Assessing and Improving Survey Data Quality in Sub-Saharan Africa

P. Linh Nguyen

A thesis submitted for the degree of
Doctor of Philosophy in Survey Methodology

Institute for Social and Economic Research
University of Essex
January 2026

Declaration

No part of this thesis has been submitted for another degree.

I am the sole author of this whole thesis.

Acknowledgements

This research was funded through several channels. I received a studentship in Advanced Quantitative Methods by the Economic and Social Research Council (ESRC) and a funded part-time contract to complete my thesis through the Female Professors Programme III (PPIII) of the German federal government at the University of Mannheim. Data collection for the three waves of the Zambian Savings Group Panel (ZamSaP) was funded by the International Initiative for Impact Evaluation (3ie) awarded to Markus Frölich (University of Mannheim) [grant number TW4.1025]. The majority of my part-time contracts as the manager for ZamSaP was funded by the Collaborative Research Centre 884 “Political Economy of Reforms” at the University of Mannheim funded by the German Research Foundation (DFG). The behavioural coding for Chapter Three was funded through the Research Methods Initiative 2020 awarded to myself by the Innovations for Poverty Action (IPA) and the Methods Cluster of the Global Poverty Research Lab at Northwestern University. Also, I would like to thank the joint review panel of the World Association for Public Opinion Research (WAPOR) and the American Association for Public Opinion Research (AAPOR) for selecting my Chapter Three for the Janet A. Harkness Prize 2025.

Apart from financial support, my doctoral thesis was also nurtured through the invaluable guidance and support from mentors, peers, friends, and family. I would like to first express my lifelong gratitude to my three supervisors. My principal supervisor, Peter Lynn, has served as

a strong lighthouse and pillar having supported my PhD project from its early stages before I had even registered in the PhD programme for Survey Methodology at the Institute for Social and Economic Research (ISER). Joining at a later stage of my PhD journey, Tarek Al Baghal, kindly agreed to serve as my second ISER-based supervisor co-guiding me towards completion. Through the trust and responsibility Markus Frölich gave me as a co-principal investigator and project manager for ZamSaP, I could benefit from this unique data incorporating a quasi-interpenetrated design to measure interviewer effects. Apart from their outstanding expert opinions, they have also shown me compassion during this challenging PhD journey.

Next to my supervisors, I have had the unique opportunity to have mentors and collaborators who have influenced parts of my doctoral research. Fred Conrad supported various applications, from the doctoral scholarship by Fulbright Germany to the proposal to the Research Methods Initiative 2020. Only with the instructions of Wiebke Weber, I could implement the multi-trait multi-method experiment within ZamSaP. Despite the completion of my doctoral thesis, my research will continue to evolve thanks to multiple collaborators and supporters. Their support and mentorship have fortified my profile as a researcher. My special thanks go to Brady West, Florian Keusch, Arlette Simo Fotso, Géraldine Duthé, Julie de Jong, Dorothée Behr, Steve Dept, Tim Johnson, and Yfke Ongena. I am also grateful to Valérie Golaz and Laurent Toulemon, as well as the anonymous reviewers of *Population* who provided thoughtful feedback to a former version of Chapter One.

Additionally, I would like to thank my examiners, Pablo Cabrera Álvarez and Marieke van Gerner-Haan, who prepared insightful comments for my viva which helped to strengthen and refine some parts of my thesis. I am also grateful for Alita Nandi, Yfke Ongena, and Heini Väisänen for agreeing to hold a mock viva which prepared me thoroughly for my actual one.

My networks were essential for my positive well-being and I was fortunate to be part of several such networks. At my home institution, ISER, a reading group for survey methodology and the regular meetings of the ISER method group created a stimulating environment for learning and discussing innovative survey methods. From one of those instances, I can remember knowledgeable comments from Olena Kaminska which helped me for presenting my PhD research for the first time at a scientific conference. Throughout my student years, I have also relied on various help from Janice Webb as the ISER graduate administrator, Kim Bones as the finance officer, Jane Ismail as the library officer, Cara Booker as the current ISER director of graduate studies, and Paul Clarke as the former one. Apart from my office mates (Yuliya Kazakova, Liza Benny, Valerija Kolbas, Alexander Wenz, Jasmine Mitchell) and the kind Nicole James, I have benefitted greatly from interacting with my peers at the French Institute for Demographic Studies (INED), foremost Lucie Vanhoutte and Julie Maestri, and at the European Doctoral School for Demography (EDSD) cohort (especially Ibraheem Ahmed, Maria Gargiulo, and José Andrade Santacruz). I am particularly grateful to my PhD family (Ai Rene Ong, Fernanda Alvarado Leiton, Micha Fischer), who have been with me through the highs and lows of PhD life and beyond.

Last but not least, I express my immense gratitude to my dear friends and family. My parents and my sister have encouraged my passion for research throughout, and later played a key role in relieving me of some care responsibilities. My wonderful daughter and son have brought new joy and laughter to my life which have given me the energy to persevere. And nothing would have been possible without my partner who has been with me from the moment I first thought about starting a PhD to the moment I submitted my thesis. His dedication to our family has allowed me to pursue my research and his unwavering support is reflected in my achievements.

Summary

High-quality survey data is the foundation for informed and valid decision-making, policymaking, and research. In low- and middle-income countries (LMICs), surveys rely heavily on interviewers to administer questions to respondents who have varying degrees of literacy and who live in areas without stable telephone or internet access. Another characteristic of LMICs, especially in sub-Saharan Africa, lies in their multi-ethnic and multilingual nature. This doctoral thesis examines survey data quality using a Zambian panel survey through the lens of the interviewer, exploring gender-of-interviewer effects (chapter 2), connecting interviewer performance to translation quality through different modes of translations (chapter 3), and examining language switching as a problematic interaction in multilingual surveys (chapter 4).

To estimate interviewer effects, the thesis leverages the random assignment of interviewers to respondents in each of the three panel waves conducted in 2016, 2018, and 2019. The analysis combines the survey data with two additional sources of data: (a) the self-administered interviewer surveys capturing key interviewer characteristics, such as age, gender, and language proficiency; and (b) the behavioural coding of interactions between respondents and interviewers during the survey interview.

The findings of Chapter One demonstrate non-negligible interviewer effects for behavioural, knowledge, and attitudinal questions, which call for a more nuanced approach to studying social desirability bias. Chapter Two shows a considerable reduction of interviewer effects from the first wave (which had no scripted questionnaire translation) to the subsequent waves (which employed a pretested translation into the *lingua franca* of the three study regions). Chapter Three examines the frequency of language switching across different questions and two different regions demonstrating its co-occurrence with other problematic interactional

behaviours that are known to influence survey data quality. The latter two studies highlight and recommend good practices in questionnaire translation and interviewer selection and training in multilingual settings.

Contents

Introduction

Thirty years ago, Grosh and Glewwe underlined that “[w]e have systematic evidence on neither data quality nor changes in quality” for data stemming from low- and middle-income countries (LMICs) (Grosh & Glewwe, 1996: 18). About twenty years later, Jerven and Johnston explicitly affirmed that “[m]uch academic work on Africa regularly uses flawed data, but not all researchers demonstrate awareness of the flaws” (2015: 112). Consequently, the impetus for this doctoral research is motivated by the limited knowledge on survey quality in sub-Saharan Africa, in particular regarding the multilingual survey context, and pursues following overarching research question:

How do different aspects of survey design impact the quality of surveys in sub-Saharan Africa?

Albeit many researchers from different disciplines are concerned about survey data quality in LMICs, published articles are scattered across different journals ranging from demography, development economics, political science, to survey methodology.

Only few studies have embarked on studying certain aspects of survey data quality in LMICs in general and in the sub-Saharan African context in particular. With the exception of the recent emergence of scientific articles on quantifying interviewer effects in LMICs (e.g., Di Maio & Fiala, 2020; Footman, 2021; Leone et al., 2021; Rozelle et al., 2023; A. Singh et al., 2022; S. Singh et al., 2024) research on the quality of survey data from LMICs can still be considered as nascent despite receiving growing, broader interest, e.g., in the form of about 50 submissions for the recent call for the special issue of the on “Survey Research from Asia-Pacific, Africa, the Middle East, Latin America, and the Caribbean” (Berg & Edwards 2025).

Data collection practices and challenges in sub-Saharan Africa

Hughes and Lin (2018) bring forward five commonly shared challenges from which data collection for African surveys suffers, among them the multiplicity of languages. The "multilingual nature of [S]ub-Saharan societies" is evident by their ethnic and linguistic heterogeneity with their citizens disposing a large individual linguistic repertoire containing on average two languages (Buzasi 2016). Such a diverse linguistic environment leads to the common context of mostly multilingual respondents being surveyed by multilingual interviewers. Consequently, interviewer practicing on-the-fly translation (also referred to as on-the-fly or oral translation) "is common even among the larger [cross-national] surveys" due to the cost constraints regarding providing and testing multiple translations (Hughes & Lin, 2018: 539). Lau and colleagues underline that data quality is impacted when "[u]sing a nonhome language in a survey, either because of choice or constraint" (2020: 103), e.g., in the form of response errors, such as "acquiescence, item nonresponse, nondifferentiation in scales and interview length" (2020: 113). The authors further point out that, to understand the impact of multilingualism on data quality, researchers need to know both the level of proficiency and the individual preferences regarding the languages spoken by survey respondents and interviewers (Lau et al., 2020).

While there is compelling evidence from Germany and the US about the detrimental effects of on-the-fly translation and data quality (Harkness et al., 2008), only few sparsely cited studies have been dedicated to the investigation of this topic. An unnoticed article using survey data from the Philippines and Côte d'Ivoire by survey methodologists from 1988 concludes that pre-translated questionnaires are less error prone than on-the-fly translations by interviewers (Scott et al., 1988). Using Kenyan data collected in 2018 as part of the Demographic and Health Surveys (DHS), Weinreb and Sana (2009) demonstrated that 23% of measurement errors were due to on-the-fly translation. A study using Turkish DHS data showed that on-the-fly translation introduces higher level of nonresponse (Sarac & Koc, 2021).

Another important fieldwork reality from the list of Hughes and Lin (2018) is the high dependence on interviewers to collect survey data. This reliance is due to the low penetration rate of the internet, telecommunication network issues, uneven literacy rates and high illiteracy levels among the older population, and the fact that many local languages lack either a written form or a widespread knowledge of them among the population. This dependence on interviewers led Randall and colleagues (2013) to investigate "the power of the interviewer" in African surveys. These arguments underline the importance of this doctoral research to understand interviewer effects for an assessment of survey data quality in the sub-Saharan context. This doctoral research relies on the Dijkstra (1983)'s definition of interviewer effects operationalised as the proportion of the total response variance attributable to the interviewers or a selected interviewer characteristic, such as the gender-of-interviewer effect.

Assessing survey data quality in sub-Saharan Africa

Although the methodological reports of the World Fertility Surveys in the 80s have been a remarkable starting point for systematic research into survey data quality in LMICs (e.g., Cleland & Verma, 1989; O'Muircheartaigh, 1982), the comprehensiveness of the study topics and designs produced in LMICs continue to lag behind that of high-income countries. In recent years, however, different fields (such as development economics, political scientists studying LMICs, and demographers interested in the Global South) have produced a growing number of studies on survey data quality issues in LMICs.

Due to the continued high reliance of interviewers in sub-Saharan Africa, the key encompassing aspect of survey design of this thesis is the interviewer, and the impact interviewers have on data quality in a multilingual context found in many sub-Saharan countries. Therefore, this PhD project set out to answer three research questions in line with the two main fieldwork challenges mentioned in the previous section, namely the high dependence on the interviewer and the multilingual context of survey interviewing:

RQ1. To what extent does the gender of the interviewer influence responses in a face-to-face survey?

RQ2. To what extent do interviewer effects influence responses in a multilingual, face-to-face survey context?

RQ3.1: Regardless of who initiated it, is language switching associated with another problematic interactional behaviour?

RQ3.2: How is language switching associated with question, respondent, and interviewer characteristics?

This PhD project aims for a threefold contribution. Firstly, the research detects sources of survey error coming from the interviewer to avoid spurious inferences and therefore make a better value for money for surveys carried out in the sub-Saharan context. Furthermore, the findings expand the field of survey methodology to multilingualism. Finally, the summary of research findings can then be further developed into evidence-based guidelines for survey methods in developing settings which are beneficial for survey practitioners and researchers working in both developing and developed countries.

As a better understanding of survey quality allows for control and minimisation of survey error, more accurate data eliminates spurious conclusions avoiding wrong policy implications (Strauss & Thomas, 1996) and research findings. To investigate the social and economic development, policymakers and practitioners rely heavily on survey data (often collected as part of evaluation studies of development projects mostly run with public funding) as a complement to (sometimes outdated) official statistics (Devarajan, 2013).

Traditionally, policymakers and researchers alike have pointed out the challenges for policymaking and research due to infrequent administration of population censuses and population surveys which has often forced “donors [themselves] […] to collect the data as quickly as possible” (Devarajan, 2013: S14). De facto, numerous ad-hoc data collections have been funded for policymaking, government accountability, and as part of the monitoring and evaluation of mainly ODA-financed development projects as “donor organi[s]ations acknowledge the value of survey data as input for sound decision-making” (Woolfrey, 2009: 22). Apart from these surveys within development evaluation, public opinion polls have also been on the rise since the start of the new millennium but with varying quality (Conroy-Krutz, 2019).

Africa plays a unique role in the field of international development as recipient of the biggest share of net official development assistance (ODA) with ca. 59.7 billion USD in 2023, or 26.8% of foreign aid. If the growing share of ODA spent directly in the donor countries is excluded, Africa has continued to receive the largest share among the world regions since 2006 (see https://data.one.org/analysis/official-development-assistance).

Leverage unique data collection in Zambia in form of the Zambian Savings Group Panel (ZamSaP)

The main survey data used for analysis are part of an impact evaluation of a development program called Rural Financial Expansion Programme (RUFEP), mainly funded by the International Fund for Agriculture (IFAD). RUFEP aims to provide access to financial services to the people living in rural or semi-urban areas in Zambia through financial and technical support to local non-government organisations. In this case, the target group are members of selected savings groups in the Eastern, the Northern, and the Western Province which were trained to strengthen their savings, lending, and internal insurance schemes. To study the impact of RUFEP, the Zambian Savings Group Panel survey (ZamSaP) has been implemented to capture changes at the individual savings group members' level and household level. Only the data from the individual questionnaire and interviews are analysed for this thesis.

To estimate interviewer effects, I implemented a random assignment of interviewers to respondents in each of the three waves of the ZamSaP study conducted in 2016, 2018, and 2019. This procedure, known as interpenetration, has been introduced in low- and middle-income countries as part of the "Response Error Project" within the World Fertility Surveys in the seventies (C. A. O'Muircheartaigh, 1982). Likewise, Stecklov and Weinreb advocate for the interpenetration design to "reliably judge the extent to which [certain interviewer] characteristics matter" in producing interviewer effects (2010: 52). Complementary to the survey data, this thesis makes use of two critical other sources of data: (a) the self-administered interviewer surveys providing additional information on the interviewer level; and (b) the behavioural coding of question snippets of selected survey items. These two additional data sources in combination with the interviewer penetration enables a novel perspective in the study of interviewer effects.

My role as co-principal investigator for the impact evaluation of RUFEP provided the necessary opportunities to implement key design features. These include the interpenetration of interviewers and the collection of interviewer characteristics, as well as the integration of auxiliary studies such as those on behavioural coding or on a multi-trait multi-methods experiment comparing different response scales to the principal ZamSaP data collection. Through my collaboration with the local survey firms, I was involved in all trainings and pre-testing events and vetted the interviewer manuals and instructions in addition to monitoring and supervising fieldwork. Through the successful acquisition of additional funding, it was, for example, also possible to hire former ZamSaP interviewers to transcribe and translate a small number of full interview scripts. The extracts from the interview transcripts then provided examples for language switches happening within a full question-answer sequence. Based on my observations from the field, I could steer the data collection processes for subsequent waves to better adapt to the complex, multilingual context of Zambia. For example, I allocated a proportion of the budget of wave 3 to pay external translators who independently translated the survey instrument in the three *lingua franca*, documented translation discrepancies, pre-tested, and trained the interviewer in the dedicated local language.

As many of its neighbours in sub-Saharan Africa, Zambia is a multi-ethnic or multi-tribal society. This leads the average Zambian to grow up as multilingual speaker. Next to its official language, English, Zambia is home to twenty-six indigenous languages – many of them incorporating different dialects (Marten & Kula, 2008). For ZamSaP, the questionnaires were translated from its source questionnaire in English in the *lingua franca* of the respective regions: Nyanja (also known as Chewa) in the Eastern Province, Bemba for the Northern Province, and Lozi for the Western Province. These survey languages have the status of official national language, which means that they also gained prominence as teaching languages in primary education since the change of the millennium (Marten & Kula, 2008).

In this peculiar context of Zambia, this doctoral thesis assesses the data quality of ZamSaP survey data and starts its investigation through the lens of the interviewer exploring gender-of-the interviewer effects (Chapter One), connecting interviewer performance to translation quality through different mode of translations (Chapter Two), and examining language switching as a problematic interaction in multilingual surveys (Chapter Three). It concludes by summarising the main research findings and drawing practical implications, as well as providing avenues for future research.

Chapter One: Revisiting the gender-of-interviewer effects

Abstract

Survey interviewers are often indispensable for the data generation process. In low- and middle-income countries (LMICs), interviewer-administered face-to-face (F2F) surveys were the most widely used data collection tools before the COVID-19 pandemic and continue to play a major role in the post-COVID era due to several restrictions concerning internet or telephone network stability. The way interviewers pose questions has a direct influence on data quality. Thus, demographers and social scientists using survey data generated in LMICs and those interested in studying sensitive topics (e.g., abortion, domestic violence, financial behaviour, trust) have been paying more attention to interviewer effects.

This paper focuses on the gender-of-interviewer (GOI) effect and compares interviewer effects between attitudinal, knowledge, and behavioural questions using a face-to-face survey fielded in Zambia in 2016 (N=2051). In addition to investigating the GOI effect as the main effect and their interaction with the gender of the respondent, the GOI effect is compared between attitudinal questions on trust, as well as behavioural and knowledge questions on financial issues and civic engagement. The findings demonstrate non-negligible interviewer effects which threaten survey data quality. These results suggest that the social desirability bias induced by interviewers in a face-to-face setting can be present for any question type and topic.

Introduction

The need to study interviewer effects in low- and middle-income countries

Interviewers play a crucial role in data collection (either via face-to-face or via the phone), especially in low- and middle-income countries (LMICs) where the lack of widespread infrastructure and universal education may limit the possibility for representative web surveys. Even after the "push-to-web" movement, in which face-to-face surveys were suspended during COVID-19 in favour of telephone or web surveys, the dependence on interviewers to administer questionnaires remains (see Frankovic et al., 2023). For interviewer-administered surveys, both data producers and users (coming from many different domains of academia, policy, or civil society) rely on interviewers to ensure data quality so that they can derive valid answers for research, decision and policy making. The way survey interviewers pose survey questions and collect answers has, thus, a direct influence on data quality. Especially, population and social scientists often interested in sensitive topics (e.g., abortion, domestic violence, financial behaviour, trust) should pay attention to whether there is an interviewer effect in interviewer-administered survey data.

Despite ongoing growing penetration of smartphones and increasingly high percentages of internet usage in LMICs, prevailing low levels of literacy (Stecklov & Weinreb, 2010), especially among older generations, fuel the demand for personal interviewing. In sub-Saharan Africa, "[i]nterviewers often conduct interviews with poorly educated or illiterate respondents, which possibly increases their impact on how questions are understood and/or answered" (Demarest, 1997: 3). In this kind of setting in which both survey implementers and respondents are highly dependent on interviewers, it is important to analyse how interviewers influence the data collection and whether they introduce a systematic bias to survey response. The majority of LMICs are characterised by gender inequalities and high ethnic heterogeneity (Stecklov & Weinreb, 2010), and their societal composition may aggravate interviewer effects.

The comprehensive overviews and syntheses on existing literature of interviewer effects drawn mainly on studies from Western countries concluded that most empirical studies generated null findings, or they show an inconclusive trend on the relationship of certain interviewer characteristics on measurement error see (Davis et al., 2010; Schaeffer et al., 2010; West & Blom, 2017). Albeit research on interviewer effects has been prominent in the survey methods literature for Western countries, only recently have researchers focused on interviewer effects in LMICs surveys (e.g., Di Maio & Fiala, 2020; Footman, 2021; Leone et al., 2021; Rozelle et al., 2023; A. Singh et al., 2022; S. Singh et al., 2024). This recent surge has in common that it demonstrates large interviewer effects and raises concerns about survey data quality on highly gender-sensitive topics such as abortion or sexual violence. However, that body of research often calculates interviewer effects in terms of intra-cluster correlations (ICCs) of the interviewer variance without taking specific interviewer characteristics into account. This ambiguity between findings in Western countries vs. LMICs provides a justification for further investigation, especially as the recent above-mentioned studies were not able to look at the interviewer gender.

To contribute to this body on interviewer effects from LMICs, this paper focuses on gender as one of the most observable interviewer characteristics and compares interviewer effects between attitudinal, knowledge and behavioural questions. This paper extends the literature by revisiting the gender-of-interviewer (GOI) effect for selected questions on financial behaviour and attitudes from a face-to-face survey conducted in 2016 in Zambia. It investigates whether the GOI effect can be detected even for topics which seemingly are less gender-sensitive and raises the question whether questions about money can also induce socially desirable answers. The GOI effect is investigated in four ways: as (a) the main effect; as (b) an interaction effect between gender of interviewer and respondents; as (c) moderated by social distance between interviewer and survey participant (social distance is operationalised as difference between the "urban", educated interviewer interacting with respondents living in urban vs. rural areas); and (d) the GOI effect is further compared between questions on trust and questions on financial behaviour and civic engagement.

Gender-of-Interviewer effects on measurement error – findings from low- and middle income countries

More recent studies from LMICs highlight the concerningly high interviewer effect on abortion reports in surveys (Footman, 2021; Leone et al., 2021) and female domestic violence and abuse (Singh et al., 2022), as well as on indicators for household food security (Pietrelli et al., 2021) implemented random pairs of interviewers and respondents to overcome the lack of interpenetration. However, these studies focus on estimating the interviewer effects in terms of intra-cluster correlations and had often little disposal of specific interviewer characteristics, e.g., gender.

Finding inconclusive findings regarding the GOI effect from selected studies, mostly from Western countries, the synthesis on interviewer research by West and Blom (2017) can be regarded as partly outdated with studies from LMICs only constituting a very small share. Only three out of the 23 studies examining the effect of interviewer gender on response distributions considered by West and Blom come from a non-Western setting, all involving face-to-face surveys: Flores-Macias and Lawson presented findings from Mexico (2008), Liu and Stainback (2013) investigated Chinese data, and Liu and Wang (2016) covered six countries or societies from the Asian Barometer. However, research on the GOI effect needs to encompass as many different countries and contexts as possible to investigate "[h]ow respondents react to interviewer gender may vary not only by the topic of the question, but by geographic or cultural region" (Schaeffer et al., 2010: 453).

The section extends the review of gender-of-interviewer effects on measurement error from LMICs for diverse survey question topics. Therefore, it excludes research on unit nonresponse bias attributable to interviewers (see Amos, 2018), or systematic interviewer errors when providing interviewer observations (see Wu & Xie, 2024), administering the questionnaire (see Sharma et al., 2022) or non-traditional survey measures, such as measuring anthropometric data (see Dwivedi et al., 2022). As the GOI effect may differ when facing either a female or a male respondent (West & Blom, 2017), findings on interaction effects are summarised in this literature review whenever included in the studies. As interaction effects between the GOI and the gender-of-respondent (GOR) is of interest, this literature review does not encompass studies on the GOI surveying either men (Hathi et al., 2025) or women only (Becker et al., 1995; Harling et al., 2019; Kianersi et al., 2020).

The first published studies investigating the GOI effect in a LMIC context demonstrated differences in data quality between female and male interviewers. Using Nepalese data, Axinn examined the GOI effect both for demographic questions and sensitive, behavioural questions on economic activity and fertility. The author found weak evidence for female interviewers collecting better demographic data (Axinn, 1989) and stronger evidence of this female superiority when recording answers to sensitive questions on economic activity (Axinn, 1991). Relying on a Mexican survey, Flores-Macias and Lawson (2008) concluded that both the GOI effect and an interaction between gender of interviewer and respondents, as well as a moderator effect between urban and rural respondents were present in gender-sensitive questions on abortion and women's rights but not found in other questions. Consistent with findings from Western contexts (see West & Blom, 2017), (survey) researchers studying the Global South became more aware of the GOI effect in certain attitudinal, sensitive, and/or subjective questions.

