Methods Appendix to
Attrition in mobile phone panel surveys

This document provides supplementary information on an analysis of attrition in mobile phone panel surveys. Subject of the analysis was open data of the World Bank's Listening to Africa [1] initiative, a series of mobile phone panel surveys conducted in six sub-Saharan African countries. For both projects for which data was still publicly available in December 2020 we analyzed attrition, focusing on its variation over the two-year periods of data acquisition and potential biases introduced. This was done in R [2]. All scripts and results are available on Github.

Data preparation

As of December 2020, open data remained available for two of the Listening to Africa surveys: The Sauti za Wananchi (SzW) survey conducted between 2012 and 2014 in Tanzania [3] and the Listening to Senegal (L2S) survey, which was conducted between 2014 and 2017 [4]. The data from both projects each consist of a dataset from a face-to-face baseline survey and several datasets from computer-assisted telephone interviewing (CATI) survey rounds in which interviewers attempted to reach the respondents of the baseline survey on mobile phones previously distributed to all survey participants. This setup enables us to analyze demographic and occupational characteristics not only for the respondents, but also for non-respondents to the CATI rounds.

We merged the data of the individual CATI rounds to the baseline data by a universal household number (uhn). The data of all survey rounds of SzW that were available online (namely Rounds 1-8, as well as Round 10, 14 and 24) already contained an appropriate universal identifier variable. In the L2S data, household identifiers were missing in some survey rounds and not always of consistent format where present. Using the variables *census district*, *household number within the census district* and *household ID,* we were able to reconstruct an uhn that allowed us to merge most CATI rounds to the baseline. Only for the *Module Education*, the *Modul Electricité et Transport* and the first survey of the *Modul Alimentation et Sécurité Alimentaire* the creation of unique household identifiers failed. This was either due to lack of the necessary information (1st Round of *Modul Alimentation et Sécurité Alimentaire*), or the created identifiers were not unique (*Module Education*), or not all identifiers were contained in the baseline dataset (*Modul Electricité et Transport*). These datasets were therefore omitted from the analysis. Details about this process are described within the script 01b_create_IDs_and_metadata.R.

In a next step, we added a response status variable to each CATI round. The L2S datasets only contained respondents. Consequently, all observations contained in the datasets were classified as respondents, all others as non-respondents. The CATI rounds of SzW contained observations from respondents and non-respondents alike. Non-respondents were characterized by a larger number of missing values. We summed up the number of missing values for each survey participant and each CATI round. For each CATI round we then classified each observation either as respondent or as non-respondent depending on the

number of missing values. The threshold was determined by the number of occurrences of the maximum number of missing values and the sizes of the gaps between maximum and second largest value, and between second and third largest value. For details see 01c_determine_response_status.R. To ensure the algorithm worked as intended we visually validated the distinction for each round by checking the respective histogram of non-missing values and resolved unclear cases with help of the documentation and questionnaires. It should be noted that the datasets of round 7 (C*orruption*), round 10 (*Political Poll*) and round 24 (*Political Poll*) of SzW contained fewer observations than the other datasets and round 7 only contained respondents.

Subsequently, reserve households were excluded from the SzW sample, as well as a few observations with values that should – according to the baseline questionnaire – have led to replacement of the respondent. These last two steps were not necessary for the L2S data, since no such observations were found in the datasets.

The following packages were used in data management: *dplyr* [5], *haven* [6], *plyr* [7] and the *tidyverse* [8]. Additionally, the *renv* [9] and *here* [10] packages were used for dependency management and file referencing, respectively.

Response rates against time

Response rates were calculated by dividing the number of respondents of the respective CATI round by the baseline sample size. For L2S the response rates had been stated in the documentation. Up to the provided accuracy, these are equal to the response rates we calculated for all rounds but the second round of the *Modul Alimentation et Sécurité Alimentaire*. According to the documentation, this round has a response rate of 93% (we calculated 90.8%), which is equal to the response rate of the first round of this module, making a copy-paste error appear plausible.

The calculated response rates were plotted against the approximate time after baseline in months. For L2S the respective time information was extracted from the documentation, which provided start and end date of the data acquisition periods. We used the time difference between the start of the data acquisition of the respective CATI round and the start of baseline data acquisition, rounded to months to improve comparability to SzW. For the SzW survey such precise information was not provided. We therefore used the month and year contained in the file names. Since this information was missing for round 7 we set its time centrically between the months of round 6 and round 8.

This plot and all other graphs in this analysis were created using the *ggplot2* package [11] and labels were added with the *directlabels* package [12].

