CHAPTER 5

Language Differences Between Interviewers and Respondents in African Surveys

Charles Q. Lau, Stephanie Eckman, Luis Sevilla Kreysa, and Benjamin Piper

Introduction

In face-to-face surveys, the survey language has important implications for data quality. Linguistic issues are particularly relevant in Africa because of its linguistic diversity and complexity. Combined, there are over 2,000 African languages, more than 30 percent of the world's languages (Eberhard, Simons, & Fennig, 2019). Although there are some relatively linguistically homogeneous countries (e.g., predominantly Arabic-speaking countries in Northern Africa), most countries have a complex, multilingual structure. Many Africans are multilingual: 61 percent of Kenyan adults, for example, speak three or more languages (Logan, 2017). There are also different types of languages. People may grow up speaking the language of their tribe or local community but often learn languages of broader communication in school (for brevity, we refer to these languages as "local languages" and "broader languages"). These broader languages may be African (e.g., Swahili) or Western (e.g., English or French) and are often used in mass media, government communications, and workplaces (Bodomo, 1996). Although both local and broader languages are used to communicate, local languages tend to be used more for verbal communication, whereas broader languages are typically used for written communication.

The linguistic diversity in Africa presents several challenges for face-to-face surveys. Survey language can lead to undercoverage if the survey is not offered in a language the respondents speak (Andreenkova, 2018). Language also can shape the respondents' cultural and cognitive frames, affecting the response formation process (see Chapter 1). In this chapter, we focus on another challenge: problems that arise if respondents or data collectors are

not proficient in the survey language (Ahlmark et al., 2014; Pearson, Garvin, Ford, & Balluz, 2010; Peytcheva, 2008).

In the simplest situation, respondents and data collectors share the same first language (also known as "home language" in Africa) and conduct the interview in that language, as is the situation in many surveys around the world. In Africa, however, the linguistic situation is more complicated for four reasons:

- The linguistic diversity of Africa means it may not be feasible to translate surveys into all languages, primarily because of cost but also because of the difficulty in securing qualified translators and data collectors. If the survey is not offered in a respondent's home language, the respondent can either (1) not participate in the survey and become a nonrespondent or (2) participate using a language other than their home language (if available).
- For most surveys, data collectors work in multiple areas within a country with different languages, and many data collectors are multilingual. If data collectors work in a region where their home language is not spoken, they may need to use their second or third language to complete an interview (if available).
- Our experience observing fieldwork in Africa suggests that some people view participating in a survey as a "formal" activity more suited to a language of broader communication than a local language. As a result, respondents and data collectors may gravitate toward using a language of broader communication.
- Surveys sometimes use terminology that is more natural in a language of broader communication. The surveys we analyze in this chapter, for example, ask questions about democracy and political attitudes. Because the word "democracy" does not exist in most African languages, data collectors are trained to use a language of broader communication, not a local language, for these questions. For these reasons, even if a respondent and data collector share the same first language, they may opt out of that local language and choose to conduct the survey in a language of broader communication.

In sum, respondents and data collectors may opt to use a language other than their home or first language. Respondents may also engage in "code switching" (i.e., changing languages within the survey)—using a broader

language for complex questions and a local language for standard questions. Using a nonhome language in a survey, either because of choice or constraint, may have implications for data quality.

Given the limited literature on linguistic issues in African survey research, this chapter describes language patterns in face-to-face surveys in 36 African countries. Our goal is to provide a broad-brush, descriptive account that sets the stage for more complex analysis in future research. Our analysis extends the excellent descriptive work conducted by Logan (2017) on linguistic issues in the Afrobarometer project by pursuing three research goals:

1. Describe the languages used by respondents and data collectors in face-to-face surveys and develop a five-category taxonomy of language patterns.
2. Describe how the taxonomy functions in three countries from different regions and linguistic backgrounds (Cameroon, Kenya, and Mozambique).
3. Investigate which respondent characteristics (e.g., age, education, urban or rural location, gender) are associated with choosing different languages.

Data and Methods

Data

We analyze data from Afrobarometer Round 6, face-to-face, paper-and-pencil surveys conducted in 2014–2015 in 36 African countries. With surveys spanning two decades, the Afrobarometer initiative is the primary source of public opinion data on Africans' political attitudes, behaviors, and beliefs (see afrobarometer.org for more information). To produce comparable data, each country used a standardized sample design, questionnaire, and fieldwork procedures. The surveys were based on clustered, multistage area probability samples and used random walk procedures to select households. The sample design included stratification by geography (e.g., state, province) and urban–rural location. One individual was randomly selected in each household; to ensure adequate representation by gender, the random selection of respondents alternated between selecting men and women (Afrobarometer Network, 2014). Response rates varied by country, ranging from 30 percent (Tunisia) to 99 percent (Zambia; Isbell, 2017).