This seemingly established differential GOI effect in factual and attitudinal questions also motivated Himelein's study in Timor-Leste which demonstrated a GOI effect for both question types with a larger GOI effect in the later ones. In that study, factual question covered agricultural practices while subjective questions touched upon corruption, law and order, and women's rights. Although an interaction between the GOI and the GOR per se was absent, the author found that female respondents were more likely to be influenced by their interviewer's inherent attitudes to the same item. Albeit the study is the only one in this literature review who did not either rely on some randomisation between the interviewer and the respondents or control for the area effect in the absence of interpenetration (Himelein, 2016), the study represents one of the few which showed significant results for the GOI effect in factual questions.

For the Chinese context, Liu and Stainback (2013) concluded that the GOI effect in marriage-related questions were present in some items but not in all, as is the case for interactions between the gender of interviewer and interviewee. In one of the few existing comparative studies, Liu and Wang (2016) revealed a GOI effect on acquiescence in 53 non-factual questions with a 4-point disagree/agree scale in the Asian Barometer Study in six out of eleven Asian countries and societies. However, the authors found no evidence for an interaction between the GOI and the GOR.

Some studies have examined the GOI in political questions (both gender-sensitive and non-gendered ones). For a study on political attitudes in Morocco, Benstead and Hatfield concluded that men gave more socially progressive responses to female interviewers than male ones showing a clear interaction between the GOI and the GOR (Benstead & Hatfield, 2014). In an

Ugandan study, Di Maio and Fiala (2020) discovered GOI main effects for questions on the support of political parties but found no interaction effects between the GOI and the GOR with some significant interactions rather when it comes to the distance between the educational level of the respondent and the interviewer. With respect to measuring support for women's political leadership, Sundström and Stockemer (2022) found a GOI effect which manifested among survey respondents in the Afrobarometer while pointing out that male respondents are more susceptible to the gender of the interviewer for gender-sensitive items.

Apart from analysing the GOI effect on actual responses, some studies focused on whether female and male interviewers had a different impact on the number of item nonresponse. The above cited Moroccan study produced strong evidence that respondents of both genders tend to skip questions when surveyed by a female interviewer (Benstead & Hatfield, 2014). For a survey in Vietnam and Thailand, Phung and colleagues found both main and interaction effects between the GOI and the GOR for explaining the number of missing values per interview (Phung et al., 2015).

All referenced studies from LMICs found some non-negligible GOI (main) effects albeit they vary greatly by topic, by their focus on either only one type of question (e.g., only attitudinal items in Flores-Macias & Lawson, 2008, vs. only factual questions for Phung et al., 2015), or the comparison of both (Himelein, 2016), and by their degree of gender-sensitive items (abortion vs. agricultural practices). In addition, the interaction between the GOI and the GOR was often present. As the cited research spans over different countries and continents, one conclusive overarching finding is the presence of a GOI effect in gender-sensitive, mainly attitudinal questions lead consistently to gender-of-interviewer effects and to the interaction effect between the GOI and the GOR except for one the study with mixed results (Liu & Stainback, 2013). This dispersed evident warrants this study on studying the GOI effect for attitudinal and factual questions such as the one by Himelein (2016) in another context and expanding the range of topics to assumingly less gender-sensitive topics on financial behaviour and attitudes.

Studying the gender of the interviewer in the Zambian context

Gender is one of the easiest interviewer traits to observe, giving a clear cue to respondents. It is either studied as the main variable of interest (see studies cited in the previous section) or used as an indispensable control variable. Hence, this paper puts forward the following research question: *To what extent does the gender of interviewer influence responses in a face-to-face survey?* This question shall be answered by investigating the following effects using data from a Zambian survey on financial inclusion: (a) the main GOI effect; (b) the interaction between the GOI and the GOR; (c) social distance using a geographical dummy variable capturing whether the interviewee lives in a rural or semi-urban area; and (d) a GOI effect between knowledge, behavioural, and attitudinal questions.

Gender-of-Interviewer. The theoretical underpinning to study gender in the survey interview lies in the interpersonal power relationships pertaining to the interaction between interviewer and interviewee which may differ and/or be more pronounced in different cultural contexts. For the US context, Carli (1999) found that men exert more influence than their female counterparts as society attributes higher levels of expert and legitimate power. In their role as an interviewer, it is perceived as more legitimate for men to influence others and expect respect and deference. Even if the interview situation is a special case of social interaction, women might not be able to play their role as expert interviewers and exert their perceived competence which comes with this role well enough to bridge the gap on legitimate power. In contrast, men can extend their lead in legitimate power while performing interviews by adding the expert bonus as interviewers. Consequently, the first hypothesis of this paper is to find evidence whether there is a GOI effect:

H1: A systematic gender-of-interviewer effects exists in this survey.

Interaction between the GOI and the GOR. Power relationships between men and women are still clearly visible in everyday life in Zambia with gender-status inequalities in favour of men (Evans, 2014). The literature from LMICs previously summarised do not lead to a conclusive trend on whether there is an interaction between the GOI and the GOR. Therefore, it is interesting to study whether the GOI effect differs between female or male respondents:

H2: The gender-of-interviewer effects are different for female versus male respondents (i.e., the interaction between the GOI and the GOR).

Social distance. As Tu and Liao (2007) already showed for the Taiwanese case, observable differences, among them gender, between the respondent and the interviewer are manifested in social distance and play a significant role in explaining certain measures of survey data quality (like numbers of "don't know" and "refusal"). In a binary concept for gender, a dissimilarity of gender between interviewer and interviewee is considered as social distance. For sub-Saharan Africa, social distance dimensions refer to "age, gender, education, wealth, urban and rural origins [that] are major issues alongside ethnicity and language" (Randall et al., 2013: 766). Especially the difference between urban versus rural and semi-urban is essential as the interview situations in this Zambian survey often follow the common pattern of other surveys in sub-Saharan Africa: "illiterate, rural farmers may feel so intimidated by the well-dressed, well-educated interviewer that they do not think they can refuse to participate, but they may well demonstrate resistance by providing inaccurate replies" (Randall et al., 2013: 779). In this sense, the dissimilarity in residence could serve as a proxy to capture other non-observable differences, such as social economic status measured by education and wealth. Therefore, the location of the survey (urban vs. rural respondents) might represent a moderator for the GOI effect by cementing social distance and aggravating the GOI effect. Presumably, having a male interviewer instead of a female one will affect a rural respondent even more than an urban one.

Findings from Mexico and Timor-Leste show that the geographic location of the survey does indeed moderate the GOI effect (Flores-Macias & Lawson, 2008; Himelein, 2016). Applied to the Zambian context, the GOI effect is expected to be different between interviewees living in rural villages or in regional cities (here referred to as semi-urban areas as they cannot be compared to heavily urban areas such as the capital). Survey participants in semi-urban areas have better access to education and the provision of goods and services, which makes them more like interviewers coming from the capital. Therefore, the third hypothesis captures the concept of social distance and is phrased as follows:

> H3: Gender-of-interviewer effects are stronger for rural interviewees compared to (semi-) urban ones.

Question Type. Although the questionnaire on economic well-being and financial behaviour in Zambia includes several types of questions, this paper focuses on knowledge, behavioural, and attitudinal questions. Knowledge and behavioural questions examined in this study are mostly considered as recognition judgements about a specific action which has or has not happened in the last 12 months (getting a loan, giving out a loan and attending a village meeting) or whether the respondent has ever heard of mobile money[1]. As all survey participants are savings group members and part of the community, they would not have difficulties retrieving actions regarding receiving or giving out loans or participating in community gatherings. Furthermore,

[1] Mobile money is a generic name widely used in LMICs for sending and receiving money via mobile phones. The financial transactions are linked to a customer's mobile money account, similar to a bank account. In the case of mobile money, the sender specifies the name and mobile number of the receiver and deposits the amount to be sent at a mobile money agent which can be thought of as mobile mini banks. To cash out the amount sent, the receiver can identify himself with his mobile number and valid identification at a mobile money agent of the same operator, even in a different location. Mobile money is considered as an effective means to extend financial services to rural areas in LMICs with no financial structure like a physical bank.

mobile money is frequently used to transfer money in sub-Saharan Africa, with Zambia ranked on the 10th place with about 14% of adults with an account (Demirguc-Kunt et al., 2018). Thus, mobile money cannot be regarded as "unfamiliar or inaccessible enough" and thus, can be regarded as a factual question (Tourangeau et al., 2000: 157). In addition, these four questions were specifically chosen as they refer to familiar concepts and do not ask about duration or frequency of behaviour. As such, those questions facilitate that "the target population are likely to have encoded the information" (Schaeffer & Presser, 2003: 69) and avoid the respondents falsely "concluding that they never experienced events of a type that they never heard of before" (Tourangeau et al., 2000: 157).

According to Schaeffer and Presser, the traditional cleavage regarding question type is between "questions about events or behaviours and questions that ask for evaluations or attitudes" (2003: 66). The differentiation between distinct types of questions is important as "[i]nstructions and design features of questions can moderate or amplify interviewer effects" (Schaeffer et al., 2010: 455). According to their review, questions which cover attitudinal and sensitive topics, as well as questions of ambiguous, complex, and open-ended nature are more prone to interviewer effects. In attitudinal questions, interviewers might reveal their own attitudes regardless of doing so intentionally or unintentionally (Himelein, 2016). Thus, an interviewer might alternate his or her interviewing behaviour depending on how strong he or she feels on certain attitudinal topics. In addition, O'Muircheartaigh (1976) found interviewer effects for attitudinal items, but not for factual ones. However, it must be noted that his paper relies on survey data from merely five interviewers.

An important dimension for question characteristics is studied by Schnell and Kreuter (2005) who found that sensitive questions in a crime survey are more prone to interviewer effects compared to non-sensitive ones, as well as non-factual ones compared to factual ones. Particularly for non-factual questions which are also judged as sensitive, the authors found a higher interviewer effect compared to factual questions which are deemed as non-sensitive. This entanglement between sensitivity of the question and the traditional differentiation between knowledge and behavioural versus attitudinal questions is essential for this paper.

In comparison to factual questions, Tourangeau et al. state that “the type of information we tap in answering attitude questions depends both on its relative accessibility and on more strategic considerations, such as our level of motivation to reach a defensible position” (2000: 195). In a face-to-face context, the interviewer can both affect the motivation of retrieving the respondent’s inherent attitude and the subjective feeling of the respondent to report a socially desirable answer. Thus, it is of key importance that any study of interviewer effects makes at least the distinction between knowledge, behavioural, and attitudinal questions in their examinations, hence the fourth hypothesis:

> H4: Gender-of-interviewer effects are more likely to be significant for attitudinal questions compared to knowledge and behavioural ones.

Methods

Data and measurement

Background of study. This paper relies on a survey conducted in 2016 on financial behaviour and livelihoods of people living in rural or semi-urban areas in Zambia. This survey was part of an impact evaluation of a development program called Rural Financial Expansion Programme (RUFEP) which aimed to provide access to financial services to the rural poor through financial and technical support to local non-government organisations (NGOs) in their interventions.

The survey was conducted in 2016 in three Zambian provinces where the eligible savings groups are located. The data collection in the Northern Province in July was done by six teams of five interviewers or interviewers accompanied by one supervisor and one quality controller per team. After the presidential election in August 2016 impeding fieldwork, the survey continued 2 months later in the Western and Eastern Province involving five teams with a similar set-up. The interviewers, supervisors and quality controllers attended a training of at least five days prior to the start of each data collection exercise, whereas interviewers with uncertainties on the questionnaire were requested to attend additional training days to match up to the expected level of interviewing proficiency[2]. In addition to the regular training, there was one day of dress rehearsal in the field one day prior to the actual start of the survey. During this pre-test day, each interviewer had to complete at least one household and adult interview with savings group members not selected for the actual survey or with members of ineligible savings groups.

Sampling procedure. The Zambian survey consists of a one-stage stratified probability sample where the respondents were chosen by simple random sampling from each savings group. For all 533 eligible savings groups, meaning mature groups formed before July 2015), four members per savings group were selected randomly for individual and household interviews using the savings group registry.

Interpenetrated assignment of interviewers. Within each survey team, interviewers were randomly assigned to savings group members to prevent any selection bias caused by interviewers choosing the interviewee. Stecklov and Weinreb explain that “if male and female interviewers are randomly assigned to respondents in the same population and the mean

[2] As an exception, one interviewer was only allowed to collect data on actual survey participants after his mock interview file was deemed satisfactory, even though the survey has already officially begun. In other cases, weaker performing interviewers were scrutinised by their team leaders in the first days of the survey to ensure the quality of the interview.

response to a given question differs by the gender of the interviewer, then the existence of […] bias can legitimately be inferred" (2010: 22). This Zambian survey relies on a quasi-interpenetrated assignment, in which interviewers in groups of five are randomly assigned to selected savings group members.

Survey administration and participation. The final dataset for analysis consists of 2,051 savings group members[3] surveyed by 30 interviewers in July 2016 and 25 interviewers in October/November 2016 using either paper and pencil or laptops. For the fieldwork, interviewer teams were assigned to districts within provinces and usually more than one team shared work in one district. Regarding survey participation, the selected savings group members were contacted by the survey teams together with the savings group leadership who was facilitating the contact and request for interview. Therefore, refusals on the individual level were nearly absent. In some rare cases, the savings groups leadership refused to participate altogether (e.g., a group consisting of Jehova's witnesses refused on religious grounds).

Interviewer characteristics. As mentioned above, 30 interviewers conducted an average of 29.3 household and individual interviews per interviewer in July 2016 and 25 interviewers administered 48.8 interviews on average in October/November 2016. 15 interviewers worked for both the first and the second part of the survey since they fulfilled the language requirement for either the Western or the Eastern Province, as well as for the Northern Province. 43% of the interviewers in the Northern Province were females while their share was at 40% in the remaining two provinces. The mean age of the interviewers in the Northern Province was lower (29.1 years) compared to the second (32.8). Drawing on the four hypotheses of this paper, the

[3] In theory, four members per saving group with a total of 533 savings groups would result in a dataset of 2132 cases. However, there are some exceptions in which only three savings group members could be interviewed, or the interviewer teams lost either the records for the household and/or the individual interviews so that the files could not be merged.

independent variables are the GOI, the GOR and residence of the respondents (i.e., whether a respondent lives in a village or in a provincial town).

Comparison between respondent and interviewer characteristics. When comparing respondent and interviewer characteristics (see Figure 1), respondents were more likely to be female. Regarding schooling, all interviewers had to have completed secondary education as a prerequisite for being recruited by the survey firm. Nevertheless, most respondents surveyed attended school. When comparing the respondents between the two parts of Wave 1, the respondents from the later one were slightly more likely to live in urban areas (+ 4%), as well as more likely to be less educated (80% instead of 90% of respondents interviewed in the Northern Provice had ever attended school). The respondents were predominantly females (72% in the Northern Province and 83% in the remaining two Zambian provinces) as international funders and their partnering, local non-government organisations often target to train females to form savings groups as a form of empowerment.

In addition,

Table 1 highlights the social distance dimension between interviewers and interviewees regarding residence in urban or rural areas. All the interviewers were residing in Lusaka, the capital, with only a few exceptions living in other towns like Kabwe or Livingstone. Although 15% selected savings groups member in the Northern Province (conducted in July) and 19% in the Eastern and Western Provinces (conducted in October and November) were living in provincial towns, the respondents could at most be considered as living in semi-urban areas as these provincial towns are not comparable with the capital in any aspect of infrastructure, e.g., regarding tarmac roads, number of supermarkets or banks. These data underline empirically the relevance of the third hypothesis on social distance.

Table 1. Respondent and interviewer mean characteristics

Variable	**Northern Province**				**Eastern and Western Province**			
	Respondents (n = 848)		Interviewers (n = 30)		Respondents (n = 1203)		Interviewers (n = 25)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Female	.72	.45	.43	.50	.83	.38	0.40	0.50
(Semi-) urban	.15	.36	1.00	0.00	.19	.39	1.00	0.00
Ever attended school	.93	.26	1.00	0.00	.80	.40	1.00	0.00

Due to the major differences in residence with between respondents from rural and semi-urban areas and interviewers residing in the capital, I tested the categorical independent variables for independence to qualify the rigour of the quasi-interpenetrated assignment. The random allocation between interviewers and interviewees aims to prevent interviewers from choosing their respondents at their own discretion and therefore, from selecting respondents due to their inherent preference. Without the interpenetrated design, there is reason to expect that certain interviewers choose certain respondents. For example, women might prefer female interview partners as they can better relate to the same sex. As a result, the GOI and the GOR would not be independent. Table 2 reports the results for the Pearson chi-squared tests of independence between the GOI and the GOR and, respectively, the GOI and the social distance with regard to residence. Savings group members in the Western Province lived both in rural and semi-urban areas. In contrast, all selected respondents in the Eastern Province live in villages. Therefore, the test for independence considers only the respondents in the Western Province.

Table 2. Tests of independence between independent variables

	Northern Province		Eastern and Western Province	
Gender of respondent	Male interviewers	Female interviewers	Male interviewers	Female interviewers
Male	28.69	26.11	17.75	16.70
Female	71.31	73.89	82.25	83.30
Pearson chi-square test	χ^2 = .6888		χ^2 = .2297	
Area of residency	Male interviewers	Female interviewers	Male interviewers	Female interviewers
			Only for Western Province	
Rural	85.86	83.89	38.41	40.00
Semi-urban	14.14	16.11	61.59	60.00
Pearson chi-square test	χ^2 = .426		χ^2 = .0965	

***** p< .001, ** p<.01 and * p< .05*

As attested by the Pearson chi-square test statistics, the null hypotheses of independence are not rejected, i.e., the GOI and the GOR and, respectively, the GOI and the area of residency are independent from each other. These results provide convincing evidence for the success of the randomisation of interviewer to interviewees judged on the observable characteristic of the GOI and the GOR, as well as the area of residency.

Dependent variables. To test whether the GOI introduces a systematic bias to survey response according to the four hypotheses of this paper, I rely on a set of knowledge, behavioural, and attitudinal questions as dependent variables (see Table 3 for exact question wording):

1. Objective, dichotomous questions such as whether the respondent attended a village or community meeting in the last 12 months, whether the respondent requested or gave out a loan in the last 12 months, and whether the respondent has ever heard of mobile money services;
2. Classical attitudinal questions measured on a four-point scale (“completely”, “much”, “a bit”, “not at all”) on how much trust the respondent places in certain (financial) institutions and trust in their neighbours as used as a reference category.

As this survey topic focuses mostly on financial behaviour, the battery questions on trust were the only coherent set of attitudinal questions from the questionnaire with more than 750 questions (including filter/ conditional questions)[4]. To match those five selected, attitudinal questions, I chose four knowledge and behavioural questions which all respondents had to answer, and which are general and not too specific to the savings group context. The question on village meeting attendance was modelled after the approach by Himelein to construct “a pseudo-comparison group” for the attitudinal questions (2016: 6) as factual items are generally assumed to be less prone to interviewer effects. To conclude, the questions for analysis were chosen to maintain, as much as possible, the original sample size to ensure statistical power as they were asked to all respondents.

[4] The administration of the survey took on average between 40 and 60min depending on the number of filter questions being answered as “yes” triggering more follow-up questions. The data collection software used did not allow for an exact measure of interviewer duration.

Table 3. Question wording and answer options of questions for analysis

Type of question	Question wording	Answer options
Behavioural	Have you requested any such loans in the last 12 months?	(yes/no; dichotomous)
Behavioural	Have you given out any such loans in the last 12 months?	(yes/no; dichotomous)
Knowledge	Have you ever heard of mobile money (MM) services, like Airtel Money, MTN Money and Zoona?	(yes/no; dichotomous)
Behavioural	Have you attended a village meeting in the last 12 months? (yes/no; dichotomous)	(yes/no; dichotomous)
Attitudinal	Item battery: To what degree do you trust…	
	… government banks?	Answer categories of item battery: completely, much, a bit, not at all
	… private banks?	
	… microfinance institutions?	
	… non-government organisations (NGOs) in general?	
	… your neighbours?	

Results

Graphical analysis of main and interaction effects

This section starts with a graphical analysis of the GOI effect. The second part of the section encompasses the presentation of the statistical models to test the four hypotheses developed in the previous section. As the systematic GOI effect per item manifests itself in the unequal response distributions of yes vs. no answers depending on female or male interviewer, the first two graphs depict those responses per item in bar charts with their respective confidence intervals. In absence of a GOI effect, the bars differentiating between female and male interviewers are expected to have the same bar length as the random assignment of the quasi-interpenetrated design would render no difference in the response distribution between male and female interviewers.

Erreur ! Source du renvoi introuvable. depicts the fraction of those respondents separated by the GOI who answered “yes” to the knowledge and behavioural questions. More respondents affirm to female interviewers that they requested a loan in the last 12 months. On the contrary, more survey participants reported to male interviewers that they attended a village meeting in the last 12 months compared to their female counterparts. Regarding whether they gave out a loan in the last 12 months or their knowledge about mobile money, no clear GOI effect was detected as the confidence intervals slightly overlap.

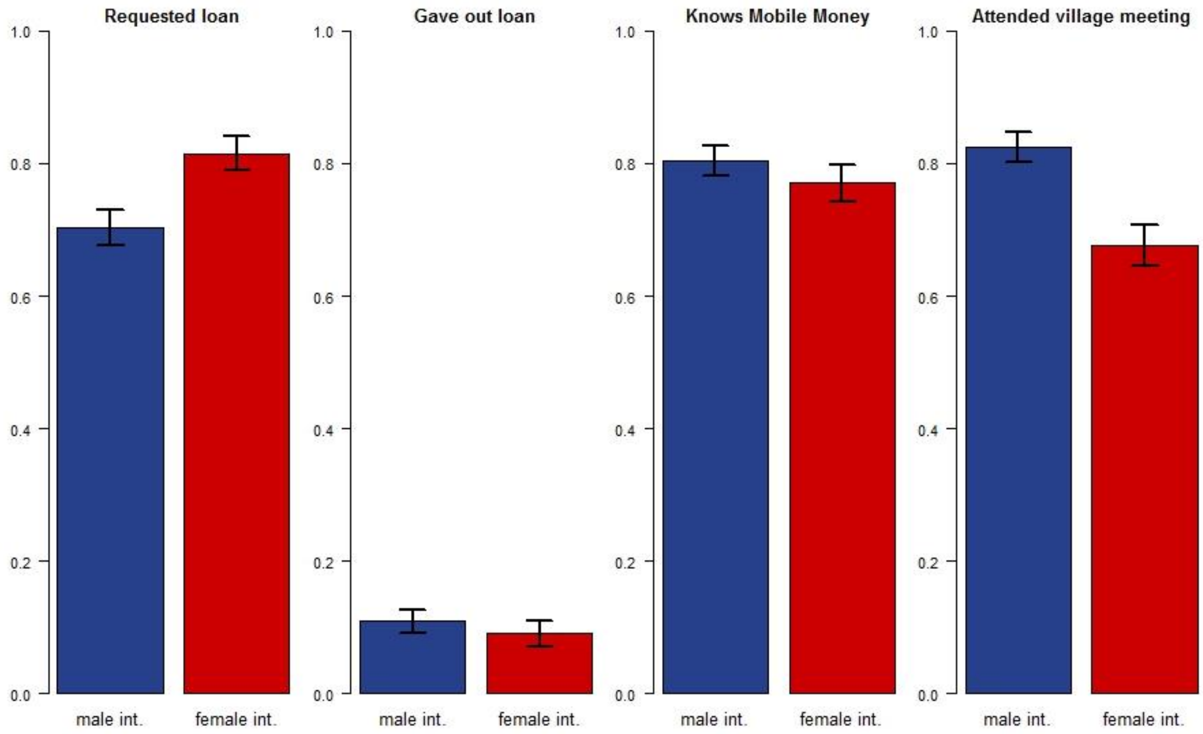

Figure 1 The GOI effect in knowledge and behavioural questions

The following graphs below distil the GOI effect depending on the gender-of-respondent (GOR) to demonstrate whether the GOI effect is driven by the GOI only or whether male and female respondents are susceptible to the GOI effect differently by comparing the bars of male and female respondents separated by male and female interviewers. Regarding the detected GOI effect for the question on having requested a loan in the last 12 months, there seems to be no interaction between the GOI and the GOR as all confidence intervals overlap. So regardless of the respondents' gender, they would disclose this financial behaviour more willingly to female interviewers than to male ones. Potentially requesting loans is considered as a sensitive topic so that respondents feel more comfortable to speak to women.

For the question on having attended village meetings, it becomes clear that the main effect is driven by female respondents being more susceptible to the GOI. Concretely, women reported more frequently on their village meeting attendance towards male interviewers to whom they provide more socially desirable answers. The interaction between the GOI and the GOR becomes apparent with the knowledge question on mobile money. Female and male respondents reported more often about their awareness of this financial service to male interviewers than to their female counterparts. Certainly, the socially desirable answer is to appear knowledgeable.