Composition Changes

Joint test of orthogonality of response status against a set of participant characteristics

For each CATI round, we tested if response status is correlated to a set of demographic and occupational variables: age and gender identity of the participants, whether they live in an urban area, whether they have completed secondary education and whether they work in agriculture. For this purpose, we estimated a multivariate linear OLS regression for each

CATI round with the response status as outcome variable and the set of demographic and occupational variables as input variables, using the *lm* function from the *stats* package [2]. We extracted the coefficients and heteroskedasticity-consistent standard errors using the *coeftest* function from the *lmtest* package [13] in combination with the *vcovHC* function from the *sandwich* package [14, 15] with the estimation type parameter *HC1*. For each CATI round we tested the null hypothesis that all coefficients are zero, using the *waldtest* function from the *lmtest* package [13] – again in combination with the *vcovHC* function from the *sandwich* package [14, 15] with the estimation type parameter *HC1*. Coefficients, standard errors and waldtest p-values of all CATI rounds are reported together with the baseline sample mean and standard deviation of the respective variables in an html table created with the *kableExtra* package [16]. The results for SzW and L2S, respectively, can be found in the 04_output folder on Github.

Tests of equivalence between survey rounds with respect to individual variables

We tested for all CATI rounds and all the variables mentioned above individually if the sample mean of the CATI respondents is equivalent to the baseline sample mean. This was done using a series of OLS linear regressions without intercept, regressing the respective demographic or occupational variable against a categorical variable of membership in either the group of respondents to the individual CATI rounds or the baseline sample. Heteroskedasticity-consistent standard errors and p-values of the equivalence test were again calculated using the *coeftest* function from the *lmtest* package [13] in combination with the *vcovHC* function from the *sandwich* package [14, 15] with the estimation type parameter *HC1*.

For all binary variables, the results are presented in form of a plot of the sample means against the survey round, with 95% confidence intervals as calculated from the heteroskedasticity-consistent standard errors. CATI sample means that significantly differ from the baseline average are marked with asterisks: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. These plots were generated using the *ggplot2* package [11]. Labels were added with the *geom_dl* function from the *directlabels* package [12]. These plots are available in the 04_output folder for SzW and L2S, respectively, together with the equivalent graphs for SzW and L2S with approximate time after the baseline survey as abscissa. The age distribution of the baseline sample and of the respondents of each CATI round were visualized as violin plots using the *geom_violinhalf* function from the *see* package [17]. These plots for SzW and L2S are available in the 04_output folder.

The script also generates a simple csv table of baseline means, mean differences for all CATI rounds and respective t-test p-values as output. The tables are provided for SzW and L2S in the 04_output folder on Github. They were generated using the *t.test* function from the *stats* package [2].

References

[1] The World Bank, „Listening to Africa,“ [Online]. Available: https://www.worldbank.org/en/programs/listening-to-africa.

[2] R Core Team, „R: A Language and Environment for Statistical Computing,“ Vienna, 2021.

[3] Twaweza, Sauti za Wananchi survey (Rounds 1-24, October 2012-September 2014), www.twaweza.org.

[4] Survey - Listening to Senegal, Year 2014-2017, National Agency for Statistics and Demography (ANSD) of the Republic of Senegal.

[5] H. Wickham, R. François, L. Henry und K. Müller, „dplyr: A Grammar of Data Manipulation,“ 2021.

[6] H. Wickham und E. Miller, „haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files,“ 2021.

[7] H. Wickham, „The Split-Apply-Combine Strategy for Data Analysis,“ *Journal of Statistical Software,* Bd. 40, p. 1–29, 2011.

[8] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo und H. Yutani, „Welcome to the tidyverse,“ *Journal of Open Source Software,* Bd. 4, p. 1686, 2019.

[9] K. Ushey, „renv: Project Environments,“ 2021.

[10] K. Müller, „here: A Simpler Way to Find Your Files,“ 2020.

[11] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, 2016.

[12] T. D. Hocking, „directlabels: Direct Labels for Multicolor Plots,“ 2021.

[13] A. Zeileis und T. Hothorn, „Diagnostic Checking in Regression Relationships,“ *R News,* Bd. 2, p. 7–10, 2002.

[14] A. Zeileis, S. Köll und N. Graham, „Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R,“ *Journal of Statistical Software,* Bd. 95, p. 1–36, 2020.

[15] A. Zeileis, „Econometric Computing with HC and HAC Covariance Matrix Estimators,“ *Journal of Statistical Software,* Bd. 11, p. 1–17, 2004.

[16] H. Zhu, „kableExtra: Construct Complex Table with 'kable' and Pipe Syntax,“ 2021.

[17] D. Lüdecke, M. S. Ben-Shachar, P. Waggoner und D. Makowski, „see: Visualisation Toolbox for 'easystats' and Extra Geoms, Themes and Color Palettes for 'ggplot2',“ *CRAN,* 2020.