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

Matthew K. Ribar†

September, 2025

Abstract

Conjoint experiments ask respondents to consider multiple 'treatments' simultaneously, leading respondents to make errors in their responses. Field conjoint experiments introduce another error point because enumerators may make mistakes in administering the conjoint. Calculating intra-respondent reliability (IRR) at the enumerator level provides researchers a tool to monitor enumerator performance. Other common proxies for enumerator quality do not correlate with IRR, which stabilizes after as few as 15 completed surveys. Using related conjoint experiments in Côte d'Ivoire and Sénégal, I show that respondent characteristics fail to predict differences in switching rates between enumerators. Researchers should include IRR in their regular quality monitoring to proactively diagnose problematic enumerators.

Word Count: 2,577 (including references)

Key words: Conjoint experiments, survey methods, survey design

*This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1656518. 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 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 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. The Stanford University IRB approved the human subjects portion of this research under protocols number IRB-68012 and IRB-72215.

†Postdoctoral Fellow, Department of Political Science and the Weidenbaum Center on the Economy, Government, and Public Policy, Washington University in St. Louis mribar@wustl.edu

Researchers increasingly use field conjoint experiments to reach populations in developing countries. Garbe et al. (2024) use such an experiment to study willingness to register for a biometric ID program in Kenya. Baldin, Kao, and Lust (2025) 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 with a conjoint experiment in India which asks opinions about Kashmir. Auerbach and Thachil (2018) study clientelism and brokers in Indian slums. In such surveys, research teams go to respondents and administer the survey in person, 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 who rely on enumerators to implement conjoint experiments should include enumerator-specific averages of intra-respondent reliability (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 IRR to calculate it. Switching error occurs when respondents provide an answer to the conjoint experiment that does not reflect their true preferences. Conjoint experiments often have multiple rounds; by adding an additional round which repeats or inverts a previous round, researchers can calculate the likelihood that respondents provide the same answer to identical conjoint rounds. Averaging this statistic at the enumerator level captures the fidelity with which the enumerators implement the conjoint.

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 ethnicity (farmer or herder in Sénégal, autochthone or

[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.

allochthone in Côte d'Ivoire), whether the party had given the traditional chief a gift, and whether the party possessed a written title for their land.[2] Using these data, I first use respondent's demographic information to predict their likelihood of switching (i.e. the probability that their response to the same question differs across rounds). I then show that enumerators are balanced across these predicted propensities to switch. If respondent characteristics do not predict switching error, then enumerators must be responsible.

1 Intra-Respondent Reliability

The canonical conjoint experiment design presents 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 show multiple rounds to increase statistical power. Bansak et al. (2018) show that increasing the number of conjoint rounds does 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. Conjoint designs ask respondents to hold large amounts of information in their head and profiles combine multiple treatments arms.[3] These quirks of survey design make respondents vulnerable to 'switching error:' respondents may unwittingly choose profiles that do not correspond to their true preferences (Clayton et al. 2023).[4] This error can attenuate estimates and complicate sub-group analysis. 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

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

[3] 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.

means.[5] These corrections rely on 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, the survey presented respondents with six total paired conjoint tasks, the sixth being an 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

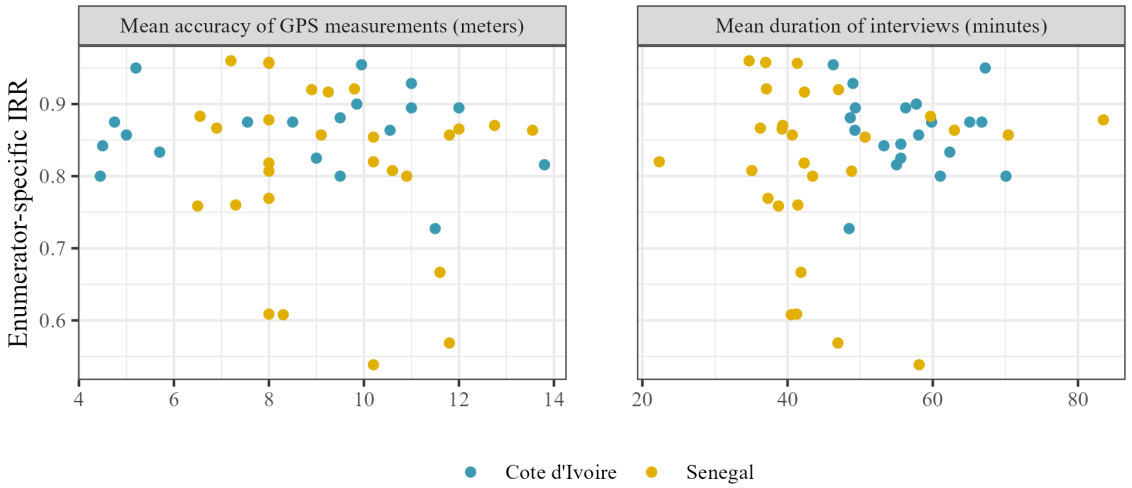
IRR is an excellent tool to monitor enumerator quality, beyond its role in correcting for measurement error. Monitoring the average IRR per enumerator allows survey researchers conducting high frequency data checks during data collection to identify which enumerators are effectively delivering the conjoint experiment's content and which enumerators are not.