In Round 6, 53,935 interviews were completed with approximately 1,200 completed per country, except for Nigeria, Kenya, Uganda, South Africa, Ghana, Malawi, and Tanzania, which had approximately 2,400 interviews each. Of the 53,935 completed interviews, our analytic sample consisted of 53,596 cases; we excluded cases in which the respondent was younger than 18 or older than 120 ($n = 294$) and cases for which the variables on respondent or interviewer language were missing ($n = 45$).

Across the 36 countries, the interview had a median length of 59 minutes. The questionnaire asked about complex topics, including conflict and crime, democracy, elections, gender equality, governance, identity, macroeconomics and markets, political participation, poverty, public services, social capital, and tolerance.

Language Measure

The data include three language variables: (1) the respondent's first language, which the interviewer asked at the beginning of the questionnaire (example in Kenya: "Which Kenyan language is your home language?"); (2) the data collector's first language; and (3) the language of the interview. Across the 36 countries, there were 414 unique respondent home languages, 203 unique interviewer home languages, and 104 survey languages.

Using these variables, we created a five-category taxonomy that describes the combinations of respondent first language, data collector first language, and interview language (see Table 5-1). This table describes each category and provides an example from Kenya. The first two categories (common language and opt out of first) occur when a respondent and data collector share the same first language. The remaining three categories (data collector compromises, respondent compromises, third language as bridge) occur when a respondent and data collector do not share the same first language.

Languages Used in Afrobarometer (Research Goal 1)

Figure 5-1 shows the distribution of cases across the five categories in the taxonomy. The figure is sorted (ascending) by the percentage of interviews completed in the common language of the interviewer and respondent. The last row shows all countries combined. The figure shows that there is substantial variability across countries in how the language of the interview is chosen. In Côte d'Ivoire, Liberia, and Tanzania, interviews are never conducted in a common language shared by the interviewer and respondent. In 15 countries, the common language category is the majority category.

Table 5-1. Language taxonomy

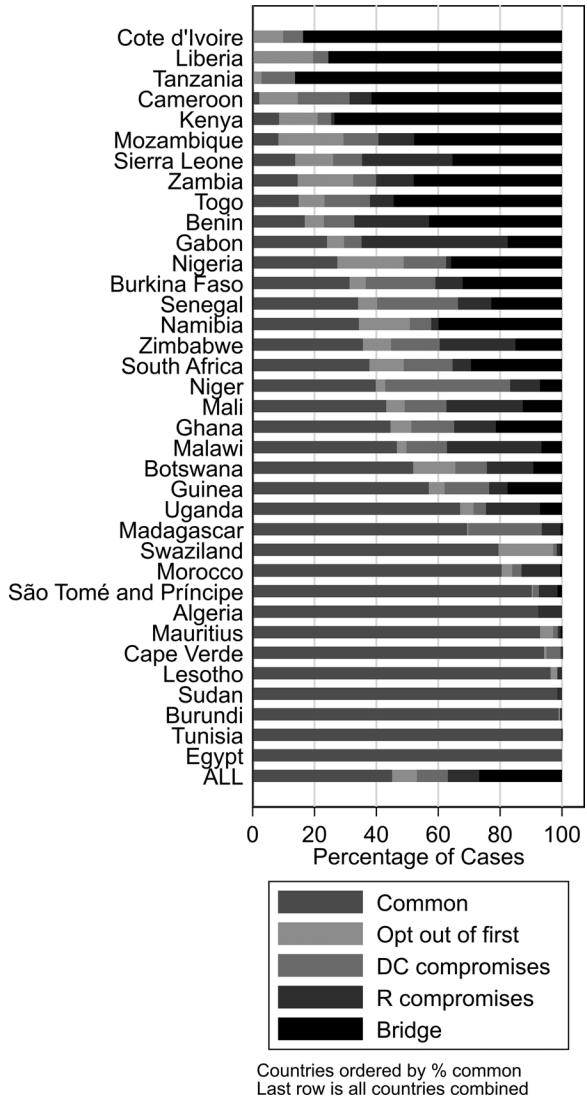
		Example (from Kenya)		
		First Language		
Name	Description	Respondent	Data Collector	Interview Language
A. Respondent and Data Collector Share First Language				
Common Language	The respondent and interviewer share a first language. The interview is conducted in that language.	Dholuo	Dholuo	Dholuo
Opt Out of First	The respondent and interviewer share a first language, but they conduct the interview in another language.	Dholuo	Dholuo	Kiswahili
B. Respondent and Data Collector Do Not Share First Language				
Data Collector Compromises	The respondent and interviewer have different first languages. The interview is conducted in the *respondent's* first language.	Dholuo	Kikamba	Dholuo
Respondent Compromises	The respondent and interviewer have different first languages. The interview is conducted in the *interviewer's* first language.	Dholuo	Kikamba	Kikamba
Third Language as Bridge	The respondent and interviewer have different first languages, and they conduct the interview in another language.	Dholuo	Kikamba	Kiswahili

