators is that the measurement error captured by IRR is necessarily co-produced. Enumerators may shirk, but respondents may be inattentive. Either of the two could contribute to poor IRR, and we obtain only one IRR measurement per respondent. Any given enumerator could encounter a series of inattentive respondents and therefore have comparatively low IRR. However, figure 2 shows that enumerator-specific IRRs smooth over time and change only minimally after about the 15th administered survey. This figure also show that a number of clear outliers emerge—while some enumerators have final IRR ratings well-above 0.9, a number are barely above 0.5, an IRR that implies respondents are answering as good as randomly.

Anecdotes from the implementation of the two field conjoint experiments reinforce the utility of adding enumerator-specific IRR to high frequency data checks. In Côte d'Ivoire, I assigned extra field supervision to the enumerator with the lowest IRR. The supervisor discovered that—contra the survey training—the enumerator was reading conjoint profiles aloud to the respondent rather than showing the tablet screen, on which the conjoint profiles were displayed with imaged. Respondents surveyed by this enumer-

Figure 2. Enumerator-specific IRR smooths for enumerators over time



ator were holding much greater amounts of information in their head, and as a result were making inconsistent selections. Alternative strategies to monitor enumerators would not have detected this anomalous behavior, but checking enumerator IRR allowed me to correct the enumerator's behavior.

Enumerators for both conjoint experiments presented respondents with a printed guide for the different attribute levels, to help respondents keep track of the different information. In Sénégal, examining enumerator-specific IRR revealed a handful of enumerators with low IRRs. After assigning additional supervision to these enumerators, the supervisors reported that enumerators were not using this printed guide. The supervisors corrected this error, and IRR improved. In both cases, including enumerator-specific IRR in regular data checks allowed me to identify errors which would not have been detected using other methods.

Beyond enumerator monitoring, respondent-specific IRR could also be used for post-hoc corrections, such as dropping observations from enumerators with an IRR be-

low some pre-specified (and presumably pre-registered threshold). For example, a pre-registration plan could specify that data from any enumerator whose final IRR is below 0.6 will be dropped. In the case of the two conjoint experiments in Sénégal and Côte d'Ivoire, dropping poor performing enumerators does not affect the statistical significance of my results.[7] However, in these surveys, enumerators with poor IRRs were targeted for increased supervision and coaching. In surveys where IRR was not monitored, dropping poor-performing enumerators may have a greater effect.

3 Conclusion

This letter advocates for using enumerator-specific IRR as a tool to monitor enumerator quality in field conjoint experiments. This extension of Clayton et al. (2023)'s results will be relevant to any survey researchers who implement conjoint experiments in contexts which require enumerators to administer the survey. It also builds on previous literature which shows how enumerator-specific attributes can bias survey results (Adida et al. 2016; Di Maio and Fiala 2020).

References

Adida, C. L., Ferree, K. E., Posner, D. N., and Robinson, A. L. (2016), "Who's Asking? Interviewer Coethnicity Effects in African Survey Data", *Comparative Political Studies*, 49 (12): 1630–60.

Auerbach, A. M., and Thachil, T. (2018), "How Clients Select Brokers: Competition and Choice in India's Slums", *American Political Science Review*, 112 (4) (): 775–91.

Baldwin, K., Kao, K., and Lust, E. (2023), "Is authority fungible? Legitimacy, domain congruence, and the limits of power in Africa", *American Journal of Political Science*, First view: 1–16.

[7] Specifically, I test the difference in marginal means between two subgroups: the lowest quartile of trust in formal institutions and the highest quartile of trust in formal institutions. Dropping 1-3 enumerators does not affect this t-statistic; dropping more reduces it due to the decrease in sample size.

Bansak, K., Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2018), "The Number of Choice Tasks and Survey Satisficing in Conjoint Experiments", *Political Analysis*, 26 (1) (): 112–9.

Clayton, K., Horiuchi, Y., Kaufman, A., King, G., and Komisarchik, M. (2023), "Correcting Measurement Error Bias in Conjoint Survey Experiments", Working Paper.

Di Maio, M., and Fiala, N. (2020), "Be Wary of Those Who Ask: A Randomized Experiment on the Size and Determinants of the Enumerator Effect", *The World Bank Economic Review*, 34 (3): 654–69.

Ferree, K. E., Honig, L., Lust, E., and Phillips, M. L. (2023), "Land and Legibility: When Do Citizens Expect Secure Property Rights in Weak States?", *American Political Science Review*, 117 (1): 42–58.

Garbe, L., McMurry, N., Scacco, A., and Zhang, K. (2024), "Who Wants to be Legible? Digitalization and Intergroup Inequality in Kenya", *Comparative Political Studies* (), Publisher: SAGE Publications Inc: 00104140241276971.

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2014), "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments", *Political Analysis*, 22 (1), Publisher: Cambridge University Press: 1–30.

Ribar, M. K. (2023), "Who wants property rights? Conjoint evidence from Senegal".

Zhou, A. (2024), "The Cost of Compromise: Territorial Disputes and Domestic Public Opinion".