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. Such data checks can identify enumerators who shirk or who otherwise do not collect high-quality data. To that end, many 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.

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

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

Figure 1. Proxies for enumerator shirking and enumerator-specific IRR are poorly correlated



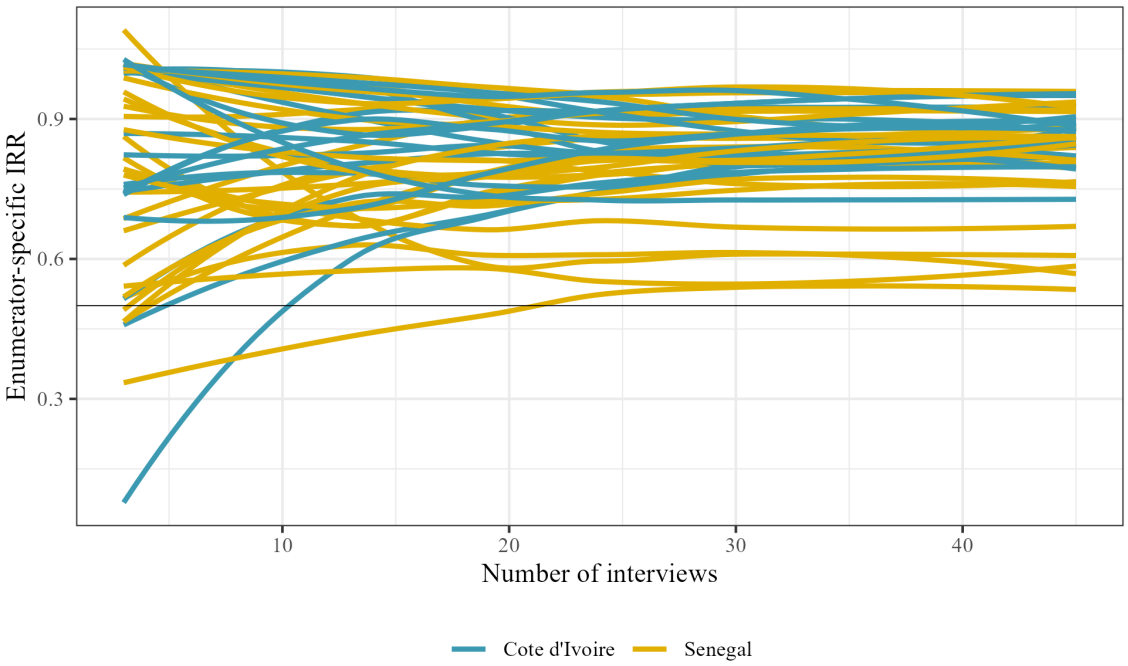
The vertical axis of this figure shows enumerator-specific IRR; the horizontal axes show the mean accuracy of GPS coordinates (high values mean imprecise measurements) and the average survey duration, both calculated at the enumerator level. The lefthand panel excludes an extreme outlier in GPS accuracy that does not affect the overall conclusion.

While it is difficult to automate GPS based checks, many researchers also require a minimum accuracy threshold for any GPS measurements. Lower accuracy show enumerators fail to adequately record their position.

These tools detect enumerator shirking in general, but adding enumerator IRR to regular quality checks directly monitors the fidelity and effectiveness with which enumerators implement the conjoint experiment. Because of the complexities of the conjoint design, it may be the case that enumerators successfully administer other parts of the survey but compromise data quality for the conjoint itself. Figure 1 demonstrates that enumerators who do not rush through the survey and who take accurate GPS coordinates, two common metrics by which researchers monitor enumerators, may nevertheless do a poor job administering the conjoint. Both mean duration and mean GPS accuracy are poorly correlated with IRR.

Any given enumerator could encounter a series of inattentive respondents and there-

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



The vertical axis of this figure shows enumerator-specific IRR; the horizontal axis shows the enumerators' count of completed interviews. While initial estimates of enumerator-specific IRR are noisy, they smooth after approximately 15 interviews.

fore 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 shows that a number of clear outliers emerge—while some enumerators have final IRR ratings well-above 0.9, some enumerators have IRRs barely above 0.5, an IRR that implies respondents are answering as good as randomly.

Anecdotes from the conjoint experiments in Sèègal and Côte d'Ivoire evidence the utility of adding enumerator-specific IRR to high frequency data checks. In Côte d'Ivoire, I assigned an extra supervisor to oversee 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, which displayed images of the conjoint profiles. Respondents surveyed by this enumerator were hold-

ing 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.

Likewise, in Sénégal, examining enumerator-specific IRR revealed a handful of enumerators with low IRRs. 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. 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, IRR could also be used for post-hoc corrections, such as dropping observations from enumerators with an IRR below some pre-specified (and presumably pre-registered) threshold. 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, I targeted enumerators with poor IRRs for increased supervision and coaching. In surveys where IRR was not monitored, dropping poor-performing enumerators may have a greater effect.