Overall, when the interviewer and respondent do not share a home language, the two are about equally likely to compromise (in 10 percent of cases, the data collector compromises; in 10 percent of cases, the respondent compromises).

Description of Three Countries (Research Goal 2)

To provide a closer look at how respondents and interviewers choose which language to use in the interview, we look in-depth at three countries: Cameroon, Kenya, and Mozambique. In Figure 5-1, Cameroon is the fourth country from the top, and Kenya and Mozambique are the fifth and sixth countries from the top. We explore these countries for various reasons. First, they represent multilingual countries where data collectors and respondents could share more than one common language to choose from when conducting an interview. Second, they each have different languages of

Figure 5-1. Taxonomy of language choice, by Afrobarometer country



broader communication: English and French in Cameroon, Kiswahili and English in Kenya, and Portuguese in Mozambique. For each country, we cross-tabulated the language taxonomy with the survey language. This analysis makes the taxonomy more concrete and provides suggestive evidence for respondents and data collectors' language choices.

Table 5-2 shows the survey language by language taxonomy in Cameroon, Kenya, and Mozambique. For each country, we show the distribution of language taxonomy (e.g., in Kenya, 9 percent of all interviews used a common language, and 74 percent used a third language as a bridge). Then we report the specific languages used within each language taxonomy category. An example from Kenya: 18 percent of all interviews in which a common language was used were conducted in Gikuyu. Similarly, 51 percent of opt out of first interviews were conducted in English.

In the first column, the common language consists of respondents and data collectors speaking the same first language. In the case of Kenya and Cameroon, the common language category consisted of local languages—not broader languages. Nine percent of cases in Kenya and only 2 percent of cases in Cameroon fell into this category; it is rare for data collectors and respondents to speak the same language and do the interview in that language. In Kenya, interviews conducted in a common language were typically conducted in Dholuo (47 percent), Kikamba (21 percent), and Gikuyu (18 percent). In Cameroon, most interviews in a common language were conducted in Foufouldé (84 percent). In the case of Mozambique, fewer than 1 in 10 interviews (8 percent) was conducted in a shared first language; however, unlike Kenya and Cameroon, most surveys conducted in a common language were in a broader language (Portuguese).

Across the three countries, it was more common to find instances in which the data collector and respondent spoke a common language but chose to do the interview in a different language: respectively, 12 percent, 12 percent, and 21 percent of cases in Cameroon, Kenya, and Mozambique resulted in the opt out of first category. In nearly all these cases, they chose a broader language for the interview. This is interesting because, theoretically, the conversation could have been done in their first and common language, but a broader language may have been used because the survey was perceived as a more formal activity or the words in the questionnaire were easier to use in a broader language.

In instances in which the first language was not shared by the respondent and data collector, we see that the data collector compromises and respondent compromises categories were rare in Kenya (4 percent and 1 percent, respectively). In contrast, in Mozambique, the data collector compromised in 11 percent of cases, and the respondent compromised in 12 percent of cases. When the data collector compromised, the survey language was Portuguese 58 percent of the time; when the respondent compromised, they almost

Table 5-2. Survey language in Kenya, Mozambique, and Cameroon, by language taxonomy