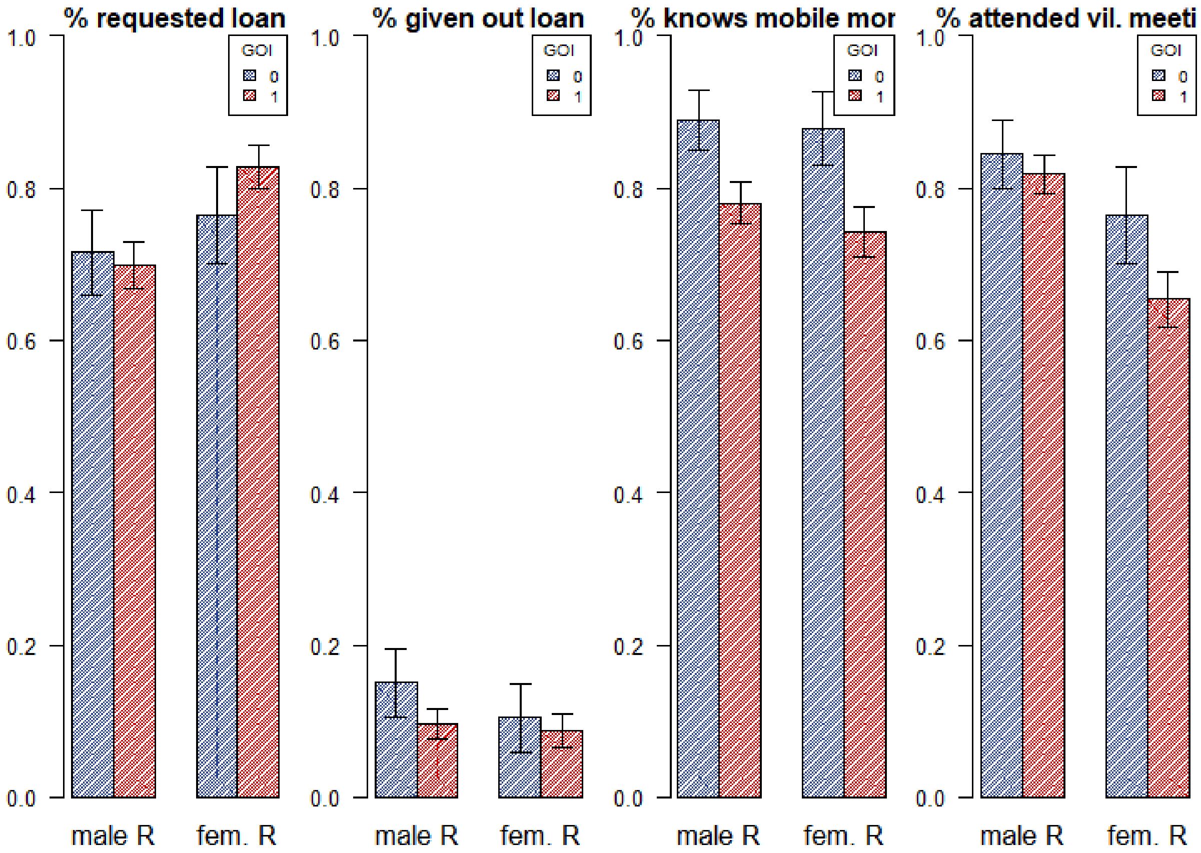

Figure 2 The GOI effects in knowledge and behavioural questions depending on gender of respondent (with 0/blue signifying male respondents and 1/red signifying female respondents)

Regarding attitudinal questions, the items showing an GOI effect are the trust levels in private banks and microfinance institutions (MFIs). More respondents affirm to trust those two financial institutions at least on some level (one out of the three response options of “a bit”, “much”, “completely”) when asked by a female interviewer in contrast to a male one.

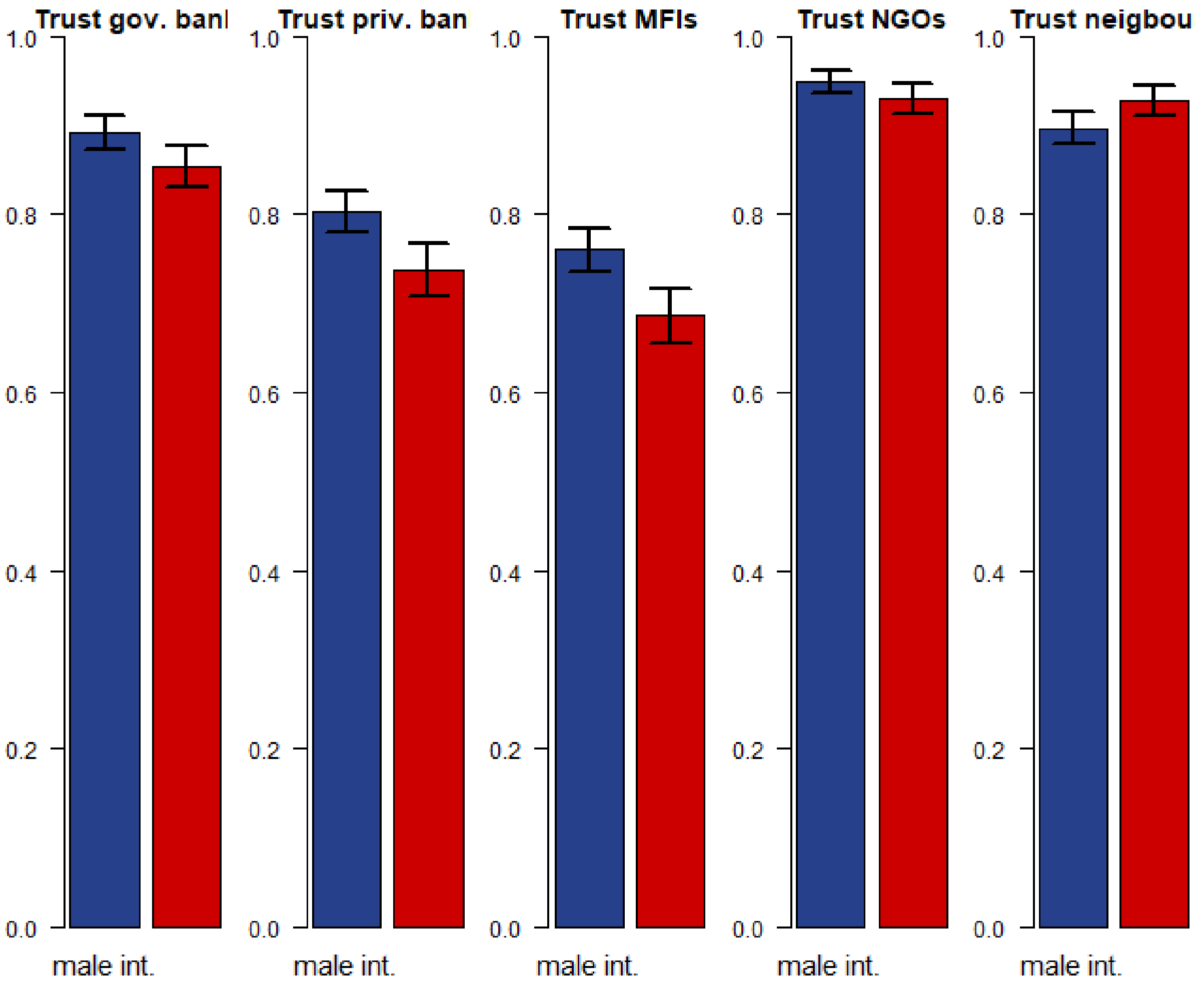

Figure 3 Interviewer effects in attitudinal questions

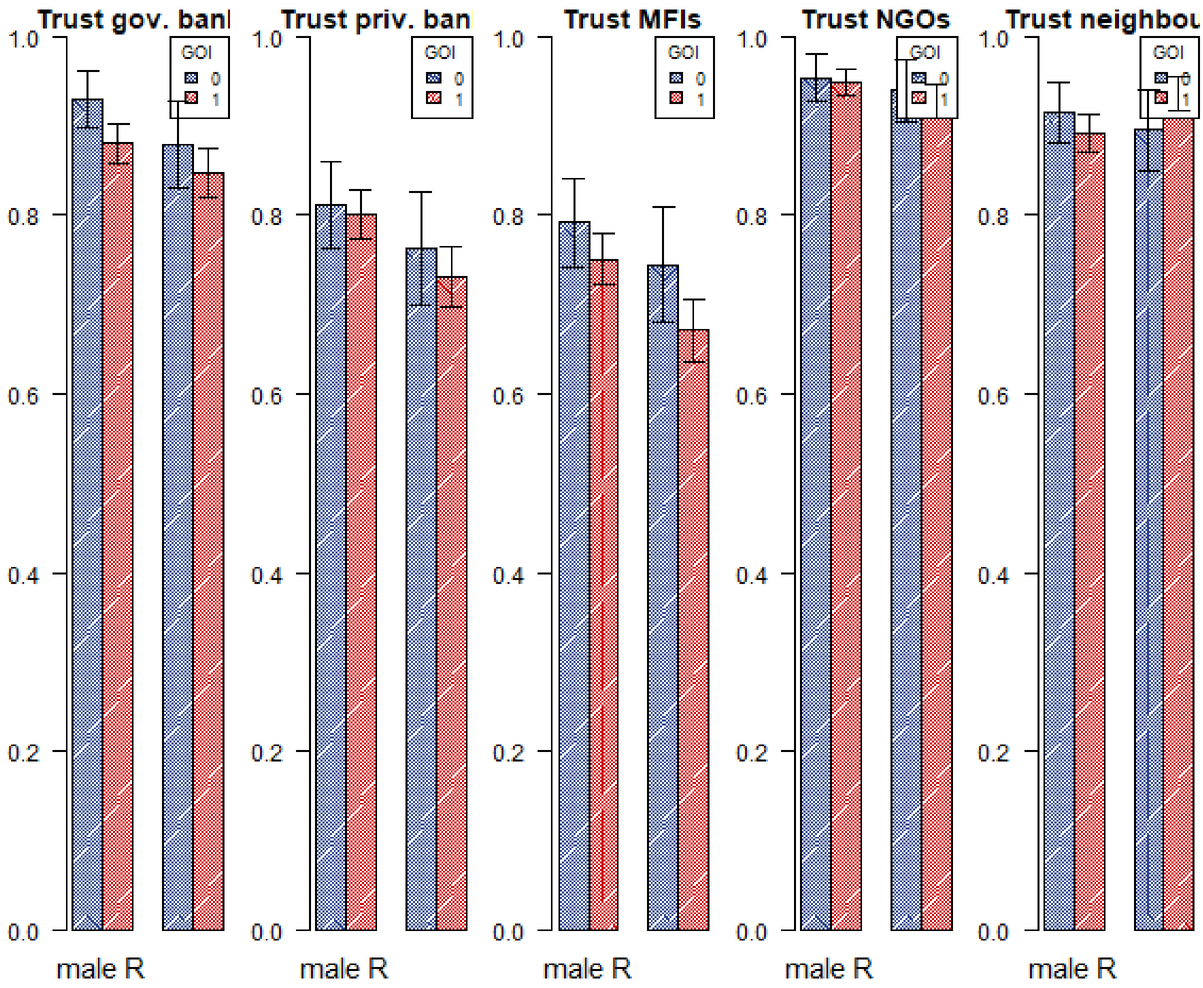

Figure 4 The GOI effects in attitudinal questions depending on gender of respondent (with 0/blue signifying male respondents and 1/red signifying female respondents)

Surprisingly, no interaction between the GOI and the GOR can be detected for attitudinal questions as all the confidence intervals overlap. The GOI effect is the main driver for those selected trust questions, suggesting that respondents more willingly

Using the multivariate analysis of variance (MANOVA) method

This paper uses the multivariate analysis of variance (MANOVA) method to group four knowledge and behavioural items and five attitudinal items together in a linear combination to address the issue of multiple testing (Haase and Ellis, 1987). Due to the trade-off between solving the problem of multiple testing and accounting for more control variables, previous studies on interviewer effects, however, opted for multivariate (logit or probit) regressions to the expense of the multiple testing problem (e.g., Flores-Macias & Lawson, 2008; Himelein, 2016; Liu & Stainback, 2013). As the study design is a randomised experiment based on a quasi-interpenetrated assignment of interviewers (i.e., each interviewer per group was randomly assigned to a selected savings group member), the dataset allows for a rigorous examination of the GOI free of usual selection bias induced by interviewers when choosing the respondent by themselves. The random assignment between interviewers and respondents can be understood as the GOI being a treatment which is randomly allocated to a random subset of the population (here, selected savings group members). Consequently, the main explanatory variable of interest, the GOI, is a binary or dichotomous variable (0 for male, 1 for female interviewers).

To address the dichotomous nature of the dependent variable in the form of yes-no-questions - an important assumption of MANOVA, the four dependent variables on the behavioural and knowledge items are converted using the arcsine-square root transformation: *t(x) = arcsine (squared(x))* following standard practices (Jaeger, 2008). The attitudinal questions regarding level of trust for certain institutions are not transformed as they are on a four-point scale (“completely”, “much”, “a bit”, “not at all”). To test null hypotheses from a probability distribution, Wilks’ lambda as one key statistic of MANOVA is translated into a variable following an approximate F distribution following a Rao transformation. If the obtained F-statistics exceed the critical value for $p < .001$, the null hypothesis can be rejected.

Table 3 summarises the models and hypotheses analysed in this paper. To confirm the first hypothesis on a systematic GOI effect, the F-statistics of the main effect of the GOI should exceed the critical value at a significance level of $p < .001$. To investigate the second and third hypotheses, F-statistic for the interaction term between the GOI and respectively the GOR or the area of residence should exceed the critical value at a significance level of $p < .001$. Instead of a null hypothesis test, the fourth hypothesis is tested by comparing whether the F-statistics for the GOI is consistently higher in the model with the knowledge and behavioral questions compared to one with the attitudinal items.

Table 4. Statistical models testing proposed hypotheses

Hypotheses	**Model**	**Null hypotheses of interest tested**	**Decision criteria**
H1: A systematic gender-of-interviewer effects exists in this survey.	$Y_{ij,q} = \mu_{j,q} + \alpha_{j,q} + \varepsilon_{ij,q}$	1. $\mu_{1,q} = \mu_{2,q}$	$F_{A,q}$ > critical F for p < .001
H2: The gender-of-interviewer is different for female versus male respondents (i.e., the interaction between the GOI and the GOR).	$Y_{ijh,q} = \mu_{j,q} + \alpha_{j,q} + \beta_{h,q}$ $+ \beta_{h,q}*\alpha_{j,q}$ $+ \varepsilon_{ijh,q}$	1. $\mu_{1,q} = \mu_{2,q}$ 2. Simple effects of the GOI are consistent across female and male respondents; or "Simple effects of factor A are consistent across all levels of factor B" (Huitema, 2011: 490)	$F_{A,q}$ > critical F for p < .001; $F_{B*A,q}$ > critical F for p < .001
H3: Gender-of-interviewer effects are stronger for rural interviewees compared to	$Y_{ijg,q} = \mu_{j,q} + \alpha_{j,q} + \gamma_{g,q}$ $+ \gamma_{g,q}*\alpha_{i,q}$ $+ \varepsilon_{ijg,q}$	1. $\mu_{1,q} = \mu_{2,q}$ 2. Simple effects of the GOI are consistent across rural and semi-urban respondents	$F_{A,q}$ > critical F for p < .001;

(semi-) urban ones.		$F_{C*A,q}$ > critical F for p < .001
H4: Gender-of-interviewer effects are more likely to be significant for attitudinal questions compared to knowledge and behavioural ones.	$Y_{ij,f} = \mu_{f} + \alpha_{j,f} + \varepsilon_{i,f}$ versus $Y_{ij,a} = \mu_{a} + \alpha_{j,a} + \varepsilon_{i,a}$	$F_{A,a} < F_{A,f}$ for all models

Legend**:**

i is the index denoting the *i*-th subject;

j is the index indicating the GOI (j=1 stands for male interviewers and j=2 for female interviewers);

h is the index for the GOR (h=1 stands for male respondents while h=2 for female respondents);

g is the index for the area of residency (where g=1 represents rural areas and g=2 semi-urban areas);

$Y_{ijhg,f}$ is the vector comprising the dependent variable score for the *i*-th subject in the *j*-th treatment (here the treatment is the GOI) for either q=f standing for the set of knowledge and behavioural questions or q=a standing for the set of attitudinal questions;

μ_{q} is the vector comprising the overall population means (i.e., the mean of the individual population means) of each dependent variable for either q=f standing for the set of knowledge and behavioural questions or q=a standing for the set of attitudinal questions;

$\alpha_{j,q}$ is the effect of GOI, where j=1 if the interviewer is male and j=2 if the interviewer is female for either q=f standing for the set of knowledge and behavioural questions or q=a standing for the set of attitudinal questions;

$\beta_{h,q}$ is the effect of the GOR, where h=1 if the respondent is male and h=2 if the respondent is female for either q=f standing for the set of knowledge and behavioural questions or q=a standing for the set of attitudinal questions;

$\gamma_{g,q}$ is the effect of the area of residency, where g=1 if the respondent is living in rural areas and j=2 if the respondent is living in semi-urban areas for either q=f standing for the set of knowledge and behavioural questions or q=a standing for the set of attitudinal questions; and

$\varepsilon_{ijhg, q}$ is the error component associated with the *i*-th subject for the respective models containing $\alpha_{j,q}$, $\beta_{h,q}$, $\gamma_{g,q}$ and for either q=f standing for the set of knowledge and behavioural questions or q=a standing for the set of attitudinal questions;

$F_{A,q}$ denotes the F-statistic for the main effect of the GOI for either q=f standing for the set of knowledge and behavioural questions or q=a standing for the set of attitudinal questions;

$F_{B*A,q}$ denotes the F-statistic for the interaction effect of the GOI and the GOR for either q=f standing for the set of knowledge and behavioural questions or q=a standing for the set of attitudinal questions;

$F_{C*A,q}$ denotes the F-statistic for the interaction effect of the GOI and the area of residence for either q=f standing for the set of knowledge and behavioural questions or q=a standing for the set of attitudinal questions.

Table 5 shows empirically the underlying connection or the system of the five items. As to be expected, all five attitudinal questions have a statistically positive correlation among each other. In other words, the level of trust the respondent attaches to one item, such as government banks, is correlated with the level of trust he or she reports for another item, such as private banks. Consequently, the correlation coefficients are the highest among financial institutions such as government banks, private banks, and microfinance institutions. In contrast, the correlation coefficients attached to neighbours and the remaining four items are the lowest since neighbours are not regarded as institutions.

Table 5. Matrices for dependent variable intercorrelation

	Dependent variables			
Binary knowledge and behavioural questions	Requesting loan	Giving out loan	Mobile money knowledge	Attended village meeting
Requesting loan in last 12 months		- .009	.081***	- .046*
Giving out loan in last 12 months			.125***	- .015
Ever heard of mobile money				- .006
Attended village/ community meeting in last 12 months				

Ordinal attitudinal questions regarding trust towards …	Government banks	Private banks	MFIs	NGOs	Neighbours
Government banks		.570***	.459***	.278***	.160***
Private banks			.635***	.269***	.193***
Microfinance institutions (MFIs)				.323***	.267***
Non-government organisations (NGOs)					.215***
Neighbours					

**** p< .001, ** p<.01 and * p< .05*

Model results

As detailed in the previous section, the data analysis relies on one one-factor and two two-factorial MANOVA (model 1, and respectively model 2 and 3) to test the proposed four hypotheses and to overcome the multiple testing problem due to four knowledge and behavioural questions and five items of an attitudinal questions.

Systematic GOI effect. In all six models specified in Table 3 for attitudinal, knowledge and behavioural questions respectively, a systematic GOI effect could be detected according to the high and significant level of the F-statistics which confirms hypothesis 1. Regardless of adding more independent variables such as gender of respondent or location of interview, the systematic GOI effect persists.

No GOI-GOR interaction. However, no interaction between the GOI and the GOR could be established in any of the three models, neither for the attitudinal questions nor for the knowledge and behavioural questions (hypothesis 2 not confirmed). Therefore, it can be concluded that the GOI effect is equal for female and male respondents although the answers between female and male respondents (significant coefficient for the GOR variable) are statistically different.

Table 6 Summary of F-statistics (based on Wilks' lambda) and level of significance for main and interaction effects in knowledge and behavioural questions with arcsine-root transformed dependent variables

	Model 1		**Model 2**		**Model 3**	
	Effect size	F statistic	Effect size	F statistic	Effect size	F statistic
Overall model		26.22***		12.81***		29.82***
GOI	0.0494	26.22***	0.0242	12.50***	0.0234	12.10***
GOR			0.021	10.28***		
GOR x GOI			0.0041	2.06		
Urban					0.1086	61.47***
Urban x GOI					0.0038	1.91

**** p< .001, ** p<.01 and * p< .05; effect sizes calculated as $\eta^2 = 1 - \lambda$ using Wilk's λ*

Effect of social distance for attitudinal questions. The third hypothesis focuses on the social distance beyond gender measured by a dummy variable capturing whether the interview was done with rural or urban respondents. Regardless of type of question, the coefficient for location of interview is significant which leads to the conclusion that responses given by urban respondents are statistically different than the ones of rural respondents. Assessing whether there is an interaction between the GOI and location, the statistical difference is only maintained for attitudinal questions but not for knowledge and behavioural ones (hypothesis 3 confirmed only for attitudinal questions).

Type of question. Regardless of question type, the single effects for the GOI, the GOR and location, as well as the interaction between the GOI and the GOR behave similarly for attitudinal, knowledge and behavioural questions regarding whether the effect attains statistical significance or not. However, the interaction between location and the GOI is an exception, and the location of the interview moderates the GOI effect for attitudinal questions. This result is comparable to the findings for attitudinal questions on women's rights and abortion in Mexico where male respondents in Mexico City behave differently than their counterparts in the rest of the country even though the Mexican study only analysed attitudinal questions and not factual questions (Flores-Macias & Lawson, 2008). Furthermore, it is notable that both the effect sizes and the F-statistics for the GOI coefficient are larger for knowledge and behavioural questions than for attitudinal questions. Larger F-statistics indicate larger confidence in the statistical difference and a greater chance of the true effect being non-zero for knowledge and behavioural questions. This finding goes against the expectation of the fourth hypothesis and thus, against earlier studies which demonstrate that attitudinal questions are more prone to interviewer effects.

Table 7 Summary of F-statistics (based on Wilks' lambda) and level of significance for main and interaction effects in attitudinal questions

	Model 1		**Model 2**		**Model 3**	
	Effect size	F statistic	Effect size	F statistic	Effect size	F statistic
Overall model		7.52***		4.29***		6.33***
GOI	0.0181	7.52***	0.0128	5.28***	0.0218	9.08***
GOR			0.0105	4.34***		
GOR x GOI			0.0017	0.68		
Urban					0.0152	6.28***
Urban x GOI					0.0117	4.82***

**** p< .001, ** p<.01 and * p< .05; effect sizes calculated as $\eta^2 = 1 - \lambda$ using Wilk's λ*

To conclude, a systematic GOI effect can be detected in this Zambian survey in attitudinal, knowledge, and behavioural questions. It is important to note that this analysis does not account that the interviews are nested in interviewers. However, considerably high inter-class correlations among the interviewers across selected attitudinal questions presented in the subsequent chapter[5] suggest that further analysis of interviewer effects demands the multilevel modelling accounting for the nesting of respondents within interviewers.

Discussion

Randall et al. put forward the dilemma for interviewers from sub-Saharan cultures as follows: "they are often caught between two different value systems: their professional position and their role as a culturally attuned member of their society. The former requires them to ask

[5] Analyses show that the intra-class correlations for interviewers for attitudinal items about trust in different institutions vary between 0.225 to 0.337.

questions in a specific way, often on subjects which are rarely spoken about openly and where power relations between interviewee and interviewer may influence acceptable responses in different, but unknown ways" (2013: 784). In addition, "survey concepts are often different from local concepts and […] interviewers are socio-economically distant from respondents" (2013: 764). Due to these generic reasons, interviewer effects will remain as a concern in any LMIC survey despite advances in technology for monitoring interviewers becoming more sophisticated (see Lupu and Michelitch, 2018).

Relying on a face-to-face survey with more than 2000 Zambian respondents, this study demonstrates that there is a systematic GOI effect both in knowledge, behavioural, and attitudinal questions although the effect sizes are larger there is a greater chance of the true effect being non-zero in the former than in the latter. These results are surprising as they contradict previous findings (see e.g., West and Blom, 2017 or Himelein, 2016) on interviewer effects being larger in subjective compared to objective questions. This contrast poses questions on whether socially desirable answering regarding socially desirable behaviours (such as attending village meetings) follow different dynamics in specific context such as in this Zambian survey.

This Zambian study demonstrated that the divide between rural and semi-urban plays a role on how the GOI effect can manifest. Even though the responses between urban and rural residents, as well as female and male respondents are statistically different, these differences are independent of the GOI. The only exception is that an interaction effect between location and the GOI was detected for attitudinal items but not for knowledge and behavioural ones. While there is no evidence of an interaction between the GOI and the GOR, the effects of the latter are susceptible to the urban vs. rural residency of the respondents. As the GOI effect differs among knowledge, behavioural, and attitudinal questions, future studies on social desirability bias induced by the GOI need to be analysed regarding strength and direction on the level of question item per topic.