3 What predicts IRR?

One important question is whether the enumerators or the respondents drive this variation in IRR. Enumerators who are going an excellent job may nevertheless face inattentive respondents. Because enumerators are generally assigned geographically—one team implements the survey in region A, another in region B—respondent characteristics such as education or occupation could be systematically different across enumerators. If education predicts a respondent's IRR and average education of respondents differs

[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.

across enumerators, then enumerator IRRs would not be an accurate way to monitor enumerator performances.

To identify the extent of this problem, I use a logit model to predict the likelihood of respondent switching their responses.[8] The outcome is a binary indicator of switching (i.e. what is averaged to find enumerator-specific IRR). The predictors are the respondent's demographics: age group indicators, education, sex, and the respondent's relationship to the head of the household. I also include the country in which the survey took place. Figure 3 shows these results in three ways. The lefthand panel shows the distribution of the predicted propensities to switch; the center panel shows the distribution of switching itself; and the righthand panel shows the distribution of the absolute value of error (i.e the absolute value of the actual switching minus the estimated propensity to switch). For all statistics I first convert to enumerator averages and then to z-scores. A value to the right of the vertical line indicates a statistically significant outlier.

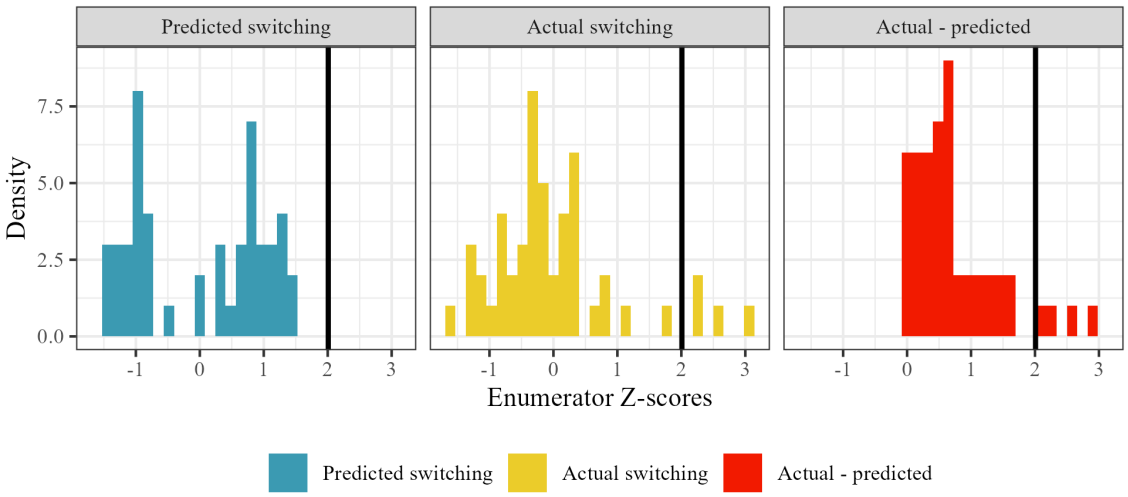
The lefthand panel of figure 3 shows that the predicted propensity of switching error is reasonably balanced among enumerators; there is no enumerator for whom the average predicted propensity to switch is more than two standard deviations away from the mean. On the other hand, the central panel shows four enumerators with high values of average switching error (the same four enumerators that figure 2 identifies). The righthand panel shows that the same enumerators have abnormally high values of prediction error: these outliers persist even when accounting for respondent's propensity to switch. In other words, the high enumerator-specific IRR is not predicted by respondent characteristics, but by the enumerator themselves.

4 Conclusion

This research note advocates for using enumerator-specific IRR as a tool to monitor enumerator quality in field conjoint experiments. IRR captures whether the respondent

[8] I also predicted switching error using a linear probability model (LPM) and a random forest, as well as weighted ensemble of the logit, the LPM, and the random forest. These alternatives produced no appreciable improvement on prediction errors (measured through root mean squared error) over the logit in out-of-sample predictions.

Figure 3. Predicted propensity to switch and actual switching by enumerators



All statistics are averaged at the enumerator level and transformed into z-scores. Predictions are from a logit regression of demographic variables on binary switching error at the respondent level. The vertical lines show critical values for a t-distribution with 48 degrees of freedom and $\alpha = 0.05$.

provides identical answers to identical questions posed at the start and end of the conjoint experiment. Averaging these values at the enumerator level can help researchers to monitor whether enumerators are correctly implementing the conjoint experiments. Respondent characteristics fail to meaningfully predict IRR, suggesting that enumerators drive deviations in IRR, rather than respondents. 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.

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. (2025), "Is authority fungible? Legitimacy, domain congruence, and the limits of power in Africa", *American Journal of Political Science*, 69 (1): 314–29.

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*, 58 (9): 1803–53.

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): 1–30.

Ribar, M. K. (2023), "Who wants property rights? Conjoint evidence from Senegal", CEPR Working Paper no. 072.

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