Survey Language	Share First Language		Do Not Share First Language		
	(1) Common Language	(2) Opt Out of First	(3) Data Collector Compromises	(4) Respondent Compromises	(5) Third Language as Bridge
Cameroon					
Percentage of all cases (row %)	2	12	17	7	62
Languages					
English	0	8	0	78	8
French	0	76	88	0	82
Foufouldé	84	0	17	19	3
Pidgin	0	16	0	0	7
Ewondo	4	0	0	0	0
Other	12	0	0	2	0
Total	100	100	100	100	100
Kenya					
Percentage of all cases (row %)	9	12	4	1	74
Languages					
English	0	51	2	68	30
Kiswahili	0	48	57	0	69
Gikuyu	18	0	10	0	0
Dholuo	47	0	17	16	0
Luhya	0	1	2	0	0
Kikamba	21	0	4	4	0
Kalenjin	0	0	2	0	0
Kisii	0	0	1	0	0
Somali	13	0	1	12	0
Other	1	0	4	0	0
Total	100	100	100	100	100
Mozambique					
Percentage of all cases (row %)	8	21	11	12	48
Languages					
Portuguese	60	93	58	91	93

(*Continued*)

Table 5-2. Survey language in Kenya, Mozambique, and Cameroon, by language taxonomy (*Continued*)

	Share First Language		Do Not Share First Language		
Survey Language	(1) Common Language	(2) Opt Out of First	(3) Data Collector Compromises	(4) Respondent Compromises	(5) Third Language as Bridge
Makhuwa	11	2	7	1	2
Sena	4	0	11	1	1
Ndau	7	0	6	2	0
Changana	17	0	15	5	0
Other	0	4	3	0	4
Total	100	100	100	100	100

always (91 percent) used Portuguese. In Cameroon, data collectors compromised more than respondents did (17 percent versus 7 percent, respectively). When respondents compromised, they used English most often. In contrast, when data collectors compromised, they used French most often. This pattern suggests that the Cameroonian data collectors mostly speak English as their first language.

In all three countries, when the data collector and respondent had different first languages, they most often chose to use a third language as a bridge for communication (62 percent of the time in Cameroon, 74 percent in Kenya, and 48 percent in Mozambique). The bridge language was nearly always a broader language. In Kenya, among the cases that relied on a bridge language, 30 percent used English and 69 percent used Kiswahili. In Mozambique and Cameroon, the majority of cases relied on Portuguese and French, respectively.

Analysis of Language Choice (Research Goal 3)

The previous analyses focused on aggregate patterns of languages across countries and within three countries. Next, we seek to understand language patterns on a micro level, that is, between the respondent and data collector. To analyze Research Goal 3, we focused on two issues. First, when respondents and data collectors speak the same first language, why do some interviews occur in that first language (common language) and others occur in a different language (opt out of first)? Second, when respondents and data collectors do not speak the same first language, why do respondents compromise in some cases, whereas data collectors compromise in other cases?

Respondents and data collectors may opt out of their common language in favor of a different language for a variety of reasons. In some instances, they may opt out because the questionnaire was not translated into the first language. Alternatively, respondents and data collectors may be accustomed to using technical terms in a language of broader communication. The decision to opt out of first language could also reflect the interviewer's discomfort with reading the local language. Data collectors may be accustomed to speaking a mother tongue (e.g., Dholuo) but feel more comfortable reading in a broader language (e.g., English or Swahili). Finally, there may be social benefits for respondents in showing that they can participate in an interview in English or Swahili, for example.

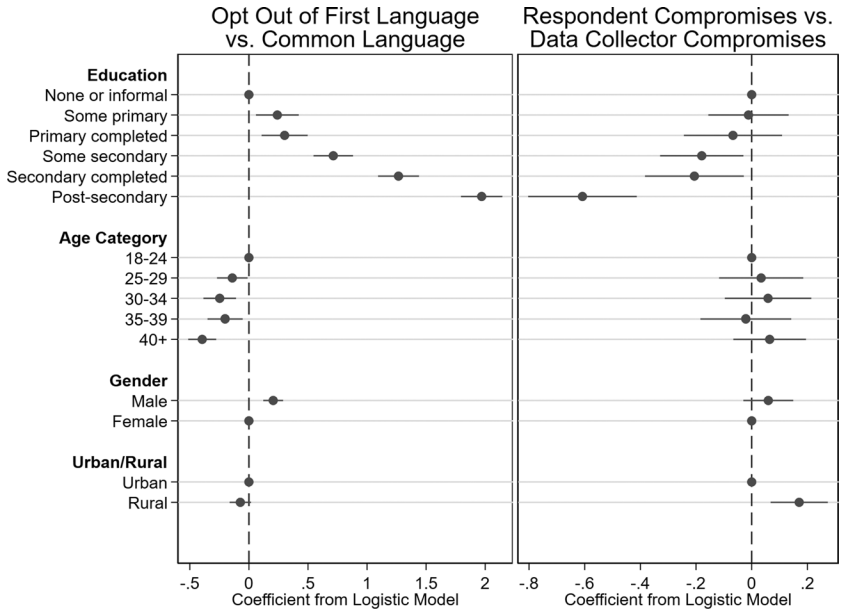
For cases in which respondents and data collectors do not speak the same language, the available languages may constrain the decision of who (respondent or data collector) compromises. If a respondent is multilingual and the data collector speaks only one language (the respondent's second language), then the respondent compromises. Other dynamics may be at play, however. For instance, data collectors may attempt to accommodate respondents, out of politeness or to secure cooperation, by using the respondent's language. Alternatively, some interviewers may insist on their own first language to exert power over respondents or because they are more comfortable administering the survey in that language.