More specifically, this paper shows that it is insufficient to consider only one dimension of the question characteristics when analysing the differential interviewer effect from one item to another. It means that the traditional distinction between factual and attitudinal questions or sensitive vs. non-sensitive is insufficient (for literature reviews on the impact of sensitive questions and interviewers see Schaeffer et al., 2010; West & Blom, 2017). Rather the intertwining nature between several dimensions of question characteristics (in this case factual sensitive items such as attending community meetings) needs to be taken into account. Certainly, the strategy of studying interviewer effects on items by combining different question characteristics is a more nuanced approach (see Schnell & Kreuter, 2005). As shown, however, in this paper, some question items with the same topic and characteristics (for example receiving versus giving out a loan) do not produce a similar interviewer effect (while the GOI effect is present in the former item, it is absent for the latter). This finding points into the direction that social desirability bias needs to be studied on a case-by-case basis.

The concept of social desirability in cross-cultural comparisons remains to be complex without quick, straightforward remedies to survey researchers and practitioners concerned about interviewer effects in general, and about the GOI effect in particular. Johnson and van de Vijver study the differential impact of social desirability across different cultures as a source of measurement error and conclude that “social desirability is a personality characteristic with an influence on what a respondent wants to transpire in a survey” (2003: 203). In other words, this personality characteristics can be triggered differently by various factors, such as the characteristics of both interviewer and interviewee, as well as different question topics. However, this study could demonstrate that the GOI effect is important also in a Zambian context on survey topics which do not seem to invoke social desirability at the first look.

The findings presented here have several implications for social surveys. Firstly, they encourage researchers to enlarge the scope of the GOI effect analysis for knowledge and behavioural, as well as attitudinal items which have yet to be studied apart from evidently gender-sensitive question topics. Also, the GOI effect induced by the social distance between interviewer and respondents should be identified and controlled for whenever suspected theoretically.

To conclude, this paper provides additional evidence on the GOI effect using data with a random assignment between interviewer and interviewee. The research both echoes and complements the results of previous relevant studies by comparing the GOI effect between knowledge and behavioural questions on financial behaviour and civic engagement and attitudinal questions on trust. Due to the analysis of multiple questionnaire items, the analysis addresses the multiple testing problem by using MANOVA and shows graphically how the GOI effect differs between knowledge and behavioural vs. attitudinal questions. Finally, the findings suggest the social desirability and interviewer effects are specific to the question topic rather than to the question type. Most importantly, this paper concludes that sensitive questions and socially desirable answers are not limited to topics such as abortions or sexual violence. Thus, researchers need to investigate interviewer effects, and among them, the GOI effect, to ensure survey data quality and improve questionnaire design to tackle the issue of sensitivity and social desirability. For example, researchers started to test new abortion questions and their functionality through survey experiments (placement of question items within a survey, the level of detail of the question, etc.) (see e.g., Lindberg et al., 2022; Mueller et al., 2023). This approach should apply beyond highly sensitive question topics to also seemingly less sensitive topics.

Chapter Two: Revisiting the interviewer effects in a multilingual context

Abstract

Survey methodologists and researchers using interviewer-administered survey data show a never-ending interest in investigating the impact of interviewers on survey data quality. This applies when studying survey data collected in high-income countries or in low- and middle-income countries (LMICs). In LMICs, interviewer-administered surveys remain the principal data collection tool for representative surveys mainly due to lower levels of literacy. This study marks the beginning of a strand of research accounting for the complex multilingual context by demonstrating a link between interviewer effects and different modes of translation and translation quality.

It documents the decrease in interviewer effects by making use of the stepwise improvements in translation quality of a longitudinal questionnaire fielded in a three-wave household panel in Zambia between 2016 and 2019. The survey began with on-the-fly translation from the source questionnaire in English. In the following wave, a written questionnaire script was provided which was translated in-house by the local field organisation in the specific *lingua franca* of each of the three study regions. In the last wave, questionnaire translation quality was assured by relying on a team-based, iterative translation process accompanied by on-site pre-testing in the study regions. Due to the quasi-interpenetration of interviewer assignments to respondents, interviewer effects in the form of response variance attributable to interviewers is measured using intra-cluster correlations (ICCs) from cross-classified multilevel regressions.

The first findings show that the non-negligible interviewer effects for questions on trust in various institutions were not explained by adding respondent-level and subsequently interviewer-level characteristics into the cross-classified multilevel regressions. However,

interviewer effects reduce dramatically for the later waves when a scripted translation was available to the interviewer, limiting the necessity of on-the-fly translation.

Introduction

Survey methodologists and researchers using interviewer-administered survey data show a never-ceasing interest in investigating the impact of interviewers on survey data quality – both while studying survey data collected in high-income countries (Olson et al., 2020; West and Blom, 2017) or low- and middle-income countries (LMICs). In LMICs, interviewer-administered surveys were the principal data collection tool before the outbreak of the COVID pandemic (see Lupu and Michelitch, 2018 for evidence in political science) and have remained so even in the post-COVID period (Frankovic et al., 2023 for opinion polls around the world). Thus, researchers coming from development economics, demography, or political science have recently devoted more attention to interviewer effects (Di Maio and Fiala, 2020; Leone et al., 2021; Sundström and Stockemer, 2022). Before this surge, research on interviewer effects in LMICs was rather periodical and often only focused on single, observable interviewer characteristics (instead of investigating interviewer effects in form of the share of response variance attributable to interviewer), such as the sex or the religious attire of the interviewer (see for example Axinn, 1991; Becker et al., 1995; Blaydes and Gillum, 2013; Flores-Macias and Lawson, 2008; Liu and Stainback, 2013; Liu and Wang, 2016).

Analysing qualitative data in the form of published definitions, interviewer manuals and qualitative interviews with respondents, data producers and consumers in Burkina Faso, Senegal, Tanzania, and Uganda, Randall et al. (2013) demonstrated how "the power of the interviewer" plays a role in measurement error in the sub-Saharan context where interviewer-administered surveys are the norm. Lower levels of literacy rates, especially among the older population, as well as lacking spread of internet and phone networks to rural areas and poorly functioning postal services has impeded the use of other modes of administration (phone, web, paper questionnaire) until the recent COVID-19 outbreak. Representative surveys in LMICs with low literacy rates, especially among higher age groups, might still have to rely on interviewer efforts even though the mode of administration has changed from face-to-face to phone interviews.

Due to the significant role of interviewers in LMICs, interviewer effects research in LMIC contexts has been flourishing since 2021, with more than ten studies published in demographic and development economic journals. Yet literature has not examined the effects of multilingual contexts in a comprehensive manner despite multiple LMICs being multilingual in nature. Specifically, the question arises of how survey data quality might be affected by the decision on whether, in which local languages and through which translation method the questionnaire is translated. Even large international survey programs do not provide sufficient information to answer this question (Hughes and Lin, 2018).

This study marks the beginning of research accounting for complex multilingual contexts as a source of survey error by demonstrating a link between interviewer effects and different modes of translation and translation quality. It investigates the relationship between translation quality and interviewer effects by making use of stepwise improvements in the translation quality of a longitudinal questionnaire fielded in a three-wave household panel in Zambia between 2016 and 2019. The survey started with oral translation from the source questionnaire in English followed by providing a written questionnaire script translated in-house by the local field organisation in the specific *lingua franca* of each of the three study regions. In the last wave, questionnaire translation quality was assured by relying on a team-based, iterative translation process accompanied by on-site pre-testing in the regions. Due to the quasi-interpenetrated assignment of interviewers to respondents, interviewer effects in the form of response variance attributable to interviewers is measured using intra-cluster correlations (ICCs) from cross-classified multilevel regressions. The results show a clear reduction in interviewer effects for the last wave when the questionnaire translation was produced through the best practice of team-based translation.

Adapting the role of the interviewer to a multilingual context

Survey research is still in its initial stages regarding the exploration and investigation of the impact of multicultural, multi-ethnic, and especially multilingual contexts which may be relevant in LMIC contexts. The book "*The essential role of language in survey research*" (Sha and Gabel, 2020) and relevant chapters in the book "*Advances in comparative survey research*" (Johnson et al., 2019) can be regarded as a starting point in filling this gap as previous research focusing on multicultural contexts has either focused on cross-national comparisons or studied exclusively bicultural and bilingual survey contexts (Peytcheva, 2020).

However, the nature of multilingualism in LMICs, especially in Sub Saharan Africa, complicates the interviewing process as respondents and interviewers speaking multiple languages with potentially various levels of proficiency lead to situations and challenges of multiple language combinations for communication. For the Afrobarometer Round 6, a face-to-face survey conducted between 2014 and 2015 in 36 African countries, Lau et al. (2020) found a diverse mix of language interview situations. Without standardised and clear rules for the choice of the interview language, leading to questions regarding who decides on the survey language and for what reasons when several language combinations among respondent and interviewer are possible, the authors reported that the choice of interview language was decided by each interviewer-respondent pair. Even in cases when the respondent and interviewer shared their first or home language, certain pairs still opted for another common language instead of the shared first language. Due to data constraints and the specific focus of the study, the authors do not touch on the issue of oral translation in the field done using the interviewer on-the-fly technique (also referred to as on-the-fly translation by survey practitioners working in LMICs).

However, as early as in 2009, Weinreb and Sana pointed out that it is common for interviewers in LMICs to translate a questionnaire during field work "on the fly", meaning without a scripted, standardised questionnaire translation – a practice "rarely acknowledged in the literature but quite common in the field" (2009: 249). Hughes and Lin (2018) discussed the diverse and unstandardised approaches of major cross-national surveys in LMIC contexts in their decision of how many of and which local languages the source questionnaire must be translated to. Except for the Demographic and Health Surveys which were translated into any local languages spoken by more than 10% of a sample, the remaining survey programs evade set guidelines for the decision of questionnaire translation while some either rely partly on on-the-fly translation from the start (such as the Living Standards and Measurement Study) or even replace respondents for whom the pool of interviewers is unable to offer to administer the questionnaire in the respondent's language of choice (such as the Afrobarometer) (Hughes and Lin, 2018).

In multilingual contexts, on-the-fly translation is sometimes indispensable due to reducing the costs of translating a source language questionnaire (usually English, French, or the *lingua franca* as main languages of broader communication) (Weinreb and Sana, 2009). Randall et al. are especially concerned that "in the linguistic melting pot of African cities, interviewers often have to work in languages that they have only half mastered" (2013: 778) – a concern shared by Hughes and Lin (2018) for the broader linguistic context in sub-Saharan Africa and by Lau et al. (2018) for phonetically close but differently scripted languages in India.

Thinking from the perspective of the interviewer, on-the-fly translation reflects an additional cognitive task and includes behaviours going beyond translating the actual questions to giving clarification and/or probing in a non-scripted language through on-the-fly translations, and even back-translating answers given in a different language into the pre-defined survey language. In the lack of standardised rules around on-the-fly translation, the interviewer's discretion and interpretation heavily influence the quality of the questionnaire translation and therefore, survey data quality in general and the measurement of key demographic concepts in particular. Based on the example of Senegal where the census guidelines specify the definition of a household in the principal local languages, Randall et al. (2013) assume that such linguistic clarifications led to more precise measurement of household size and the fact that Senegalese households are the largest ones in Africa.

Evaluating oral translation – evidence from Germany

Although several research studies lament the lack of studying the impact of this multilingual context on survey data quality (Harkness et al., 2010; Hughes and Lin, 2018; Lau et al., 2020), the scarce literature brought forward two diverging opinions when it comes to how questionnaire translation can impact data quality through the interaction of multilingual respondents and interviewers. Certain studies demonstrated the pitfalls of on-the-fly translation either using data from the Demographic and Health Survey in 1998 in Kenya (Weinreb and Sana, 2009) or from a study in 2005 assessing data quality for Swiss and German interviewers orally translating an English source questionnaire (Harkness et al., 2008). A pre-translated, scripted questionnaire helps the interviewer to read out the question verbatim, eliminating the task for the interviewer to orally translate the questions on the fly while administering a complex standardised questionnaire.

Using excerpts of transcribed interviewer-respondent-interactions, Harkness et al. (2008) demonstrated diverse ways of deviating from standardised interviewing (e.g., simplifications or omissions in administering the complete survey questions or a failure to probe). The availability of a scripted questionnaire translation reduces interviewer burden and interviewers' varying ability to translate orally may then be connected to reduced interviewer effects (Harkness et al., 2008). In contrast, Bignami-Van Assche et al. (2003) downplay the importance of questionnaire translation as they argue that careful selection, intensive training, and close supervision of interviewers are adequate compensators. However, the authors' study design is inadequate to investigate translation quality, and their deliberations are superficial. For example, they neglect the fact that the resources necessary for sufficient preparation and training to ensure the high-quality delivery of on-the-fly translation might even offset any arguments of cost reduction from limiting the translation into local languages (Harkness et al., 2010).

Harkness and her co-authors strongly argue against the use of on-the-fly translation since the disadvantages outweigh the advantages (Harkness et al., 2008; Harkness et al., 2010). This firm opposition might be the reason that oral translation did not receive any attention from survey methodologists except for the study by Harkness et al. (2008) which demonstrates the short-comings and error sources when interviewers translated questions on-the-fly using data from Germany. The authors list as disadvantages the lack of standardisation, pretesting and documentation. Through analysing transcripts of interviews in which German speaking interviewers translating on-the-fly from English, Harkness et al. (2008) demonstrated the additional cognitive burden for interviewers working in more than one language which leads them to omit or simplify words during an on-the-fly translation or keeping their translation too close to the linguistic structure of the source language. "More frequent or more dramatic repairs than in scripted interviews, false starts, long pauses, self-discourse, and apologies are examples of" threats to survey data quality as consequences of the respondent's miscomprehension to wrong translations (Harkness et al., 2008: 239). The authors could also document that after an inaccurate on-the-fly translation, interviewers "repeat responses incorrectly, present answer categories in biased fashions, and they fail to probe and negotiate suitable responses." (Harkness et al. 2008: 244)

Clearly, the quality of the on-the-fly translation depends on the translation skills of the individual interviewer which introduces variation into interviewer performance and leads to the de-standardisation of the interviewing process and the delivery of the survey question. Harkness et al. showed, however, that on-the-fly translation systematically exhibited more errors in the form of deviations from the question meaning in the source language despite idiosyncratic differences across interviewers. Therefore, they argue that "[o]rally translated interviews introduce manifold differences that work against the principles of standardi[s]ation, the notion of carefully worded questions, and against good interviewing practice." (Harkness et al., 2008: pp. 248-249)

Based on the existing findings on interviewer effects, especially in LMICs, the objective of this study is to revisit interviewer effects and to investigate how different interviewer characteristics help to explain interviewer variability. Consequently, the first research question put forward for this paper is as follows: To what extent do interviewer effects influence responses in a multilingual face-to-face survey?

To answer this first question, we rely on three hypotheses.

H1. Different interviewers systematically collect different answers from a randomly selected sample.

H2. Even after controlling for respondent-level characteristics (age and gender of the respondent), systematic interviewer effects persist.

H3. Interviewer characteristics (age and gender of the interviewer) explain a part of the interviewer variance.

Conducting a household survey in a multilingual country such as Zambia

Background of study. The survey data used for analysis is part of an impact evaluation of a development program called Rural Financial Expansion Programme (RUFEP), funded by the International Fund for Agriculture (IFAD). RUFEP's objective was to improve the financial inclusion of Zambians living in rural or semi-urban areas through diverse services, for example by providing financial and technical support to local non-government organisations (NGOs). Those NGOs then offered training and other comparable capacity building activities to strengthen so-called savings groups in their activities revolving on savings, lending, and internal insurance schemes. Through those savings groups, Zambians who are located far from financial institutions, such as banks or micro-finance institutions, gained access to financial services to invest in their businesses and improve their livelihoods.

The impact evaluation targeted RUFEP's savings groups in three Zambian provinces and 8 districts: 3 districts in the Northern Province, 3 districts in the Eastern Province, and 2 districts in the Western Province. The survey sample consists of members of active and experienced savings groups. To study the impact of RUFEP on the beneficiaries' lives, the Zambian Savings Group Panel survey (ZamSaP) has been implemented to capture changes at the individual and household level of selected savings group members. We use the first wave of ZamSaP conducted in 2016 to test the first three hypotheses.

The language situation in Zambia. While English is the only official language, Zambia is home to 26 Bantu language groups exhibiting more than 80 varieties or dialects and with seven indigenous Bantu languages among them recognised as national languages (Marten and Kula, 2008). Those seven national languages are characterised as the Bantu languages "with orthographies based on the Latin alphabet" (Kaani and Joshi, 2023: 414). According to the 2000 Zambian census, more than half of the inhabitants speak Bemba, Nyanja, Tonga, or Lozi as their first languages while two-thirds use English, Bemba, or Nyanja as their secondary language. Therefore, Marten and Kula assess the country's language use as a "complex, dynamic multilingual" situation (2008: 296) with evidenced "multilingualism and code-switching, where speakers employ a number of different languages in different contexts" (2008: 298). Each Zambian province usually has a *lingua franca* or a "main language of wider communication" (Marten and Kula, 2008: 297).

Translation approach. Due to on-the-fly translation being a widespread practice in Zambia in 2016, the data collection firm conducting wave 1 provided no translation for the questionnaire into local Zambian languages. Thus, interviewers in wave 1 conducted the interview with on-the-fly translation. However, for the subsequent two waves, we recognised the strong need for standardising the questionnaire translation. Thus, the questionnaire was translated into the *lingua franca* of each of the three study provinces in Zambia from wave 2 onwards. In addition, we invested more resources for external, professional translators to translate the questionnaires from the pre-test the questionnaire translation in the field before revising the questionnaire translation in a team-based approach with the interviewers in wave 3. While in wave 2, the questionnaire translation was provided internally by the data collection firm.

With this setting of survey administration, we can address a second research question linking interviewer performance and translation quality: How interviewer effects change within the three waves of the panel survey given the different translation modes (on-the-fly vs. translated internally vs. translated externally)?

For our additional fourth hypothesis, we assume that these steps towards an increased standardisation of questionnaire translation lower the idiosyncratic part of questionnaire translation and therefore decrease interviewer effects from one wave to the next. Consequently, our fourth hypothesis is as follows:

H4. Systematic interviewer effects decrease with subsequent survey waves providing a (better) scripted questionnaire.

Methods

Measuring interviewer effects through interpenetrated survey data in Zambia

Sampling procedure. ZamSaP consists of a one-stage stratified probability sample where the respondents were chosen by simple random sampling from each of the savings groups. All 529 savings groups eligible for RUFEP were included. Four members per savings group were selected randomly using the savings group member numbers according to the savings group member registry. In some exceptions, the sample includes three or five respondents interviewed in a savings group, either due to missing interview recording files or non-selected respondents insisting on participating in the survey. This sampling procedure resulted in a dataset with more than 2000 respondents for wave 1 in 2016 and wave 2 in 2018 with minimal sample attrition due to non-contact. About 500 respondents consisting of one additional fifth member per savings group were added after a random draw in wave 3 in 2019.

Interpenetrated assignment of interviewers. Within each interviewer team consisting of five interviewers in total, four interviewers were randomly assigned to savings group members to prevent any selection bias caused by interviewers choosing the interviewee. It is important to note that this partial interpenetrated assignment differs from a pure one explained and refined by (Fellegi, 1964, 1974) in the sense that the sub-samples between the interviewer teams are not partitioned randomly. The geographic area in which each interviewer team worked was decided based on an organisational consideration for time and cost efficiency. However, the core step, being the random assignment between each interviewer in a team and randomly selected respondents, was rigorously followed with minor exceptions.

Interviewer numbers and continuity (see table 6). In the first wave, 39 interviewers were considered for the analysis while wave 2 and 3 included 25 interviewers. The data collection in wave 1 had to be split into two – the Northern Province was visited before the national elections in the summer of 2016 while the interviews for the remaining two provinces happened after. Due to the pressing election, the survey company of wave 1 hired 30 interviewers for the Northern Province. After the elections, 20 interviewers worked in the Eastern Province while 5 travelled to the Western Province for fieldwork. Nearly all but four interviewers administered the questionnaire in both the Northern and the Eastern Province in wave 1 while one interviewer worked in both the Northern and Western team. Due to the high retention from wave 1 Northern interviewers working in either the Eastern or Western Province, the total number of interviewers in wave 1 is 38 compared to 25 interviewers for wave 2 and 3. Similarly, the retention rate between wave 2 and 3 is high as the data collection firm changed from wave 1 to wave 2 and 3. It was in the discretion of the survey firm whom to select for the interviewer training even though interested interviewers from wave 1 were encouraged to apply to the second service provider.

Table 8 Interviezer teams per province

	Wave 1		Wave 2		Wave 3		
Number of interviewers	Total	Retention from wave 1 Northern	Total	Retention from wave 1	Total	Retention from wave 1	Retention from wave 2
Northern Province	30	reference	10	1 out of 10	10	1 out of 10	4 out of 10
Eastern Province	20	16 out of 20	10	2 out of 10	10	1* out of 10	6* out of 10
Western Province	5	1 out of 5	5	0	5	0	5 out of 5
Total per wave	38 interviewers		25 interviewers		25 interviewers		

** one interviewer worked in all three waves in the Eastern Province team*

Interviewer teams per province. For the subsequent wave when all three survey regions were fielded at the same time, two interviewer teams worked in the Northern and Eastern provinces, respectively. The sample for the Western Province was half the size of the other provinces so only one team of interviewers was assigned to that region. In two further located districts (one in the Northern Province and one in the Eastern Province), only one interviewer team administered the interviews to save transportation costs (as referred to as district X in the explanatory diagram in Figure 6).

Effectuated interviews. As the RUFEP partnering NGOs were introducing the ZamSaP survey and interviewers to the savings group and motivating their partners to participate, unit item-nonresponse was rare as the survey participation was ensured with the savings group leaders through the help of the trusted NGOs[6]. Apart from natural attrition due to the passing away of a panel participant or non-contact due to migration, sample attrition was negligible. The interviewer workloads ranged from 25 to 81 interviews. In the Northern Province, the average number of interviews per interviewer was 28.3, while interviewers in Eastern and Western Province surveyed 41.1 and 76.2 respondents respectively. In this three-wave panel, the quasi-interpenetrated design continued also for wave 2 and 3 without striving for interviewer continuity with panel members, meaning that interviewers were randomly assigned to respondents even in the subsequent waves.

Table 9 Interviewer and respondent characteristics

	Wave 1		**Wave 2**		**Wave 3**	
Female proportion	Int.	Res.	Int.	Res.	Int.	Res.
Eastern Province	40%	81.3%	50%	83.1%	40%	83.5%
Northern Province	47%	72.4%	40%	74.3%	50%	73.2%
Western Province	60%	85.8%	60%	86.9%	60%	87.6%

The results of Chapter One highlight the importance of the gender of both the interviewer and the respondent for an examination on data quality. The above table presents the female proportion among both respondents and interviewers. Even after replacing non-traceable panel members with newly drawn savings group members in the subsequent waves and adding the booster sample consisting in a random selection of a 5th respondents per savings group,

[6] Refusals on the savings group level happened in one case, for example, as one savings group consisted of Jehovah's Witnesses who connected participating in a survey as going against their religious beliefs.

proportion of the female respondents remains stable. Regarding the interviewer team compositions, the ratio between female and male interviewers are rather balanced.

Dependent variables. To investigate whether interviewers introduce a systematic bias to survey response according to the four hypotheses of this paper, a set of attitudinal questions as dependent variables are selected for analysis. The exact wording of the question stem was "In what degree do you trust...?" (1.) government banks, (2.) private banks, (3.) microfinance institutions, (4.) non-governmental organisations (NGOs) (in general), and (5.) the respondent's neighbours. Measured on a four-point scale ("completely", "much", "a bit", "not at all"), those trust questions ask about how much trust the respondent places in certain (financial) institutions and about the level of trust in their neighbours.

Modelling interviewer effects on subjective questions

To investigate the relationship between response and interviewer and respondent characteristics, the analysis methods first need to account for the clustering of respondents (lower-level units) by interviewer (higher-level units) using multilevel models and separate them from area-effects (C. O'Muircheartaigh and Campanelli, 1999; Schnell and Kreuter, 2005). Those models usually treat the individual interviewers as random effects (Dijkstra, 1983).

To be more precise, multilevel modelling is used to examine the sources of variance across the ZamSaP interviewers regarding their collected answers on attitudinal questions. This approach allows researchers to estimate between-interviewer variance and to investigate which variables at the respondent or interviewer level explain differences in variance. In addition, responses collected by one interviewer are assumed likely to be correlated with each other. This violates the independence assumption of linear models, particularly that of ordinary least squares regression. Thus, it is necessary to use crossed random multilevel effects models which partition variance to reflect that respondent are hierarchically nested within savings groups and interviewers. However, interviewers are crossed with savings groups (see Figure 5 and Figure 6).

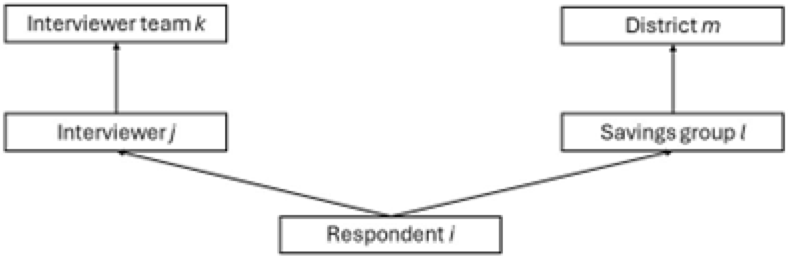

Figure 5. Nested structure to be captured by the multilevel model

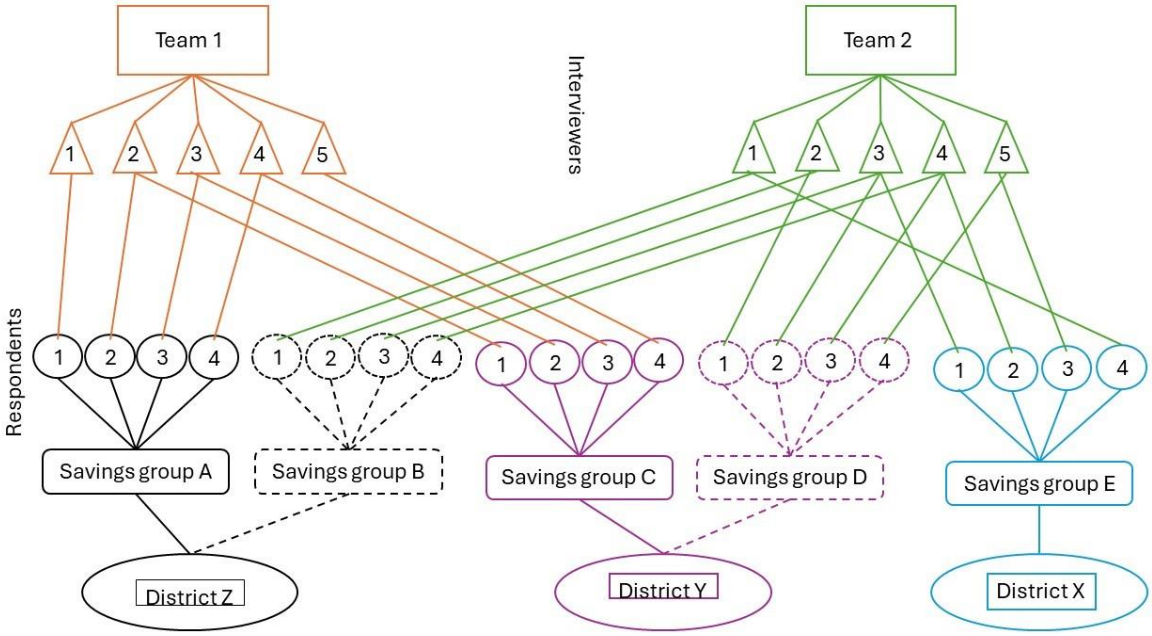

Figure 6. Example of fieldwork for ZamSaP wave 2 with more than one team per province

Following a similar design proposed by Ong et al. (2018), the step-up strategy as model building approach was employed. The "empty model" (Model 1) contains only random intercepts for different interviewers, a fixed effect for districts and a random error. The value of the dependent variable TRUST_{ij} is a response on one of the four question items for trust in different institutions given by the respondent i, interviewed by a certain interviewer *j.* This respondent i also belongs to a certain savings group *l,* located in a specific district m. As interviewers are nested within interviewer teams, the statistical model should provide clustered random errors on the team-level. As savings groups are nested within districts, respondents within districts might give more similar answers than respondents in other districts, district needs to be included as fixed effect to properly disentangle a potential area effect from the interviewer effect.

Consequently, the general empty model without the district as fixed effects has the form as follows:

$$\text{TRUST}_{ij} = \beta_0 + \beta_1 * \text{dist}_{ij}$$

$$+ u_j + \varepsilon_{ij}$$

where β_0 is the overall mean for the level of trust towards a certain institution, dist_{ij} represents the fixed effects for the 8 districts, u_j denotes the random intercept across interviewers, and ε_{ij} represents the residual error at the respondent level. All random effects are assumed to follow a normal distribution with zero means. In addition, the residuals associated with the level of trust observations are assumed to be independent of the random effect and arise from the normal distribution.

For each model fitted for the five selected attitudinal items, the intra-class correlations (ICCs) or as defined more specific intra-interviewer correlations (IICs) (West et al., 2018) are calculated as follows:

$$IIC = \frac{\sigma_j^2}{\sigma_j^2 + \sigma^2}$$

, where σ_j^2 denotes the variance of the random interviewer effects.

Results

The IICs of model 1 without any covariates are also referred to as "estimate[s] of the raw between-interviewer variance in the random interviewer effects" (Ong et al., 2018: 1786). Through the step-up model building approach from model 1 to model 3 the IICs show how the variance attributable to interviewers changes when accounting for respondent-level covariates including sex and age (model 2) and both respondent-level and interviewer-level covariates including sex and age (model 3).

Table 10. Variance of random interviewer effects, standard errors (SEs) and IICs for wave 1

	Model 1 – empty model		Model 2 – with respondent charact.		Model 3 – with interv. & respond. charact.	
	Variance (SE)	IICs	Variance (SE)	IICs	Variance (SE)	IICs
Government banks	0.260 (0.065)	22.5%	0.264 (0.066)	23.0%	0.223 (0.062)	19.6%
Private banks	0.370 (0.089)	30.2%	0.371 (0.090)	30.2%	0.375 (0.097)	29.6%
MFIs	0.382 (0.092)	33.7%	0.384 (0.092)	33.9%	0.388 (0.099)	33.1%
NGOs	0.189 (0.047)	22.5%	0.190 (0.064)	22.7%	0.181 (0.049)	22.5%
Neighbours	0.312 (0.076)	30.6%	0.314 (0.077)	30.8%	0.275 (0.073)	30.3%

The raw IICs of model 1 are large across the attitudinal items suggesting that 22.5% to 33.7% of the variance in the level of trust can be attributed to between-interviewer differences. These results with IICs of around 30% are on pair with recently published ones for items measuring abortion occurrences in Mali and Malawi (Leone et al., 2021), regulating own's menstruation in Côte d'Ivoire and Nigeria (Footman, 2021), intimate partner violence, sexual violence, the refusal for conjugal sex (for India: Singh et al., 2022; Singh et al., 2024). Accounting for respondent-level fixed effects in model 2 does not lead to an evident drop in IICs. This absence of change confirms that the interpenetrated design with random assignment of the respondent to the interviewer worked correctly. However, when adding fixed effects of both respondent-level and interviewer level covariates into model 3, the IICs decreased only in the trust questions on government and private banks.

Nevertheless, the step-up model as a model building approach using crossed random effects on interviewer-level and savings group-level provide evidence for large, non-negligible interviewer effects in all trust items that persist even after accounting for respondent-level and interviewer-level covariates. Due to the quasi-interpenetrated design of this study and to the fact that interviewers did not recruit the respondents, the IICs calculated here reliably reflect interviewer effects on measurement.

To conclude, the analyses confirm the first and second hypotheses on non-negligible interviewer effects that are not explained with respondent-level characteristics in a quasi-interpenetrated design. Nevertheless, including interviewer-related characteristics into the multilevel models does not decrease the extent of variance attributable solely to between-interviewer variance.

Table 11 Interviewer effects in comparison from wave 1 to wave 3

	Empty model wave 1		**Empty model wave 2**		**Empty model wave 3**	
Trust in…	Variance (SE)	IICs	Variance (SE)	IICs	Variance (SE)	IICs
Government banks	0.260 (0.065)	22.5%	0.162 (0.065)	12.0%	0.055 (0.020)	5.4%
Private banks	0.370 (0.089)	30.2%	0.145 (0.067)	13.2%	0.081 (0.028)	7.5%
MFIs	0.382 (0.092)	33.7%	0.095 (0.036)	8.1%	0.102 (0.039)	8.7%
NGOs	0.189 (0.047)	22.5%	0.203 (0.064)	18.0%	0.069 (0.024)	6.2%
Neighbours	0.312 (0.076)	30.6%	0.060 (0.020)	9.3%	0.057 (0.019)	9.5%

Table 3 presents the visibly decreasing interviewer effects of the empty model from wave 1 to wave 2 after the introduction of a scripted, translated questionnaire. For three out of five items, the IICs also drop drastically between wave 2 (internal translation) and wave 3 (pre-tested external translation). The decrease clearly reflects the efforts put into the improvement of the questionnaire translation and confirms, thus, the last hypothesis on interviewer performance being enhanced with improved questionnaire translation. Figure 7 compares visually the decrease in IICs in wave 1 using the above-mentioned step-up modelling approach with the changes in IICs due to the different modes of translation (on-the-fly translation vs. scripted translated questionnaire with internal vs. with external translation). The results show that translation quality plays a key role in the extent of interviewer effects, which tends to be overlooked by both survey administrators and researchers alike.

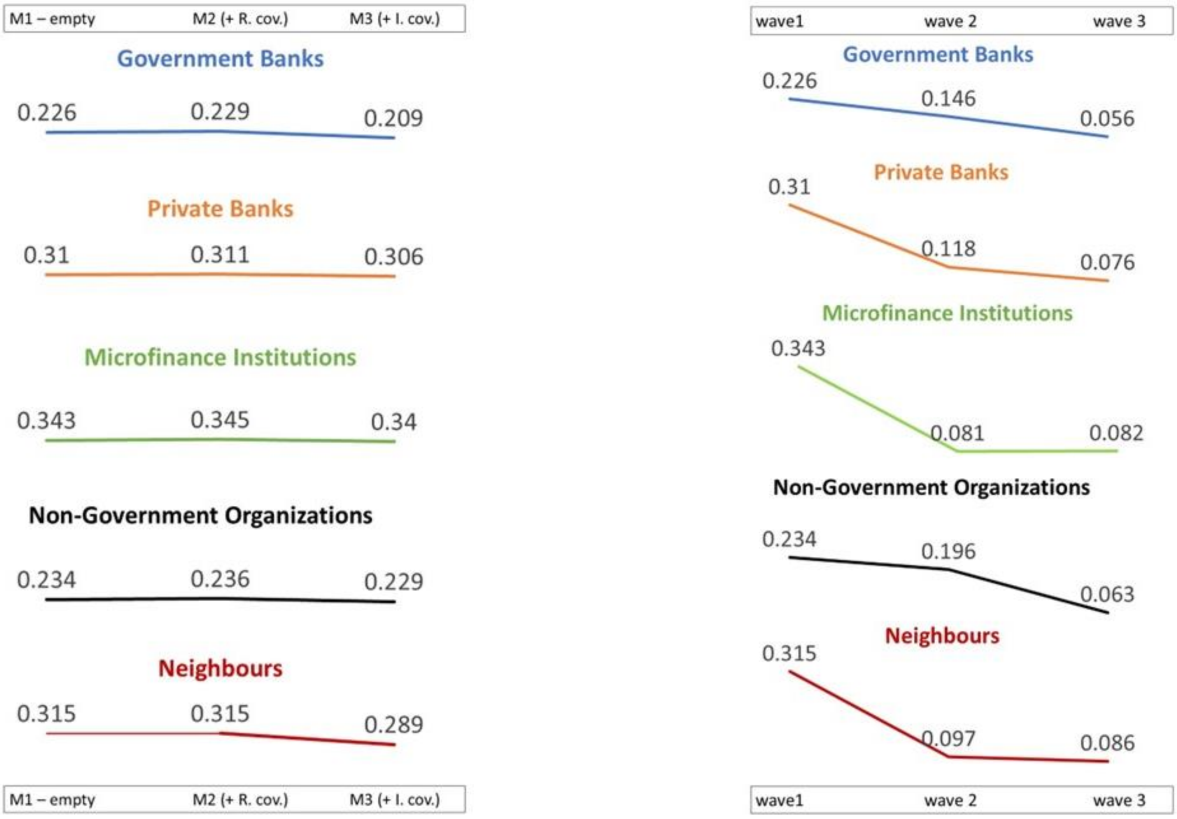

Figure 7. Comparison of reduction of IICs in wave 1 with on-the-fly translation only (left panel) to IIC reductions across three waves (right panel)

Discussion

This study first investigates whether observable interviewer characteristics evidence by previous studies to influence responses (see e.g., West and Blom, 2016) can explain large interviewer variances found in attitudinal questions on trust in different institutions in this Zambian study – ZamSaP. Surprisingly, one sees little effect in adding explanatory variables on the respondent and interviewer level – the IICs remain consistently high. Future studies aiming for a more thorough explanation of large interviewer effects should measure and include also unobservable interviewer characteristics as explanatory variables, such as interviewer motivation or prior survey experience.

Furthermore, using a panel survey allowed for linking the interviewer effects to impact of translation quality. However, the subsequent decrease in interviewer variance with each wave could be potentially confounded with either interviewers or respondents becoming more experienced, or even with a temporal change in attitudes. Effects of interviewer learning, and experience should have been more pronounced between the two later waves as most interviewers remained for both waves. Nevertheless, these reductions in interviewer variance are not as large as the ones between the first two waves.

Despite the shortcomings of this study design regarding the separation of the translation quality effect from any panel conditioning effects, the mere magnitude in interviewer variance reduction from wave 1 to 2 coupled with the drastic change from unscripted to scripted mode of translation suggests a non-negligible role of translation quality. Future studies focusing on distilling panel conditioning and interviewer learning effect from the pure translation mode effect need to consider experimental designs comparing interviewer variances of two experimental interviewer groups using either on-the-fly or scripted translation.

Although we made no explicit expectation regarding the interviewer effect reduction between the latter two waves, the implicit expectation would have been that the enhancements resulting from a pre-tested, external translation might have avoided situations in which interviewers felt the need to change the scripted questionnaire translation in the field. However, the IICs of two out of five trust items did not drop (trust in micro-finance institution and one's neighbours). One explanation might be that the translation for those two questions was already of "high" quality in wave 2 and did not call for a change for wave 3. The question on trust in micro-finance institutions has contained a definition in the original source questionnaire since the version used in wave 1. This predefined definition which was then included in the scripted translations might have helped the interviewer to administer the question in a standardised manner in wave 2 and wave 3.

Also, the trust question regarding neighbours touches upon on the familiar concept of people living next door whose translation has little leeway to be improved from the internally provided translation in wave 2 to the external, team-based translation in wave 3. Nevertheless, the actual amount of on-the-fly translation in wave 2 and wave 3 needs to be measured to investigate the relationship between translation quality and interviewer performance which exceeded the possibility of this study design.

Despite shared similarities with other African countries regarding the multilingual and multi-ethnic setting, generalisation of findings from this Zambian survey to a broader scope of sub-Saharan Africa or other LMICs might be debatable. At the expense of generalisation, most surveys implemented in a development context focus on only a part of the population (usually, the targeted beneficiaries of the development project). Likewise, the ZamSaP survey is part of an impact evaluation of a development program. Consequently, neither were interviewers working in the data collection randomly selected from a larger population of Zambian interviewers nor were the study savings groups a random selection of all savings group in that geographic province of Zambia. Generalisation to all interviewers might, thus, be limited. Furthermore, the absence of sample weights to account for the complex sampling design further hinders generalisation. Nevertheless, this study is the first of its kind to connect interviewer performance and effects to translation quality.

In addition, the survey context is special as respondents might connect their participation to future service delivery by donors or non-government organisations – a common pitfall of surveys in sub-Saharan Africa (Randall et al., 2013). While respondents and their savings groups might expect potential service delivery, the interviewers were specifically trained not to mention any future provision of services during the survey introduction. Rather, they were instructed to explain that the survey objective is to understand the economic situation and well-being of savings group members.

This study fits well into the recently flourishing body of interviewer effect from LMICs cited in this paper. Researchers have started to investigate other forms of interviewer effects beyond response variance for specific items. For example, Sarac and Koc (2021) demonstrated that on-the-fly translation can also increase item nonresponse in the case of missing answers to birth data questions when Turkish interviewers translate into Kurdish and Arabic on-the-fly.

Nevertheless, this study marks the beginning of research accounting for complex multilingual contexts as a survey error by demonstrating a link between interviewer effects and different modes of translation and translation quality. Future research needs to investigate why, and under which circumstances interviewers deviate from standardisation and thus, might influence interviewer effects and data quality. Also, more studies in measuring translation quality are needed as "many of the concepts and definitions used in data collection are difficult to translate into local languages: some have close parallels, whereas, in other languages, detailed explanations need to be given" (Randall et al., 2013: 764). Various aspects of questionnaire translation and their uptake by interviewers in delivering the correct meaning of the questions in any language are key to the respondents’ comprehension and survey data quality.

Chapter Three: Studying language switching in multilingual survey interviews

Abstract

Survey researchers and practitioners usually assume that an interview is administered continuously in one language. However, in multilingual settings found in multi-tribal, multi-ethnic countries, this assumption does not always hold, as both the interviewer and the respondent have several languages at their disposal to communicate with, in case they encounter conversational difficulties in the designated survey language. Most low- and middle-income countries (LMICs) are multi-ethnic and multilingual, leading to the situation that both interviewer and respondent may even converse with each other in combinations multiple local languages switching from one to another language.

I analyse the relationship between language switching defined as a deviation from the survey language and several indicators of problematic interactional behaviours using the interactional analysis of over eight hundred recordings in two local languages (Bemba and Chewa) from a Zambian survey. In a second analysis step, I link language switching to interviewer effects, which is the proportion of response variance attributed to the interviewers.

This study's objective is to document the process of switching language and its implications to survey data quality. The results show that the frequencies of language switching differ among study regions, as well as over different questions, and that language switching does not occur as a single phenomenon but rather co-occurs with other problematic interactional behaviours. Thus, language switching can be categorised as another problematic behaviour indicating the breakdown of the cognitive answer process. In this sense, language switching can also be considered as a disruption to the ideal question-answer sequence.

Introduction

Multilingual contexts as potential error source in survey interviews

Standardised interviewing remains widely used in all parts of the world and for all major, large-scale population surveys due to its timeliness, precision of measurement, and control over certain survey errors, namely interviewer error (Beatty, 1995). Politicians, practitioners, and scholars alike are dependent on high-quality survey data for decision-making, policymaking, and research. To avoid invalid findings from survey data, a paramount objective is minimizing survey error. In survey research, survey errors have been identified from various sources (such as the respondent or the interviewer) and may arise at various stages of the data collection, from designing a questionnaire to recording the answer (Groves et al., 2009).

In low- and middle-income countries (LMICs), interviewer-administered surveys remain the principal data collection tool, whether mainly in the form of face-to-face surveys as before the outbreak of the COVID global health pandemic (see the study of Lupu and Michelitch, 2018 for evidence in political science) or mainly in the form of telephone interviews as in the post-COVID period (see Frankovic et al., 2023 for data collection of opinion polls around the world). Employing interviewers allows survey designers to account for varying levels of literacy due to the lack of universal education, particularly among rural populations.

As most LMICs are multi-ethnic and multilingual, the need for standardisation becomes even more paramount. However, survey managers are often restricted by cost constraints so that the questionnaire is commonly translated in a few local languages, focusing on one official language and/or one *lingua franca* per study region. Just defining a fixed number of languages to be scripted among multiple languages might be insufficient when encountering respondents who are less proficient in that specific survey language. Except for when re-interviewing the same respondent within a panel study, neither the survey designer nor the interviewer knows in advance the language proficiency and the language preference of a respondent. However, this unknown factor puts forward the concern of added sources of survey errors within a multilingual context. For example, Wenz et al. (2021) showed that item nonresponse is higher when the respondent is not interviewed in their mother tongue in a general UK population survey. Jacobsen (2022) found similar results for a German refugee survey for which the questionnaire was not translated for all native languages spoken by the targeted respondents.

Notably, the multilingual context of most LMICs renders the interviewing process extraordinarily complex, as both respondent and interviewer can communicate in multiple languages – each of them with potentially distinct levels of proficiency. Consequently, I expect that some multilingual respondents will exhibit more cognitive processing problems while answering questions in a certain pre-defined survey language, as they may not be sufficiently proficient in that language. This might be especially pronounced for those with lower education.

This paper's objective is to document this process of switching language and its implication to survey data quality. Based on the recording of ten questions from more than eight hundred respondents from a probabilistic household survey in Zambia, the frequency of language switching is estimated between 2 and 7 percent depending on the question. Here, a language switch is defined as using one sentence or more than three words in a different language than the one the respondent or interviewer started with.

Using behavioural coding to understand language switching

As survey data quality is one of the key interests for data producers and users, some researchers advocate to understand the response process through interviewers' and respondents' behaviours and their interaction (see Ongena and Dijkstra, 2007; Schaeffer et al., 2016). Dykema et al.

(2016) and Olson et al. (2018) make a particularly compelling case about how certain respondent behaviours (e.g., expressing uncertainty in words or through pauses, request for clarification, providing non-substantive or problematic substantive answers) indicate a problem in understanding the question or providing a valid answer. In addition, interviewers have a direct influence on how the respondent performs their task in answering through certain behaviours, such as whether they read the questions verbatim or whether they are probing after receiving an uncodable answer as shown by Dykema et al. (1997) – one of the earliest works on the topic.

Previous research usually links question characteristics to respondent-interviewer-interaction, as all three factors play a key role in the question-answer-process (Dykema et al., 2016, 2019; Holbrook et al., 2006; Olson et al., 2018). Essential question characteristics to consider, such as question type or response format, induce comprehension troubles for the respondent and interviewer which then become apparent by problematic interactions (for a taxonomy on question characteristics see Dykema et al., 2019). Using interaction coding data, Holbrook and colleagues (2006) have established connections between respondent comprehension and the type of question. This includes the response format (agree/disagree scales, open-ended numeric, yes/no categorical, etc.) and type of judgement (subjective, behavioural, or knowledge), as well as the question's sensitivity and level of difficulty, often operationalised as reading difficulty. They found that interviewers are more likely to have problems reading longer and more difficult questions. In contrast, respondents are more likely to show comprehension problems for difficult questions and specific response formats. However, the authors do not find evidence that inaccurate interviewer readings are related to respondents' comprehension problems or to respondents giving nonvalid answers.

Language switching as a problematic interactional behaviour

Researchers and practitioners usually assume that an interview is administered in one language, i.e., the interviewer asks questions, and the respondent answers them in the same language. However, in multilingual settings found in multi-tribal, multi-ethnic countries, this assumption does not always hold, as both the interviewer and the respondent speak more than one local language and may vary in their degrees of fluency. Thus, they can choose among several languages or even different language combinations to communicate with each other.

Administering interviews and individual questions in multiple languages and thus deviating from the question wording scripted in the main regional language might introduce measurement error. This might particularly be the case for multilingual interviews and cases when an interviewer relies on their personal on-the-fly translation when using a local language which is not the *lingua franca* of that province.

As multilingual respondents differ in their proficiency in designated survey languages, some will exhibit more cognitive processing problems evidenced by audible manifestations of problematic interactional behaviours during the interview (i.e., through seeking clarification or repeating the question). Such difficulties will be especially pronounced among respondents with lower education. Faced by comprehension challenges, interviewers might adapt their behaviour to ensure that respondents understand the question as it was intended, thus deviating from standardised interviewing in the predetermined survey languages. Such problematic behaviours by interviewers include questioning, giving clarification and/or probing in a non-scripted language, and/or translating non-conform answers into the survey language (see Table A1 for some examples from translated transcriptions of survey interviews in Zambia).

This study investigates little-understood sources of problematic interactional behaviours between interviewers and respondents in the form of language switches. Language switches are defined as interviewers starting in one language, typically in one of the designated survey languages, and the respondent or interviewer changing to another language, usually not among the designated survey languages.

Respondents might switch languages to better communicate their survey answer or out of habit or convenience, as they are more fluent in a non-survey language. In contrast, interviewers are usually trained to refrain from deviating from the survey language against their own habits and convenience. An interviewer may ignore the instructions (commonly given as part of the training for standardised surveys) only if they perceive that the respondent might have comprehension difficulties which might obstruct their task to retrieve an answer. For example, when the interviewer is faced with a respondent who is not adequately proficient in the designated survey language and shows signs of comprehension problems, they might decide to deviate from standardised interviewing techniques and translate survey questions without a script. This deviation from standardised interviewing might lead to interviewer error.

From a theoretical perspective, language switching initiated by the respondent can be understood as a signal for a problematic interaction, while a language switch by the interviewer as a reaction to a respondent's behaviour can be regarded as a symptom of a problematic interaction (see Conrad and Schober, 2021, citing Clark 1991). Therefore, this paper aims to investigate the role of language switching in problematic survey interactions and sets out to answer two research questions (RQs) on whether language switching is co-occurring with other, established interactional problems and on what question, respondent, and interviewer characteristics are associated with language switching:

RQ1: Regardless of who initiated it, is language switching associated with another problematic interactional behaviour?