Answering these two questions requires knowledge of all languages spoken by the respondents and interviewers and all languages in which the questionnaire was available in each country. Information on the languages each party speaks would help us understand the choices available to the respondent and the data collector (the demand side of the language-choice decision). Unfortunately, the Afrobarometer Round 6 data only include information about first languages. Information on the questionnaire languages would help us understand the supply side. These details were not available to us at the time of this writing. Without both pieces of information, we cannot fully model the choices the respondent–data collector pairs make.

We can make progress toward answering the two questions posed earlier, however, by understanding the respondent characteristics that predict whether a case is opt out of first rather than common language (to answer the first question) and whether a case is respondent compromise rather than data collector compromise (to answer the second question).

Figure 5-2 shows parameters from two multilevel logistic regressions with opt out of first (left-hand panel) and respondent compromises (right-hand

Figure 5-2. Coefficients from multilevel logistic regression models predicting language choice



panel) conditions as the dependent variables. Both regressions use the following respondent characteristics as independent variables: education, age, gender, and urban or rural residence. The regressions pool all countries and include a random effect for country. The figure shows estimated beta coefficients from a logistic model (not odds ratios). Note that the figure includes points for the reference categories for completeness. We include 95 percent confidence intervals for the estimates. Estimates where the confidence interval does not cross zero are considered statistically significant.

In the left-hand panel, respondents with more education are more likely than their less educated peers to opt out of their first language. Similarly, younger respondents are more likely to opt out than older respondents. Men and urban residents are also more likely to opt out. In Table 5-2, we saw that most opt-out interviews were conducted in a broader (rather than local) language. These characteristics (more education, younger, male, and urban) are all markers of social advantage, suggesting that these respondents may have better skills in a broader language.

In the right-hand panel, we see that education is the only statistically significant predictor of whether a respondent, rather than the data collector,

compromises. When respondents have post-secondary education, it is less likely that the respondent compromises and more likely that the interviewer compromises. Possibly, respondents have greater bargaining power for language choice as their education increases. Alternatively, more educated respondents may also know more languages, increasing their linguistic options. When respondents have lower levels of education, both forms of compromising are equally likely.

Discussion

This chapter provides a broad-brush, descriptive account of linguistic issues in a major study of public opinion surveys across 36 African countries. We developed a taxonomy to illustrate the relationships between three language variables: the interview language, the respondent's first language, and the data collector's first language. Our analysis reveals considerable variation across countries in language usage.

When respondents and data collectors share the same first language, we find that the parties sometimes opt out of that first language and choose to use another language for the interview—often a language of broader communication. This opting out phenomenon is interesting because the parties could have used a common language but chose not to. The reasons for opting out are not apparent from our data. We speculate that data collectors and respondents may choose to opt out because they view broader languages as more appropriate for a survey or because technical terms may be easier to discuss in a broader language. Opting out of a first language in favor of a language of broader communication may affect survey estimates. In the case of Afrobarometer surveys, choosing to conduct the survey in English (versus a local language) may lead respondents to report more favorable views toward the international community. Testing this idea would require an experiment that randomly assigns respondents to a local or broader language to evaluate the impact of language on survey estimates.