RQ2: How is language switching associated with question, respondent, and interviewer characteristics?

Reflected in the first research question is our interest in establishing whether there is a relationship between language switching and other problematic interactions. In other words, does language switching occur concurrently with other behavioural signals for uncertainty or misunderstanding or does it happen inherently due to other reasons. As language switching can be regarded as a cue for interactional issues between respondent and interviewer, it can be expected that it should appear more commonly together with other problematic interactional behaviours. Table A1 in the appendix provides two excerpts of interview transcripts the Zambian survey used for analysis. It provides examples for instances when either the interviewer or the respondent switches languages during the survey interview.

While language switching is the dependent variable, the independent variables listed in Table 1 were chosen to build upon the comprehensive work of Olson et al. (2019), who link various question and respondent characteristics (controlling for a few interviewer characteristics) to six respondent outcome behaviour in their first turn (coded as adequate answer, qualified answer, uncodable answer, don't know, refusals, and clarification requests). The authors find that problematic respondent outcome behaviours are foremost linked to question characteristics rather than interviewer or respondent traits (Olson et al., 2019). Using their finding as the starting point and the answer to the first research question which investigates whether language switching co-occurs with other problematic interactional behaviours, I then model the association of language switching to "static", pre-defined characteristics regarding the question, the respondent, and the interviewer (see Table 1).

Table 12. List of question, respondent, and interviewer characteristics

Level	Characteristic	Operationalisation
Question	Question length	Word count
	Question type	Attitudinal vs. behavioural vs. knowledge
	Response option format	Open-end numeric (date) Closed-ordinal Yes/no
	Number of response options	Count of response options
	Fatigue	Question sequence
Respondent	Age	Reported age as continuous variable
	Proficiency in survey language	Binary coding of speaking survey language at home
	Female	Binary coding of being female
Interviewer	Female	Binary coding of being female
	Age	Continuous age
	Proficiency in survey language	Binary coding of speaking survey language at home

Problematic interactional behaviours of interest

Previous research has provided evidence for the association of certain interactional expressions with cognitive processing and measurement error. To answer the first research question on investigating the co-occurrence between language switches and other problematic behaviours of interest, the following problematic behaviours have been selected for analysis:

1. Language switches as deviations from the designated interview language by either the interviewer or the respondent. Per our definition, a language switch occurs if one full sentence or at least three words are spoken in a language different from the interview

language. This definition aims to exclude colloquial language borrowing of a few words in a different language due to habits of everyday life.

2. Non-verbatim question reading of the scripted survey language as a variation in how the question is delivered can introduce variation in the question stimulus, potentially increasing interviewer variance.
3. Any pre-emptive and follow-up behaviours by interviewers to obtain a codable answer (such as providing explanations without being asked or feedback or probing). Notably, Mangione et al. (1992) found that probing is associated with higher interviewer variances. The developed coding script for this study specifically separates language switches from non-verbatim question reading in the survey language.
4. Seeking clarification is also considered as an interactional indicator for processing problems (Holbrook et al., 2006; Schaeffer and Dykema, 2011).
5. Apart from asking directly for clarification, respondent uncertainty might give cues for interviewers to change their behaviour (see e.g., Schober and Bloom, 2004).

Following the approach of previous research (e.g., Dykema et al., 2016, 2019; Garbarski et al., 2011), I constructed an index out of the respondents' behaviours to capture potential problems when answering the question, including providing reports, considerations, expressions of uncertainty, and other uncodable answers. More concretely, respondents' uncertainty is captured as a dichotomous indicator if the respondent expressed at least one of the following behaviours: mitigating phrases that reduce the exactness, precision, or certainty of an answer and are offered as answers or parts of answers (e.g., ‘‘I guess" or ‘‘just,'' ‘‘maybe,'' ‘‘about,'' ‘‘put,'' or ‘‘I'd say''); a hypothetical response option; repairs; paralinguistic disfluencies (‘ums", "uh"); and a report or consideration, that is either stated as an answer or offered as an explanation for an answer. These interactional behaviours were selected as they may indicate difficulties in the response process or so-called “breakdowns in the cognitive response process” (Olson et al., 2019: 294). In a multilingual context, those breakdowns may not only be connected to restricted cognitive ability but also due to limited language proficiency.

Methods

Study background and questionnaire translation

The survey data for analysis is from the Zambian Savings group Panel survey (ZamSaP), a three-wave panel on savings group members living in rural and semi-urban parts of three Zambian provinces (see Figure 8). ZamSaP consists of a one-stage stratified probability sample where the respondents were chosen in wave 1 in 2016 by simple random sampling from each of the targeted 529 savings groups. For the first wave, four members per savings group were selected randomly from their savings group member registry. In the final survey wave, the one for analysis, a fifth respondent in each savings group was randomly selected to join the previous four panel members.

For the last survey wave in 2019, the survey organisers put a considerable budget to improve questionnaire translation into the *lingua franca* of the three study provinces. For the first time, the translation was not done either on-the-fly by the interviewers (wave 1) or in-house by the survey company (wave 2) but by external translators. The English source questionnaire was translated independently by two translators and their divergences were discussed and resolved before testing the questionnaire in the three survey provinces. The translation was then improved and corrected during one day of the week-long training using a team-based approach between the main translators and the interviewers.

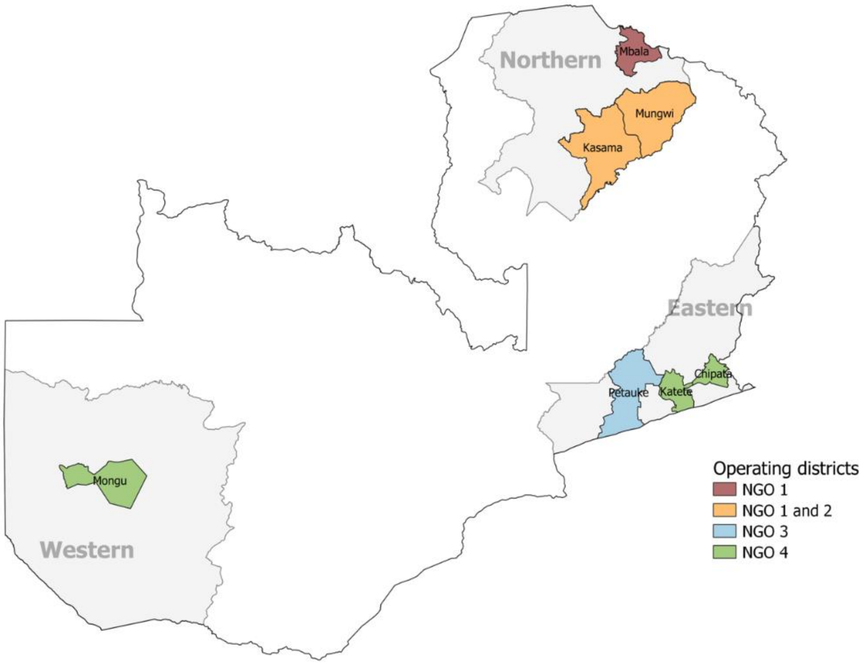

Figure 8. Operating districts of the implementing partners in Zambia[7]

Multilingual interviewing (MI) in a multilingual setting: The case of Zambia

Like most African countries and some other low- and middle-income countries (LMICs), Zambia is a country where multiple ethnicities speak different languages. Marten and Kula classify the numerous local dialects into "into twenty-six dialect clusters or 'languages', which in turn are grouped into sixteen [language] groups" (2008: 294). While English is the official language and seven local languages are recognised as national languages, language use of the predominant and the second language varies considerably around languages and regions (Marten and Kula, 2008). When deciding on a local survey language under restricted budget constraints, the regional *lingua franca*, the language spoken by the majority, is typically chosen.

[7] Figure taken from the following report: Frölich, M and Nguyen, PL, 2020. Impacts of linking savings group to formal financial service providers and strengthening their internal group insurance mechanism in Zambia, 3ie Impact Evaluation Report 121. New Delhi: International Initiative for Impact Evaluation (3ie). Available at: https://doi.org/10.23846/TW4IE121

Strikingly, during an interview, a language switch can happen within the same question-answer-sequence or even the same sentence. In fact, Boztepe (2003) differentiates between language switching (alternative use of two or more languages in one conversation) and language mixing (language switch within a sentence). Both language switching and language mixing are common in African languages and happen between local languages and the official language – the prior colonial language (Bylund and Athanasopoulos, 2014).

In this complex, multilingual setting, no published literature could be found which sheds light on what factors drive interviewers and respondents to speak in multiple languages during the survey interview and whether a new pattern in question-answer-sequences emerges. Most literature focuses on a very specific context: the bilingual interviews or surveys with bilingual interviewers and respondents in English and Spanish or Korean and English in the United States (see several relevant chapters in the book "*The essential role of language in survey research"* edited by (Sha and Gabel, 2020); or for a first published study on language switching in the Latino National Survey 2006 see Saavedra Cisneros, 2025).

For the multilingual Afrobarometer Round 6 with 36 African countries administered in 2014-2015, Lau et al. (2020) draw a comprehensive picture on how respondents and interviewers differ in their first languages and how they chose the survey language, as captured by the data collection software. Their data for Zambia show that in about one third of the interviews, the interviewer and interviewee share the same first language or home language. For the remaining interviews with no shared home language, 60% of all interviews relied on a so-called bridge which is neither the respondent's nor the interviewer's first language.

The original survey data set used for this study, ZamSaP wave 3 collected in 2019, contains 2,604 interviews. The interviews administered in the Western Province were excluded here due to the lowest prevalence of multilingual interviews, both in terms of frequencies (a total of 14 interviews) and proportion (ca. 3%). Therefore, the study includes the Northern and Eastern Province with the respective scripted survey languages, Bemba and Chewa. According to the interviewer observations, interviewers and/or respondents in ca. 33% (or 329 interviews) of all Chewa interviews and ca. 4% (or 44 interviews) of all Bemba interview used multiple languages during the survey interview.

In the ZamSaP wave 3, language proficiency was measured as a binary variable capturing whether the respondent or the interviewer spoke the survey language at home. Nearly all respondents of the Northern Province (99.4%) and all interviewers[8] spoke the survey language, Bemba, at home. In the Eastern Province, however, half of the respondents spoke the survey language, Chewa, at home while the other half did not. Similarly, only half of the interviewers working in the Eastern Province reported speaking Chewa at home. Given this more diverse language proficiency among respondents and interviewers of the survey language Chewa, it is not surprising that most language switches are found in the Eastern Province sample only.

Table 11 shows the language use in the interviews conducted in the Eastern Province. According to the interviewer observations captured at the end of each interview, 83.5% of the monolingual interviews in the survey language, Chewa. In reportedly 13.5% of the cases, the interviewer conducted the interviewer in language without a scripted questionnaire translation, Nsenga, while the respondent answered in the same non-survey language. Multilingual interviews involving Chewa happened mostly as a combination of Chewa and Nsenga (89.7%) but also as a combination of Chewa and English (in 17.4% of the cases). In 2.6% of the interviews in the Eastern Province, Bemba was used in combination with Chewa. Both Bemba and English questionnaires were scripted in the data collection software so they were in theory

[8] With missing information on one Bemba interviewer.

available for the interviewers to use in case of need. As a consequence, only in case interviewer reported posed questions in Nsenga, on-the-fly translation could be a concern for data quality.

Table 13 Language use in the Eastern Province

Languages used by respondent	**Monolingual interviews**	**Multilingual interviews involving**	**Language combination with multilingual Chewa interviews**
Chewa (= scripted *lingua franca*)	83.5% (555)	45.2% (310)	Reference: 310 interviews
Nsenga = no scripted translation	13.5% (90)	43.3% (297)	89.7% (278)
English	2.9% (19)	9.9% (68)	17.4% (54)
Bemba	0%	1.3% (9)	2.6% (8)
Other	0.2% (1)	0.29% (2)	0.7% (2)

Questions for analysis

Our primary data consists of behavioural codes from audio recordings which the data collection software automatically recorded for each interview. As part of the quality assurance protocol of the data collection firm, the data collection software recorded ten selected question snippets from each interview after July 16th, 2019. Out of the ten recorded question snippets, only those eight questions that have been asked unconditionally to all respondents are used for analysis (listed in Table A2 in the appendix).

Question K27 is one item out of an item battery of ten items measuring the external and internal locus of control. Question D05 is an essential question to this Zambian survey on financial behaviour and financial inclusion as it captures the individual amount of savings. Question D09B records the time the last saving cycle of the savings group ended and when the individual savings plus interest, the so-called share-out, was paid out; the question is conditional as only savings group members having been part of the group for more than one cycle were asked this question. Question L05 sets forth a hypothetical scenario on the functioning of the social fund, the group-inherent social insurance component, without naming it. If the respondent understands the insurance concept, that knowledge will enhance their ability to comprehend and answer this question. However, the hypothetical scenario presented is complex and thus, demands high cognitive skills from the respondents.

The last four questions pertain to a multi-trait, multi-methods (MTMM) experiment on answer scales (which tests four-, five-, six- to eleven-point scales against each other). Its objective was to study how respondents react and respond to different response scales by varying the number of the response options, with or without a middle category and reading out the option of "don't know", as well as offering either agree/disagree versus item-specific scales. Therefore, the respondents were randomly asked either X02 and X23 or X06 and X26 with the respective answer options and scales. Questions X02 and question X06 measure the degree of trust in neighbours while question X23 and question X26 are the repeated items of question K27 with different answer scales.

Recruitment and training of coders

To select the final six coders from the interviewer pool of the partnering data collection firm and to ensure their task performance, the potential candidates participated in up to three in-person training sessions with a remote trainer online (December 2021, April 2022, and July 2022). To make the training more efficient, the recruited coders were trained separately in the third training course according to the local languages (Chewa vs. Bemba). Instead of the one-directional classroom approach from the first two trainings, the coders had to collectively and individually code mock interviews which were either in English or in Bemba with translated transcripts available in English to discuss their results within the group and guided by the trainer. Several feedback sessions were scheduled to discuss their mock coding with the trainer to ensure the quality performance of the coders. The best eight performing coders were then selected for production coding forming two teams of four, one for each language. As Nsenga was the language spoken at home by most of the respondents, the Chewa coders was selected also upon their proficiency in Nsenga.

Case selection strategy for interviews to code

Due to deficient performance, one Bemba interviewer had fewer interview recordings and, thus, their interviews were omitted for analysis. In the end, the interviews of 9 Bemba interviewers and 10 Chewa interviewers were considered for analysis. The interviewer observations from the second wave in 2018 show that ca. 15% of the administered interviews exhibit multiple language combinations between interviewer and respondent. To identify the maximum number of interviews with potential language switches for coding, the selection strategy was guided by the information available in all three waves on whether the interviewer reported in their observation that either they or the respondent used more than one language. Only after those reportedly multilingual interviews were selected, a “top-up” random sample of the remaining assumingly non-multilingual interviews per interviewer was drawn for coding. This was done to get a more balanced number of interviews coded per interviewer. The interviews mainly administered in English were used for training and, thus, are not part of the sample for production coding and analysis.

Table 12 compares the sample selected for coding (813 interviews) with the full sample (2,075 interviews) split between the two provinces. The sample selected for coding (highlighted columns) is comparable to the full sample (non-highlighted columns) regarding the proportion of female respondents. Interviews administered in English were used for training, so the production coding included fewer interviews for which English was the selected survey language on the device. The above-described case selection strategy successfully included reportedly multilingual interviews for coding to maximise the chance of finding language switches in the chosen question snippet recordings. After separating those multilingual interviews into instances where the interviewer uses more than one language versus where the respondent does, it becomes clear that multilingual interviews are mostly respondent driven.

Table 14. Difference of coded sample with overall sample

	Eastern Province		**Northern Province**	
Sample	Coded	Full	Coded	Full
No. of interviews	520	994	293	1,081
Lingua franca	Chewa	Chewa	Bemba	Bemba
Female	85.19%	83.50%	72.70%	73.16%
Lingua franca as selected survey language	99.23%	94.97%	95.90%	94.84%
English as selected survey language	4 interviews (0.77%)	50 interviews (5.03%)	12 interviews (4.10%)	58 interviews (5.16%)
Interviewer reporting multilingual interviews	269 interviews (51.73%)	329 interviews (33.10%)	25 interviews (8.53%)	44 interviews (3.91%)
Interviewer reporting to have used more than one language	68 interviews (13.08%)	83 interviews (8.24%)	6 interviews (2.05%)	17 interviews (1.51%)
Interviewer reporting respondent to have used more one language	267 interviews (51.35%)	324 interviews (32.60%)	25 interviews (8.53%)	44 interviews (3.91%)

Evaluating the coding quality

The production coding was monitored rigorously through two steps or indicators: 1.) The coded interviews were cross-checked daily to see whether the coders coded all audio records pertaining to one interview; 2.) At three times during the production coding, the quality controller cross-checked whether all coders agreed in identifying the existence of a remaining question-answer sequence in the double-coded interviews. The first indicator of the quality assurance is of key importance to ensure that all existing audio files were used for coding the respective interview question and no audio files were skipped. The second indicator aims at the correct identification of paradigmatic question-answer sequences which consist only of two turns in total: one turn in which the interviewer asks the question and a second turn in which the respondent provides the answer.

Apart from avoiding omissions of audio files to be coded, the correct identification of non-paradigmatic question-answer sequences is pertinent to this study, as they are considered by the literature as an indication for problematic interviewer-respondent interactions. A question-answer-sequence exceeding two turns potentially shows either that the interviewer is having problems delivering the question or that the respondent has comprehension issues(Schaeffer and Maynard, 2005). To rigorously guarantee that the coders adhere to the key coding rules satisfying the pre-defined quality indicators, the coders were obliged to re-code the full interview each time they violated one of these two rules.

The strict re-coding obligation when failing the two quality assessment indicators prolonged the production coding and simultaneously impacted the work morale. Three coders quit out of their own volution early on and none of their work was used for analysis due to the lack of quality. Two Bemba coders were dismissed in mid-production coding which led to a restructuring of the two teams to ensure the coding of the larger sample of Chewa interviews.

Table 3 presents the share of coded interviews in the sample without interviews subjected to duplicate coding, as well as the prevalence of multilingual interviews identified per coder. Among the two dismissed coders, the coded interviews of one were discarded due to concerns of quality, while those of the other coder 109 were kept for analysis, as they passed the quality assessment (representing ca. 28% of the Bemba sample). Due to this restructuring, coder 111 ensured ca. 54% of the Bemba sample (coder 111). The third Bemba coder 110 was also proficient in Chewa and Nsenga so they changed teams during the production coding. Therefore, their share of coded interviews (disregarding duplicate coding) was the smallest among the coders per language (ca. 22% of the Chewa sample and ca. 18% of the Bemba sample).

Table 15 Share of coded interviews and share of multilingual interviews per coder using the coded sample without interviews subjected to duplicate coding

	Share of coded Chewa sample (%)	**Share of coded Bemba sample (%)**	**Prevalence of multilingual interviews coded in Chewa sample (%)**	**Prevalence of multilingual interviews coded in Bemba sample (%)**
Coder 101	43.18		53.59	
Coder 103	34.92		14.20	
Coder 110	21.90	18.37	7.55	9.62
Coder 111		53.71		8.86
Coder 109		27.92		7.24

An issue arose with the best performing Chewa coder 101 according to the performance checks. While they rarely failed to include all audio files per interview and they correctly identified paradigmatic question-answer-sequences in the duplicate coding, they coded half of their assigned interviews as multilingual while their two Chewa colleagues produced frequencies of ca. 14% and ca. 8%. When comparing the identification of a language switch the duplicate coding of the Chewa sample coded by all three coders, coder 101 identified language switches in 17 out of 35 interviews, and in one additional instance did not identify a language switch leading to a disagreement with their fellow coders in more than half of interviews assigned for the duplicate coding (results not shown in Table 3). Due to this stark discrepancy of coder 101, their coded interviews were disregarded for the calculation of the intercoder reliability, and the analysis sample was restricted to include only interviews which were consistent with the interviewer observation (see next section for details).

To compare the coding quality among the coders, 6 interviews (representing a share of ca 3% of the coded Bemba sample) and 35 interviews (equal to a share of ca. 7% of the coded Chewa sample) were subjected to duplicate coding by at least two interviewers[9]. In terms of interrater reliability on identifying a multilingual interview, the kappa score for the two main coders in Bemba is at 89% agreement while the one of the two Chewa coders is at 91%. For the duplicates, only the interviews coded by the best performing coders were retained for analysis.

Though the coding definition of a language switch was defined and explained to all coders during the training, discrepancies in coding language switching may arise due to the following reasons: (a) recording quality impedes some coders to detect the language switch; (b) the coders disagree on languages are phonetically similar (which is the case of Chewa and Nsenga); (c) some coders interpreted the definition differently (which may be the case for one Chewa coder).

Quality of the recordings of the restricted analysis sample

Due to the large discrepancy of identifying language switches driven by one coder, the analysis sample is reduced from 813 coded interviews to 505. In other words, the analysis only considers interviews for which both the interviewer reported the use of multiple languages during the interview and the coder coding a language switch. While the following results are based on the restricted sample, this conservative decision did not alter the main conclusion from the findings (see discussion).

[9] Originally, the duplicate coding was planned for 10% of the sample. Due to the reduction of coders explained above, the work allocation for duplicate coding had to be transferred to re-code the interviews previously coded by coders dismissed due to unsatisfactory coding quality.

When evaluating the quality of the recordings, the coder indicated before the actual coding whether the audio file(s) for a certain question exhibited a full question-answer sequence (QAS), only part of a QAS, or was not codable. An audio file is not codable when, for example, the recorded exchange between the interviewer and the respondent was not audible due to too many loud background noises.

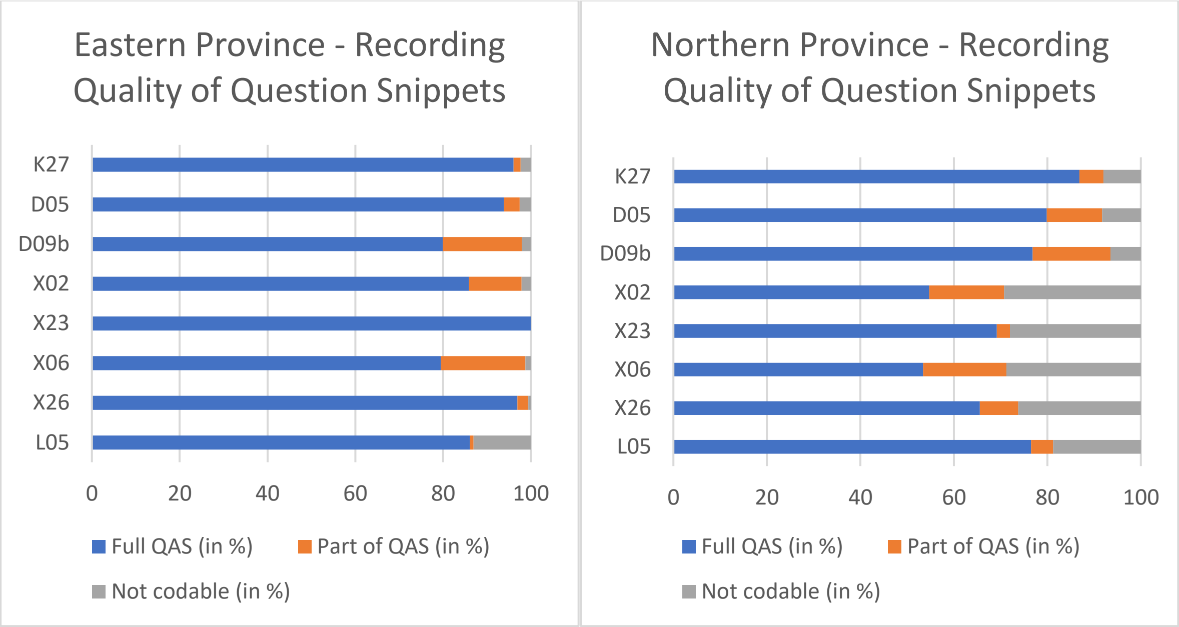

Figure 9. Recording Quality of Question Snippets per Province

Figure 9 compares the recording quality of the question per province and makes clear that the recording quality differs starkly between the two provinces. In the Eastern Province, three question snippet recordings exhibit an increased proportion of being partly recorded (D09b, X02, and X06) while the last question on the willingness to participate in a hypothetical insurance scheme (L05) produced the highest proportion of uncodable recordings. In the Northern Province, the last five questions, all situated near the end of the survey questionnaire, have concerningly high proportions of uncodable recordings, meaning that the recordings were mostly empty of spoken words.

Overall, the audio files of the Eastern Province display a higher quality for coding. The proportion of QAS which were only partly recorded are potentially connected to survey fatigue of either the respondent or the interviewer or both, or interviewers starting to ask a question before they pressed on the button to see the screen with the next question on the data collection device. In contrast, when the interviewer skips questions, it may result in uncodable recordings or partly recorded QAS which do not include the respondent's answer.

Statistical model

To answer our second research question on the association between language switching and certain traits on the question, respondent, and interviewer level, the statistical model estimates the probability that a language switch occurs. The statistical analysis includes only the 252 interviews of restricted coded sample of the Eastern Province with a prevalence of 28.6% of interviews with coded language switches. As the level of language proficiency was more varied, with only half of the respondents and interviewers speaking the survey language at home, I focus on this province. Also, Bemba coders identified language switches for only 8 interviews in this restricted analysis sample in the Northern Province. For the analysis, I pooled the question-answer sequences (QAS) of the 252 coded interviews over the eight questions of analysis which resulted in 3,245 question-answer sequences.

The statistical model consists of a three-level logistic regression model predicting the logit of the probability of language switching occurrence in a QAS of a respondent k interviewed by interviewer l. This base model or null model, where $Y_i = 1$, is composed of an overall mean (γ) and random effects on respondent- ($v\quad$), and interviewer-level ($w\quad$) with the residual variance on the QAS-level ($u\quad$). All random effects assumed to be normally distributed with a mean of zero and the respective variance.

$$logit(Pr(Y_i = 1)) \quad = \gamma \quad + u \quad + v \quad + w$$

In accordance with standard practice the proportion of the variance in $logit\ (Pr(Y_i = 1))$ due to interviewers is:

$$\rho_{interviewers} = \frac{\hat{y}_l}{\hat{y}_k + \hat{y}_l + \pi^2/3}$$

Respectively, the equation for the proportion of the variance in $logit\ (Pr(Y_i = 1))$ associated with respondents is:

$$\rho_{interviewers} = \frac{\hat{y}_l}{\hat{y}_k + \hat{y}_l + \pi^2/3}$$

To estimate the effect of the selected independent variables on the question, respondent, and interviewer level, the null model is extended as follows:

$$logit(Pr(Y_i = 1)) = \gamma \quad + \textstyle\sum_{r=1}^{m} \beta_r\, Question\_characteristics_j$$

$$+ \textstyle\sum_{s=1}^{n} \beta_s\, Respondent\_characteristics_k + \sum_{t=1}^{o} \beta_t\, Interviewer\_characteristics_l$$

$$+ v \quad + u \quad + w$$

The question characteristics included question type, response option format, number of response options, sequence number, and word count. For the respondent and interviewer, the enlarged models added both sex and a binary measure of language proficiency capturing whether the survey language was spoken at home. The models we estimated with the *melogit*-command in Stata version 18 SE.

Results

The key finding of this research is that language switching co-occurs with at least one other problematic interaction for any province and for any question. This perfect overlap between problematic interactions and language switching indicates that language switching may not be happening out of convenience or habit in a standardised interview setting. It provides initial evidence that language switching occurs when there may be a misunderstanding, as previous research has already established a link between the selected interactional behaviours and decreased data quality. Consequently, researchers can regard language switching as another problematic interactional behaviour indicating the breakdown of the cognitive response process and the survey interaction.

Figure below shows the percentages of problematic interactional behaviours per question and province. Problematic interactions vary widely between questions with more occurrences in the behavioural and knowledge questions compared to attitudinal ones. They happened foremost in the Eastern Province. The two questions emerging after the behavioural coding as most problematic for both provinces are the ones on financial behaviour involving a recall (D05 and D09B). In addition, these questions could potentially have been perceived by some respondents as sensitive (D05 capturing the amount of individual savings).

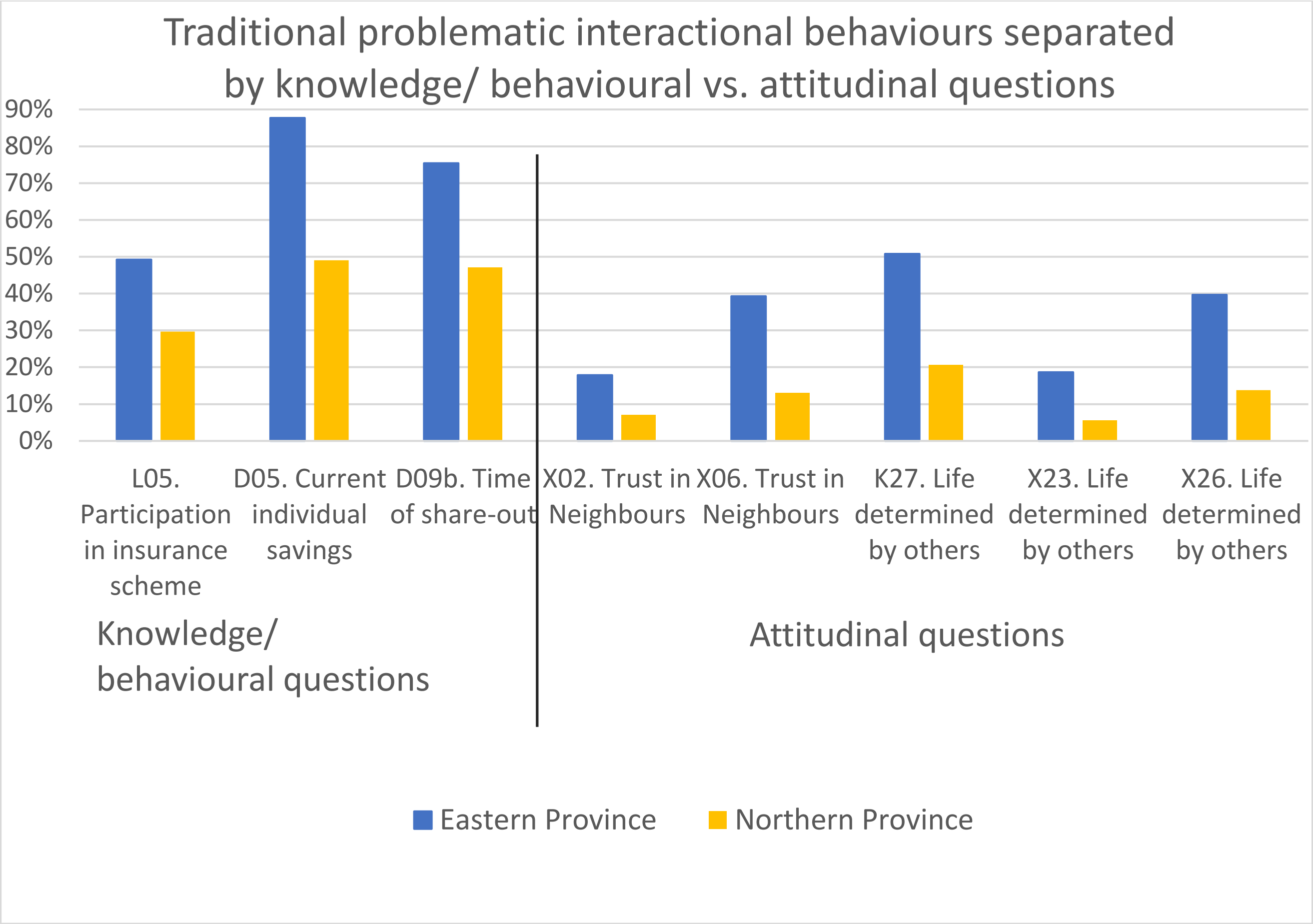

Figure 10. Frequencies of traditional problematic interactions

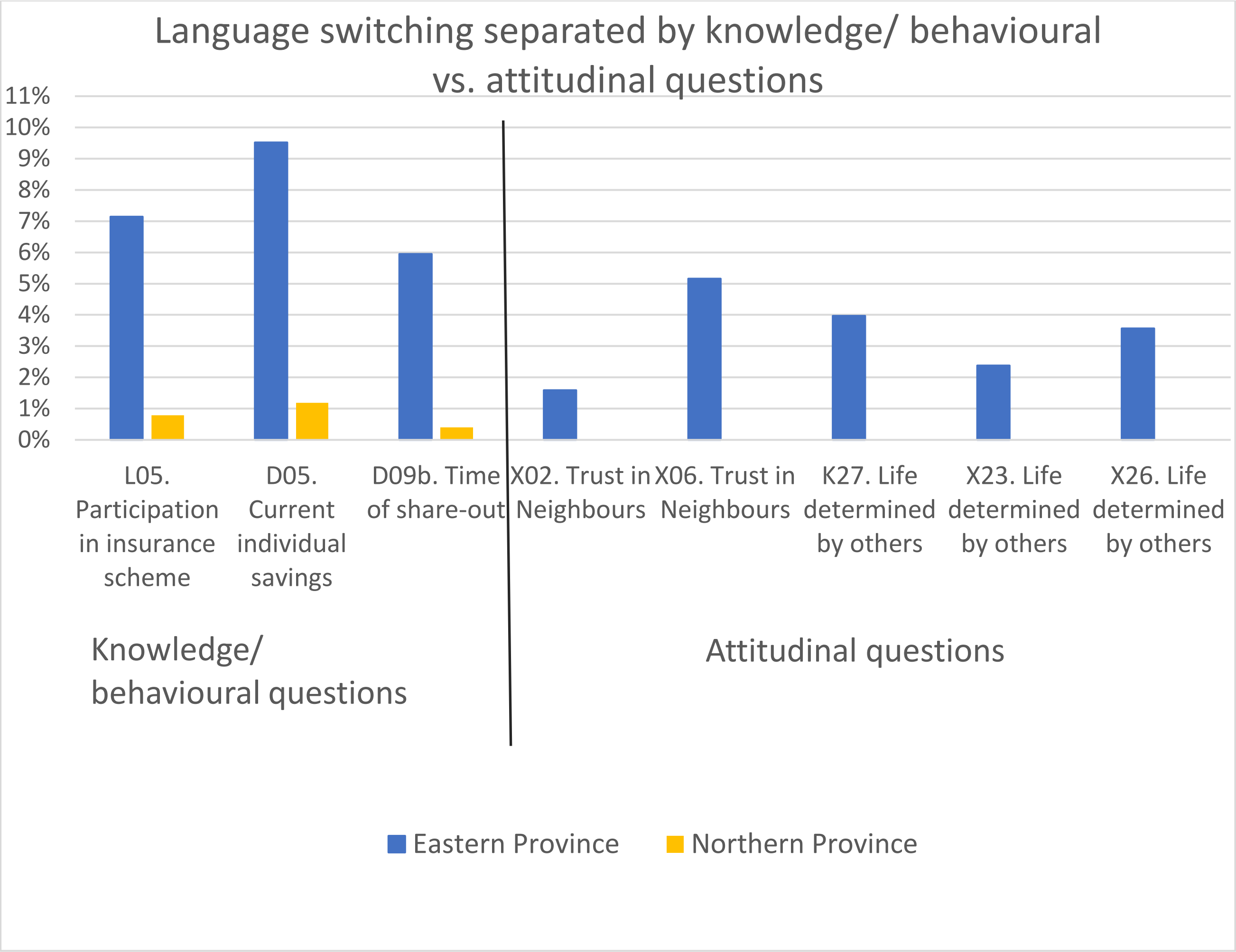

Figure 11 Language switching by question type

Figure 10 shows the frequencies of language switches in the selected question snippets for the restricted coding sample which is considerably lower than the share of multilingual interviews self-reported by the interviewer. The discrepancy suggests that the selected 6 (or 8 questions with you count the single MTMM questions) were as prone to language switching as other questions of the questionnaire which were not recorded. In congruency with the interviewer observations, language switching occurred mainly in the Eastern Province.

Albeit the percentages of language switching vary per question in restricted coding sample (see Figure 11), the frequency bars for knowledge and behavioural questions (ranging from ca. 6% to 9.5%) are visibly higher than the ones representing the attitudinal items (ranging from ca. 1.5% to 5%). The question on the amount of current individual savings in the savings group (question D05) exhibits the highest amount of language switching. In the Northern Province, no language switches are found in the selected attitudinal questions and only to a very small extent in the selected knowledge and behavioural questions.

As the attitudinal items beginning with the letter "X" are subject to the MTMM experiment (meaning that those MTMM questions were randomly allocated to the respondent based on their membership number being odd or even), they share the same question stem formulation (with minor phrasing adjustments for item-specific scales). That means that language switching should happen only when reading out the different response options and scales (refer to Table A2 for details).

For the question on trust in neighbours, the coders identified about three times as many language switches for a 11-point scale (with "0" standing for "no trust at all" and "10" signifying "complete trust") compared to a 6-point item-specific scale (ranging from complete trust, a lot of trust, some trust, a bit of trust, no trust at all). Mixed results are found for the three items measuring the external locus of control. The highest frequency of language switching was found for K27 on the extent that their life is determined by others, which was asked to all respondents using a traditional 5-point agree/disagree scale. This time, the item-specific scale with 4-points ("not at all", "a little", "very", "completely") produced slightly more language switching than the 6-point agree/disagree scale (omitting the middle category of the traditional Likert scale of "neither agree nor disagree").

Language switching associated to characteristics on question-, respondent- and interviewer-level (RQ2)

To investigate the association of language switching with selected characteristics on the question, respondent, and interviewer level, Table 14 reports on the variances and the intra-cluster correlations (ICCs) of the random respondent and interviewer effects only for the coded Eastern Province sample.

Throughout all models with and without independent variables, the ICCs capturing the variance proportion attributed to interviewers is high (more than one third of the variance is attributable to interviewers). In contrast, the variance proportion linked to respondents is negligible. The highest drop in ICCs linked to interviewers across all models is recorded for model three, the full model with all independent variables on the question, respondent, and interviewer level (5.5 percentage points). It is not surprising that the variance and ICCs fluctuate as the residual variance for logistic regression models is fixed (Baghal et al., 2014; Fielding, 2004).

Table 16. Variance and ICCs in % of respondents and interviewers (Eastern Province only)

	Null model		**Model 1 - plus question char.**		**Model 2 - plus respondent characteristics**		**Model 3 – plus interviewer characteristics**	
	Variance	ICCs	Variance	ICCs	Variance	ICCs	Variance	ICCs
Int.	2.24098	40.5%	2.44935	42.7%	2.36360	41.8%	2.07340	38.7%
Res.	0.00020	0.004%	0.00006	0.00006%	0.00006	0.00001%	0.00001	0.00002%

Discussion

This research investigates the phenomenon of language switching which may happen in multilingual context when interview partners are proficient in multiple languages and the survey language might not be the chosen language of communication during the interview. In this study, language switches are more common in the Eastern Province, where only half of the respondents and interviewers were proficient in the *lingua franca*. However, even in the Northern Province, where essentially all respondents and interviewers spoke the survey language at home, language switches still occurred for complex behavioural, knowledge questions. Translation quality was ensured by hiring external professional translators, following a team-based approach with an adjudicator resolving translation discrepancy between the two translations, piloting the translation in the survey regions with non-survey respondents from the target group and one training day between the translator and the interviewers. Given these efforts to improve understanding and lessen the cognitive burden of interviewers not translating on-the-fly, language switching could not be avoided given the multilingualism of both respondents and interviewers. Thus, it is of key importance to survey managers and methodologists who are collecting interviewer-administered data in multilingual settings to be aware of language switching.

Regardless of the province and its *lingua franca*, the results show that language switching does not occur as a single phenomenon but always co-occurs with other problematic interactional behaviours. The standardised interview setting seems to have supressed language switching out of convenience or habit. Consequently, language switching can be categorised as another problematic behaviour indicating the breakdown of the cognitive answer process and a disruption to the ideal question-answer sequence.

Nevertheless, it is also important to note that the quality of question snippet recordings, which were the base for coding interactional behaviours of both respondent and interviewer, was higher in the Eastern Province, judging by the lower frequencies of uncodable recordings.

The multilevel models showed that the variance proportion linked to the interviewer is the highest regardless of the inclusion of independent variables on the question, respondent, and interviewer level. Compared to the null model, the interviewer effects captured by the variance fraction attributed to interviewers decreased by 5.5 percentage points when including interviewer variables of sex and language proficiency. Surprisingly, the variance attributable to respondents is negligible in all models, suggesting that all variance from language switching incidences stem from the individual interviewers.

These high interviewer variances, or interviewer effects, associated with the occurrence of language switching are concerning and call for consideration whether the questionnaire should be translated into more local languages apart from the *lingua franca*. In combination with principles of standardised interviewing, interviewers encountering a respondent with insufficient level of proficiency in the survey language face a dilemma: either they stick to their training of reading the question as worded in the language of the interviewer but risk an invalid answer, or they switch to a non-survey language and deviate from standardised interviewing by translating the questions orally.

The interviewers' varying ability of oral translation in the respective non-survey language certainly is connected to the large interviewer effects. Whether the skills of oral translation can be trained so that interviewers make fewer mistakes in the field is an open question. Nevertheless, survey managers should make sure that interviewers recruited to work in a particular survey language possess the sufficient language proficiency to conduct the interviews. Foremost, more emphasis on language switching and translation during interviewer training is needed. Preferably, interviewer manuals should give guidance to the interviewers in which situations language switching is tolerated and in which instances it should be avoided.

To investigate further the extent and reason of language switching in surveys, I underline the recommendation by Saavedra Cisneros (2025) to, first, automate the capturing of language switching throughout the interview via the data collection device and systematically report the information. Although this first step is completely reliant on the interviewer diligently switching languages without any on-the-fly translation (if it can be avoided through the existence of a scripted translation), this interviewer-specific language switching may provide initial insights on at what point in time and how often language switching occurs. To be less dependent on interviewers' own reporting of their adherence to the survey protocol, artificial intelligence might enable automatic language detection from interview recordings. As behavioural coding data provide the possibility to analyse language switches by both interviewer and respondent, survey managers should consider recording a maximum of possible number of question snippets and exploit them for data quality assurance.

Overall, interactional codes could shed light on the interactions happening between respondents and interviewers during a survey interview in a multilingual setting. In particular, this study defines language switching to specifically capture instances that arose by potential misunderstandings during the survey interview. Furthermore, the unique, newly created data set of behavioural codes I present here allows one to go beyond studying interviewer effects from the given responses. As the coders were also coding the answer they heard in the audio files, future analysis using this innovative data set can investigate the language switching with data quality when recording the respondent's answer and connect this action with the problematic interviewer-respondent interactions.

Conclusion

Politicians, practitioners, and scholars alike are greatly dependent on high-quality data for decision-making, policymaking, and research. To avoid invalid findings from survey data, a paramount objective lies in minimizing survey error (Groves et al., 2009). When relying on interviewer-administered surveys, researchers continue to be both interested in and concerned about the impact of interviewers on survey data in a never-ending effort to enhance data quality (Olson et al., 2020). In standardised interviewing, the core premise to capture the "true" answer of respondents is to train and instruct the interviewers to read out the question as worded or verbatim (Fowler and Mangione, 1990). In doing so, each respondent gets asked the same questions so that the difference in answers stems from the heterogeneity of the respondents and not, for example, from different interviewers asking the questions in a different manner to different respondents, leading to interviewer effects (Dijkstra, 1983; West and Blom, 2017).

In low- and middle-income countries (LMICs), interviewer-administered surveys remain the principal data collection tool. Employing interviewers allows survey designers to account for varying levels of literacy due to the lack of universal education, particularly among rural populations. As interviewers are a key part of the data generation process, the way they pose survey questions and collect answers has a direct influence on data quality. The role of interviewers, and thus of interviewer effects, is especially pronounced LMICs, where the lack of widespread infrastructure and universal education limits the possibility for representative web surveys or partly phone surveys (modes with reduced interviewer effects).

In addition, most LMICs are multi-ethnic and multilingual, leading to interview situations in which both interviewer and respondent may converse with each other in multiple local languages. Growing up in countries with numerous local language groups and dialects, respondents and interviewers are embedded in a constant multilingual setting (at home, in the community, in school, at work, etc.). Thus, they acquire different levels of proficiency for different local languages. This type of multilingualism is fundamentally different from multilingual societies with one singular national language where individuals are rather bilingual than multilingual. Consequently, the growing literature on the role of bilingualism in survey research (see relevant chapters in Sha and Gabel, 2020) is not relevant to this Zambian context.

It is important to note that the acquisition of more than two languages is not necessary connected to a higher degree of education (as might be the case for multilingual respondents with migration background in Western countries who acquire the national language and additional foreign languages in school adding to their knowledge of their own home language). Living in multi-ethnic communities increases the constant exposure of African citizens to multiple languages. Languages switches and borrowing words from different languages are the norm in day-to-day life. Then schooling, potentially also done in a national language, fortifies the language proficiency of selected national languages. Under these diverse linguistic circumstances and given limited previous research on interviewer effects going beyond the estimation of interviewer effects, researchers and practitioners need to examine and mitigate sources of survey error stemming from the interviewer and the respondents growing up in a multilingual society.

Empirical Contributions

This PhD research set out to address the objectives stated in the title: “Assessing and improving survey data quality”. All chapters assess survey data quality from different perspective using a face-to-face panel survey in Zambia. The chapters one and two of this thesis investigate the impact of the interviewer on survey data quality in sub-Saharan Africa, first by examining the gender-of-interviewer (GOI) effect, then by estimating the magnitude of interviewer effects. Particularly, the second chapter examines how interviewer effects evolve as I implemented improvements in the scripted translation of the questionnaire in several panel waves. In doing so, it highlights the multilingual context of LMICs to pave the way for chapter 3, which investigates the phenomenon of language switching in survey interviews and their impact on survey data quality.

Below is a summary of the key empirical findings of each chapter:

- Chapter One: The findings demonstrate non-negligible interviewer effects, for behavioural, knowledge, and attitudinal questions. The results suggest that the GOI effect due to social desirability varies by question topic. Even seemingly less sensitive questions, such as village attendance in the last 12 months, may exhibit a GOI effect.
- Chapter Two: Using three panel waves with sequentially improved translation quality, the findings suggest a link between translation quality and interviewer effects. When a scripted translation into the survey languages was provided in wave 2, the interviewer effects reduced visibly over all selected 5 items on trust compared to wave 1 when interviewers were performing on-the-fly translations.
- Chapter Three: Albeit the frequencies of language switching differ among study regions, as well as over different questions, the results show that language switching does not occur as a single phenomenon, but always in co-occurrence with other problematic interactional behaviours. Thus, language switching can be considered as another problematic behaviour, indicating the breakdown of the cognitive answer process and disrupting to the ideal question-answer sequence.

The findings combined certainly highlight the special context of surveys in multilingual countries which also exhibit lower level of “survey literacy”, meaning there is a lack of understanding the purpose of standardised research (Massey, 2025). In addition, 7% to 20% of the respondents of wave 1 (depending on province) indicated that they have never been to school. For ZamSaP, the concept of survey literacy can also be somewhat extended to the interviewers, especially those with little survey experience. On the side of the local survey firms, challenges lay in the ignorance of the importance of scripted questionnaire translation into (more) local languages and of rigorous translation procedures adhering to a team-based approach as gold standard. In addition, there was an inadequate emphasis on (standardised) language proficiency and the assessment thereof for the interviewers. The impression left is that the selection of interviewers was heavily based on self-reported proficiency levels and/or superficial checking on. Certainly, a lack of interviewers with the right language skills could have also played a role as half of the interviewers for the Chewa interviews did not speak the survey language at home.

All in all, the insights of this doctoral research also pose questions on how survey methods, often primarily derived from Western context, can be replicated in a non-Western one such as Zambia. Different processes such as social desirability bias may be influenced by different societal and local power dynamics. Certain interviewer characteristics, such as the gender of the interviewer, might play a more prominent role and certainly the added cognitive burden of multilingualism for both interviewer and respondents may shape survey data in different ways. The dominant multilingual context calls for revisiting the total survey error, particularly the component of the measurement part.

Practical implications

As this PhD research is assessing survey data quality in Zambia, a multilingual country in sub-Saharan Africa, through the principal lens of interviewer performance, the implications for survey practice mostly revolve around the objective of analysing and minimising measurement error attributable to interviewers working in a multilingual setting. The findings of this doctoral thesis shed light on interviewer effects in general and the complex cognitive processes in a multilingual environment.

For any investigation focusing on interviewer performance and subsequent thorough analysis of interviewer effects, survey data need to at least provide a unique identifier for different interviewers. I strongly echo the advice provided by Stecklov and Weinreb (2010) on randomising the assignment of the interviewer to the respondents when possible. When respondents are clustered in certain units (here a savings group, but it could be villages or enumeration areas) while interviewer teams work by moving from one cluster to the next cluster, their interview assignment can be randomised within that cluster. In cases where this interpenetrated design is too costly or logistically not feasible due to the geographic spread of the study area, more complex statistical multilevel models can address the issue of separating interviewer from area effects. In addition; a dedicated interviewer survey can capture interviewer characteristics in detail, especially unobservable ones, such as the survey experience, the interviewers' own attitudes towards the survey topics, etc.