We also find scenarios where respondents and data collectors do not share the same first language; in this scenario, either the respondent compromises (using the data collector's first language) or the data collector compromises. The frequency of compromise is about the same for respondents and data collectors in the total sample, although it varies by country. As survey managers, we would prefer that respondents not compromise to avoid situations in which they do not fully understand the question or cannot

express their answers. It would be especially troubling if lower levels of respondent education increased the likelihood of respondent compromising. But fortunately, our results showed this was not the case.

Our research highlights methodological challenges in conducting research on linguistic issues in African surveys. The biggest challenge in this analysis concerns measurement of languages. The data we analyzed have information only about the respondents' and data collectors' *first languages*. Many Africans are multilingual, so a mismatch in first languages between respondents and data collectors is not necessarily a sign that they cannot communicate effectively. Additional information on the languages spoken by respondents and data collectors would provide a more accurate portrait. Most useful would be a measure of second and third languages spoken by both parties. Here, measurements of both *proficiency* and *preferences* would be relevant. Proficiency is understanding the set of language choices. Preferences would help us understand language choices given a similar choice set. Further, measures of proficiency and preferences would be useful for both spoken and written ability. Whereas parties both need to speak the language, data collectors also need to read it. Data collectors may be more comfortable reading in a language of broader communication, but both parties may be more comfortable speaking in a local language.

Another measurement issue concerns the coding of interview language. Like most surveys, the Afrobarometer codes survey language as a single response. From our experience in the field, however, we know that respondents sometimes switch between languages within an interview. Future research—perhaps based on audio recordings of interviews—would benefit from more information about how often this happens and when.

After these measurement issues are addressed, one next step is to investigate the association between language choice and indicators of data quality. We may expect language choices (particularly respondent compromises) to affect acquiescence, item nonresponse, nondifferentiation in scales, and interview length. This research would need to address several factors. First, language is not randomly assigned: language is highly correlated with ethnicity, and there is evidence that interviewer ethnicity affects responses in the Afrobarometer (Adida, Ferree, Posner, & Robinson, 2015). Second, this research would ideally be conducted separately by country to capture the unique context of each country. Third, this research should

include the full set of respondent and data collector characteristics. In the future, we plan to replicate and expand this analysis with another survey that contains additional information about respondents and interviewers.

References

Adida, C. L., Ferree, K. E., Posner, D. N., & Robinson, A. L. (2015). Who's asking? Interviewer coethnicity effects in African survey data. Afrobarometer Working Paper No. 158. Retrieved from https://afrobarometer.org/sites/default/files/publications/Working%20papers/afropaperno158.pdf

Afrobarometer Network. (2014). Afrobarometer Round 6 survey manual. Retrieved from https://www.afrobarometer.org/sites/default/files/survey_manuals/ab_r6_survey_manual_en.pdf

Ahlmark, N., Algren, M. H., Holmberg, T. Norredam, M. L., Nielsen, S. S., Blom, A. B., … Juel, K. (2014). Survey nonresponse among ethnic minorities in a national health survey–Mixed-method study of participation, barriers, and potentials. *Ethnicity and Health*, *20*(6), 611–632. https://doi.org/10.1080/13557858.2014.979768

Andreenkova, A. (2018). How to choose interview language in different countries. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 293–324). Hoboken, NJ: John Wiley & Sons.

Bodomo, A. (1996). On language and development in Africa: The case of Ghana. *Nordic Journal of Africa Studies*, *5*(2), 31–51.

Eberhard, D., Simons, G., & Fennig, C. (Eds.), (2019). *Ethnologue: Languages of the world*. (22nd ed.). Dallas, TX: SIL International.

Isbell, T. A. (2017). Data codebook for Round 6 Afrobarometer Survey. Retrieved from http://afrobarometer.org/data/merged-round-6-codebook-36-countries-2016

Logan, C. (2017). 800 languages and counting: Lessons from survey research across a linguistically diverse continent. Afrobarometer Working Paper No. 172. Retrieved from https://afrobarometer.org/publications/wp172-800-languages-and-counting-lessons-survey-research-across-linguistically-diverse

Pearson, W. S., Garvin, W. S., Ford, E. S., & Balluz, L. S. (2010). Analysis of five-year trends in self-reported language preference and issues of item non-response among Hispanic persons in a large cross-sectional health survey: Implications for the measurement of an ethnic minority population. *Population Health Metrics*, *8*, 7. https://doi.org/10.1186/1478-7954-8-7

Peytcheva, E. (2008, May). Language of administration as a cause of measurement error. Paper presented at the 63rd Annual Conference of the American Association for Public Opinion Research, New Orleans, LA.