To move beyond the mere calculation of interviewer effects towards a more fine-grained study of the effect of certain interviewer characteristics (such as gender, age, etc.), the provision of this information and their incorporation in the data set needs to be agreed upon with the survey organisation and interviewers at the onset. By doing so, findings from studies on the impact of specific interviewer characteristics may widen and deepen our understanding of interviewer effects.

As interviewer effects vary by question type and topic (see West and Blom, 2017), it is advisable to consider estimating interviewer effects for a variety of questions not limited to sensitive questions doubtlessly inducing social desirability. Especially in understudied contexts, such as in Zambia, mechanisms of social desirability found in previous studies outside of sub-Saharan Africa or from countries with lower level of patriarchy, can potentially not be replicated.

As part of good survey practice, survey practitioners should provide a scripted questionnaire translation at least in the *lingua franca* of the study area to avert the increased cognitive burden of on-the-fly translation on the interviewer. This PhD research specifically underlined the importance of translation mode interviewer effects in a multilingual setting by emphasising the connection between different modes of questionnaire translations and varying levels of interviewer variances. The objective to improve interviewer performance and avoid measurement error due to translation can be achieved by providing a well-translated questionnaire, following high standards in comparing independent translations with team-based review and documentation processes, and pre-testing properly in the study area (for good translation practices, see for example Behr, 2023; Harkness et al., 2010).

To evaluate the necessity of translating the source questionnaire in more local languages, survey managers of panel studies should capture the respondents' proficiency in the survey language and in other local languages in the first wave. If the budget does not allow for more translations in additional local languages, methods to gauge the extent of misunderstandings or limited language capabilities in answering questions should be put in place. The questionnaire instrument could incorporate interviewer observations on the respondents' proficiency in the survey language or the amount of language switching in an interview alongside with potential questions to the respondents about their individual linguistic repertoire and proficiency.

In case, survey firms offer to record a fixed number of question snippets for quality assurance, these recordings may become an additional source for analysing the black box of a multilingual survey interview. Survey managers could strategically choose to record complex questions which were flagged during the translation documentation process. These recordings are a valuable source to evaluate survey data quality and specifically language switching as an indicator for interactional problems during the interview. Researchers can also exploit those to further deepen our understanding of how cognitive models of interactions between interviewers and respondents (such as those developed by Ongena and Dijkstra, 2007) can be adapted to a multilingual context.

Finally, survey managers should design specific assessment scores based on standardised tests to measure the interviewer's language proficiency. Solely basing the selection of interviews on self-reported language competency is not sufficient. This PhD research could demonstrate that the frequency of problematic interactions, among them language switching, is highest in the study area where only half of the respondents and interviewers spoke the survey language at home. More emphasis during interviewer training and a clear rule or approach is needed in the survey protocol explaining on when language switching and on-the-fly translation is allowed.

Future research opportunities

This doctoral research has made the first step to explore survey data error in LMICs, however, future research should further explore under which conditions and through which mechanisms survey data quality may deteriorate in LMICs and in multilingual contexts. The findings of Chapter One on the GOI effect in behavioural and knowledge, as well as attitudinal questions call for a re-investigation of theoretical models on when and why an GOI effect might happen. As social desirability bias does not always reproduce across cultures (Johnson and van de Vijver, 2003), survey researchers need to expand the theoretical models and processes to understudied cultures, such as countries in sub-Saharan Africa. Although one rule of thumb for researchers is to expect social desirability bias connected to certain question topics, such as abortion, sexual behaviour, etc., which are without doubt sensitive across different cultures (for the connection between social desirability and sensitive question, see Yan, 2021). However, Johnson and Van der Vijer (2003) underline that the relationship between cultural and individual determine when and whether a social desirability bias arises. For surveys in LMICs, which are predominantly interviewer-administered, survey methods research needs to build upon more empirical evidence when sensitivity of a question is induced and how it is related to the interaction between the interviewer and respondents' characteristics.

As with any research topic, experimental settings can further strengthen the investigation of causal process and mechanisms. As Chapter Two falls short in establishing a causal relationship between improved translation quality and decreased interviewer effects, future research should examine response errors stemming from on-the-fly translation (referred to as mode A) compared to those occurring with the use of a scripted questionnaire translation of high quality (here mode B). The experimental design allows us to better understand under which circumstances interviewers stick to the scripted translated questionnaire or attempt an on-the-fly translation when faced with a respondent likely to prefer a different, unscripted language.

On potential survey design is to randomly assign interviewers to one of the two experimental branches (mode A vs. mode B). The interviewers in the mode A group should only receive one local scripted questionnaire translated into the region-specific *lingua franca* while the interviewers in the mode B group should have all scripted translation into several local language to their disposal. Lacking scripted translation in multiple languages, interviewers of the second experimental group will face the decision of translating the questionnaire on-the-fly more often. A difference in survey data quality in those two experimental groups of interviewers would establish a causal link between the mode of translation and survey data quality (for a similar experimental research design in a non-laboratory setting comparing standardised and conversational interviewing, see West et al., 2018).

For studying the circumstances of language switching in an interviewer through behavioural coding, the full interview should be recorded and coded. Through the recording of the full interview, survey researchers can make pinpoint and draw conclusions not only on which interactions between interviewers and respondents are problematic and what role language switching plays in a sequence of problematic interactions, but also whether certain questions cause misunderstandings more often than others. Studying the context of language switching together with respondents and interviewer characteristics (especially regarding their level of language proficiency) offers the possibility to disentangle the three different levels (question, respondent, and interviewer), to gain a better comprehension of this peculiar phenomenon. In a changing world, becoming more diverse and more multilingual, the African context does not only provide insights to how to identify different sources of survey error, but also produce insights which then can guide researchers beyond sub-Saharan Africa.

Bibliography

Amos, M. (2018). Evaluating the impact on the reasons for contraceptive nonuse in the Indonesia and the Philippines DHS. *Demographic Research*, *39*, 415–430. https://www.jstor.org/stable/26585335

Axinn, W. G. (1989). Interviewers and Data Quality in a Less Developed Setting. *Journal of Official Statistics*, *5*(3), 265. https://www.proquest.com/scholarly-journals/interviewers-data-quality-less-developed-setting/docview/1266808180/se-2?accountid=14570

Axinn, W. G. (1991). The influence of interviewer sex on responses to sensitive questions in Nepal. *Social Science Research*, *20*(3), 303–318. https://doi.org/10.1016/0049-089X(91)90009-R

Baghal, T. A., Belli, R. F., PHILLIPS, A. L., & Ruther, N. (2014). What Are You Doing Now? Activity-Level Responses and Recall Failures in the American Time Use Survey. *Journal of Survey Statistics and Methodology*, *2*(4), 519–537. https://doi.org/10.1093/JSSAM/SMU020

Beatty, P. (1995). Understanding the Standardized/Non-Standardized Interviewing Controversy. *Journal of Official Statistics*, *11*(2), 147. https://www.proquest.com/scholarly-journals/understanding-standardized-non-interviewing/docview/1266820500/se-2?accountid=14570

Becker, S., Feyisetan, K., & Makinwa-Adebusoye, P. (1995). The Effect of the Sex of Interviewers on the Quality of Data in a Nigerian Family Planning Questionnaire. *Studies in Family Planning*, *26*(4), 233. https://doi.org/10.2307/2137848

Behr, D. (2023). Translating Questionnaires. In C. U. Krägeloh, M. Alyami, & O. N. Medvedev (Eds.), *International Handbook of Behavioral Health Assessment* (pp. 1–15). Springer. https://doi.org/10.1007/978-3-030-89738-3_2-1

Benstead, L. J., & Hatfield, M. O. (2014). Effects of Interviewer–Respondent Gender Interaction on Attitudes toward Women and Politics: Findings from Morocco. *International Journal of Public Opinion Research*, *26*(3), 369–383. https://doi.org/10.1093/IJPOR/EDT024

Berg, E. & Edwards, B. (2025). A Message from the Editors. Journal of Survey Statistics and Methodology, 13(1), 1–2. https://doi.org/10.1093/jssam/smaf001

Bignami-Van Assche, S., Reniers, G., & Weinreb, A. A. (2003). An Assessment of the KDICP and MDICP Data Quality: Interviewer Effects, Question Reliability and Sample Attrition. *Demographic Research*, *S1*, 31–76. https://doi.org/10.4054/DEMRES.2003.S1.2

Blaydes, L., & Gillum, R. M. (2013). Religiosity-of-Interviewer Effects: Assessing the Impact of Veiled Enumerators on Survey Response in Egypt. *Politics and Religion*, *6*(3), 459–482. https://doi.org/10.1017/S1755048312000557

Boztepe, E. (2003). Issues in Code-Switching: Competing Theories and Models. *Studies in Applied Linguistics and TESOL*, *3*(2). https://doi.org/10.7916/SALT.V3I2.1626

Buzasi, K. (2016). Linguistic situation in twenty sub-Saharan African countries: A survey-based approach. *African Studies*, *75*(3), 358-380.

Bylund, E., & Athanasopoulos, P. (2014). Language and thought in a multilingual context: The case of isiXhosa. *Bilingualism: Language and Cognition*, *17*(2), 431–441. https://doi.org/10.1017/S1366728913000503

Carli, L. L. (1999). Gender, Interpersonal Power, and Social Influence. *Journal of Social Issues*, *55*(1), 81–99. https://doi.org/10.1111/0022-4537.00106

Cleland, J., & Verma, V. (1989). The World Fertility Survey: An Appraisal of Methodology. *Journal of the American Statistical Association*, *84*(407), 756–767. https://doi.org/10.1080/01621459.1989.10478831

Conrad, F. G., & Schober, M. F. (2021). Clarifying question meaning in standardized interviews can improve data quality even though wording may change: a review of the evidence. *International Journal of Social Research Methodology*, *24*(2), 203–226. https://doi.org/10.1080/13645579.2020.1824627

Conroy-Krutz, J. (2019). Surveys and Their Use in Understanding African Public Opinion. *Oxford Research Encyclopedia of Politics*. https://doi.org/10.1093/ACREFORE/9780190228637.013.820

Davis, R. E., Couper, M. P., Janz, N. K., Caldwell, C. H., & Resnicow, K. (2010). Interviewer effects in public health surveys. *Health Education Research*, *25*(1), 14–26. https://doi.org/10.1093/HER/CYP046

Demarest, L. (1997). *An Assessment of Interviewer Error in the Afrobarometer Project. CRPD Working Paper* (53), most recently retrieved from https://lirias.kuleuven.be/retrieve/462310 on 6 January 2026.

Demirguc-Kunt, A., Klapper, L., Singer, D., Ansar, S., & Hess, J. (2018). *The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1259-0

Devarajan, S. (2013). Africa's Statistical Tragedy. *Review of Income and Wealth*, *59*(SUPPL1), S9–S15. https://doi.org/10.1111/ROIW.12013

Dijkstra, W. (1983). How Interviewer Variance can Bias the Results of Research on Interviewer Effects. *Quality & Quantity*, *17*(3), 179. https://doi.org/10.1007/BF00167582

Di Maio, M., & 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), 654–669. https://doi.org/10.1093/WBER/LHY024

Dwivedi, L. K., Banerjee, K., Sharma, R., Mishra, R., Ramesh, S., Sahu, D., Mohanty, S. K., & James, K. S. (2022). Quality of anthropometric data in India's National Family Health Survey: Disentangling interviewer and area effect using a cross-classified multilevel model. *SSM - Population Health*, *19*, 101253. https://doi.org/10.1016/J.SSMPH.2022.101253

Dykema, J., Garbarski, D., Wall, I. F., Farrar Edwards, D., Assad, N., Renteria, J., Garza, R., Blixt, S., Schaeffer, N. C., & Stevenson, J. (2019). Measuring Trust in Medical Researchers: Adding Insights from Cognitive Interviews to Examine Agree-Disagree and Construct-Specific Survey Questions. *Journal of Official Statistics*, *35*(2), 353–386. https://doi.org/10.2478/JOS-2019-0017

Dykema, J., Schaeffer, N. C., Garbarski, D., Nordheim, E. V., Banghart, M., & Cyffka, K. (2016). The Impact of Parenthetical Phrases on Interviewers' and Respondents' Processing of Survey Questions. *Survey Practice*, *9*(2), 10.29115/SP-2016–0008. https://doi.org/10.29115/SP-2016-0008

Evans, A. (2014). 'Women Can Do What Men Can Do': The Causes and Consequences of Growing Flexibility in Gender Divisions of Labour in Kitwe, Zambia. *Journal of Southern African Studies*, *40*(5), 981–998. https://doi.org/10.1080/03057070.2014.946214

Fellegi, I. P. (1964). Response Variance and its Estimation. *Journal of the American Statistical Association*, *59*(308), 1016–1041. https://doi.org/10.1080/01621459.1964.10480747

Fellegi, I. P. (1974). An improved method of estimating the correlated response variance. *Journal of the American Statistical Association*, *69*(346), 496–501. https://doi.org/10.1080/01621459.1974.10482982

Fielding, A. (2004). Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models. *Quality and Quantity*, *38*(4), 425–433. https://doi.org/10.1023/B:QUQU.0000043118.19835.6C/METRICS

Flores-Macias, F., & Lawson, C. (2008). Effects of Interviewer Gender on Survey Responses: Findings from a Household Survey in Mexico. *International Journal of Public Opinion Research*, *20*(1), 100–110. https://doi.org/10.1093/IJPOR/EDN007

Footman, K. (2021). Interviewer effects on abortion reporting: a multilevel analysis of household survey responses in Côte d'Ivoire, Nigeria and Rajasthan, India. *BMJ Open*, *11*(11), e047570. https://doi.org/10.1136/BMJOPEN-2020-047570

Fowler, F. J., & Mangione, T. W. (1990). *Standardized Survey Interviewing : minimizing interviewer-related error*. Sage.

Frankovic, K., Jodice, D. A., Kizilova, K., & See Toh, W. Y. (2023). *The Freedom to Conduct and Publish Opinion Polls: A 2023 worldwide update, World Association for Public Opinion Research & European Society for Opinion and Market Research, most recently retrieved from https://wapor.org/wp-content/uploads/Freedom-to-Conduct-and-Publish-Opinion-Polls-v8-1.pdf on 6 January 2026*.

Garbarski, D., Schaeffer, N. C., & Dykema, J. (2011). Are interactional behaviors exhibited when the self-reported health question is asked associated with health status? *Social Science Research*, *40*(4), 1025–1036. https://doi.org/10.1016/J.SSRESEARCH.2011.04.002

Grosh, M. E., & Glewwe, P. (1996). Household Survey Data from Developing Countries: Progress and Prospects. *The American Economic Review*, *86*(2), 15–19. http://www.jstor.org/stable/2118088

Groves, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (R. M. Groves, F. J. Fowler, M. Couper, J. M. Lepkowski, E. Singer, & R. Tourangeau, Eds.; Second edition). Wiley.

Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, Adaptation, and Design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey Methods in Multicultural, Multinational, and Multiregional Contexts* (p. 114). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470609927.CH7

Harkness, J., Schoebi, N., Joye, D., Mohler, P., Faass, T., & Behr, D. (2008). Oral translation in telephone surveys. In J. M. Lepkowski (Ed.), *Advances in Telephone Survey Methodology* (pp. 231–249). John Wiley & Sons.

Harling, G., Chanda, M. M., Ortblad, K. F., Mwale, M., Chongo, S., Kanchele, C., Kamungoma, N., Barresi, L. G., Bärnighausen, T., & Oldenburg, C. E. (2019). The influence of interviewers on survey responses among female sex workers in Zambia. *BMC Medical Research Methodology*, *19*(1), 1–12. https://doi.org/10.1186/S12874-019-0703-2/TABLES/5

Hathi, P., Coffey, D., Thorat, A., & Nagle, A. (2025). Does matching interviewer and respondent gender improve data quality and reduce social desirability bias? Evidence from a mobile phone survey in India. *Survey Methods: Insights from the Field*. https://surveyinsights.org/?p=20157

Himelein, K. (2016). Interviewer Effects in Subjective Survey Questions: Evidence From Timor-Leste. *International Journal of Public Opinion Research*, *28*(4), 511–533. https://doi.org/10.1093/IJPOR/EDV031

Holbrook, A., Cho, Y. I., & Johnson, T. (2006). The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties. *Public Opinion Quarterly*, *70*(4), 565–595. https://doi.org/10.1093/POQ/NFL027

Hughes, S. M., & Lin, Y. (Jay). (2018). Survey Data Collection in Sub-Saharan Africa (SSA). In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), A*dvances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)* (pp. 533–554). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118884997.CH25

Jacobsen, J. (2022). If They Don't Understand the Question, They Don't answer. Language Mismatch in Face-to-Face Interviews. *Journal of Official Statistics*, *38*(2), 453–484. https://doi.org/10.2478/JOS-2022-0022

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/J.JML.2007.11.007

Jerven, M., & Johnston, D. (2015). Statistical Tragedy in Africa? Evaluating the Data Base for African Economic Development. *The Journal of Development Studies*, *51*(2), 111–115. https://doi.org/10.1080/00220388.2014.968141

Johnson, T. P., Pennell, B.-E., Stoop, I. A. L., & Dorer, B. (2019). *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)* (T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer, Eds.). John Wiley & Sons.

Johnson, T. P., & van de Vijver, F. J. R. (2003). Social Desirability in Cross-Cultural Survey Methods. In J. A. Harkness, F. J. R. van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (1st ed., pp. 193–209). Wiley. https://research.tilburguniversity.edu/en/publications/cross-cultural-survey-methods

Kaani, B., & Joshi, R. M. (2023). Literacy Practices in Zambia: Becoming Literate in a Multilingual Classroom. In R. M. Joshi, C. A. Bride, B. Kaani, & G. Elbeheri (Eds.), *Handbook of Literacy in Africa* (pp. 405–437). Springer.

Kianersi, S., Luetke, M., Jules, R., & Rosenberg, M. (2020). The association between interviewer gender and responses to sensitive survey questions in a sample of Haitian women. *International Journal of Social Research Methodology*, *23*(2), 229–239. https://doi.org/10.1080/13645579.2019.1661248

Lau, C. Q., Eckman, S., Sevilla Kreysa, L., & Piper, B. (2020). Language Differences Between Interviewers and Respondents in African Surveys. In M. Sha & T. Gabel (Eds.), *The Essential Role of Language in Survey Research* (pp. 101–115). RTI Press. https://doi.org/10.3768/rtipress.bk.0023.2004

Lau, C. Q., Marks, E., & Gupta, A. K. (2018). Survey Research in India and China. 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. 583–596). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118884997.CH28

Leone, T., Sochas, L., & Coast, E. (2021). Depends Who's Asking: Interviewer Effects in Demographic and Health Surveys Abortion Data. *Demography*, *58*(1), 31–50. https://doi.org/10.1215/00703370-8937468

Lindberg, L. D., Maddow-Zimet, I., Mueller, | Jennifer, & Vandevusse, A. (2022). Randomized experimental testing of new survey approaches to improve abortion reporting in the United States. *Perspectives on Sexual and Reproductive Health*, *54*(4), 142–155. https://doi.org/10.1363/PSRH.12217

Liu, M., & Stainback, K. (2013). Interviewer gender effects on survey responses to marriage-related questions. *Public Opinion Quarterly*, *77*(2), 606–618. https://doi.org/10.1093/poq/nft019

Liu, M., & Wang, Y. (2016). Interviewer Gender Effect on Acquiescent Response Style in 11 Asian Countries and Societies. *Field Methods*, *28*(4), 327–344. https://doi.org/10.1177/1525822X15623755

Lupu, N., & Michelitch, K. (2018). Advances in Survey Methods for the Developing World. *Annual Review of Political Science*, *21*(Volume 21, 2018), 195–214. https://doi.org/10.1146/ANNUREV-POLISCI-052115-021432/1

Mangione, T. W., Fowler, F. J., & Louis, T. A. (1992). Question Characteristics and Interviewer Effects. *Journal of Official Statistics*, *8*(3), 293. https://www.proquest.com/scholarly-journals/question-characteristics-interviewer-effects/docview/1266807067/se-2?accountid=14570

Marten, L., & Kula, N. C. (2008). Zambia: One Zambia, One Nation, Many Languages. In A. Simpson (Ed.), *Language and National Identity in Africa* (pp. 291–313). Oxford University Press. https://soas-repository.worktribe.com/output/428838

Massey, M. (2025). Survey Practice in Non-Survey-Literate Populations: Lessons Learned from a Cognitive Interview Study in Brazil. *Survey Practice*, *19*(SI). https://doi.org/10.29115/SP-2024-0035

Mueller, J., Kirstein, M., VandeVusse, A., & Lindberg, L. D. (2023). Improving abortion underreporting in the USA: a cognitive interview study. *Culture, Health & Sexuality*, *25*(1), 126–141. https://doi.org/10.1080/13691058.2022.2113434

Olson, K., Smyth Amanda Ganshert, J. D., OLSON is the Leland, K. J., Olson Associate Professor, D. H., & Chair, V. (2019). The Effects of Respondent and Question Characteristics on Respondent Answering Behaviors in Telephone Interviews. *Journal of Survey Statistics and Methodology*, *7*(2), 275–308. https://doi.org/10.1093/JSSAM/SMY006

Olson, K., Smyth, J. D., & Cochran, B. (2018). Item Location, the Interviewer–Respondent Interaction, and Responses to Battery Questions in Telephone Surveys. *Sociological Methodology*, *48*(1), 225–268. https://doi.org/10.1177/0081175018778299/SUPPL_FILE/BATTERY_TABLES_REV2_APPENDIX.PDF

Olson, K., Smyth, J. D., Dykema, J., Holbrook, A. L., Kreuter, F., & West, B. T. (2020). *Interviewer Effects from a Total Survey Error Perspective*. CRC Press.

O'Muircheartaigh, C. A. (1976). Response errors in an attitudinal sample survey. *Quality and Quantity*, *10*(2), 97–115. https://doi.org/10.1007/BF00144162/METRICS

O'Muircheartaigh, C. A. (1982). *Methodology of the Response Errors Project*. World Fertility Survey, London, GB. http://hdl.handle.net/10625/8807

O'Muircheartaigh, C., & Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(3), 437–446. https://doi.org/10.1111/1467-985X.00147

Ong, A. R., Hu, M., West, B. T., & Kirlin, J. A. (2018). Interviewer effects in food acquisition surveys. *Public Health Nutrition*, *21*(10), 1781–1793. https://doi.org/10.1017/S1368980018000137

Ongena, Y. P., & Dijkstra, W. (2007). A model of cognitive processes and conversational principles in survey interview interaction. *Applied cognitive psychology: The official journal of the society for applied research in memory and cognition*, *21*(2), 145-163.

Peytcheva, E. (2020). The Effect of Language of Survey Administration. In M. Sha & T. Gabel (Eds.), *The essential role of language in survey research* (pp. 3–21). RTI Press.

Phung, T. D., Hardeweg, B., Praneetvatakul, S., & Waibel, H. (2015). Non-Sampling Error and Data Quality: What Can We Learn from Surveys to Collect Data for Vulnerability Measurements? *World Development*, *71*, 25–35. https://doi.org/10.1016/J.WORLDDEV.2013.11.008

Pietrelli, R., d'Errico, M., & Dassesse, K. (2021). Measuring household food security through surveys: Do the characteristics of the enumerators matter? *Development Policy Review*, *39*(6), 911–925. https://doi.org/10.1111/DPR.12534

Randall, S., Coast, E., Compaore, N., & Antoine, P. (2013). The power of the interviewer. *Demographic Research*, *28*, 763–792. http://www.jstor.org/stable/26349970

Rozelle, J. W., Meyer, M. J., McKenna, A. H., Obaje, H., & Kraemer, J. D. (2023). The effect of interviewer-respondent age difference on the reporting of sexual activity in the Demographic and Health Surveys: Analysis of data from 21 countries. *Journal of Global Health*, *13*, 04002. https://doi.org/10.7189/JOGH.13.04002

Saavedra Cisneros, A. (2025). Language choices in surveys: how switching language of interview highlights both identity and acculturation. *Survey Practice*, *19* (SI). https://doi.org/10.29115/SP-2025-0001

Sarac, M., & Koc, I. (2021). Sources of nonresponse error in the translation process of survey instruments: the impact of language mismatch and on-the-spot translation on the quality of birth date data. *International Journal of Social Research Methodology*, 1–13. https://doi.org/10.1080/13645579.2020.1785088

Schaeffer, N. C., & Dykema, J. (2011). Response 1 to Fowler's Chapter: Coding the Behavior of Interviewers and Respondents to Evaluate Survey Questions . In J. Madans, K. Miller, A. Maitland, & G. Willis (Eds.), *Question Evaluation Methods: Contributing to the Science of Data Quality* (pp. 23–39). Wiley Blackwell. https://doi.org/10.1002/9781118037003

Schaeffer, N. C., Dykema, J., & Maynard, D. W. (2010). Interviewers and Interviewing. In P. Marsden & J. D. Wright (Eds.), *Handbook of Survey Research* (2nd ed., pp. 437–470). Emerald.

Schaeffer, N. C., Garbarski, D., & Dykema, J. (2016). Introduction to Special Issue of Survey Practice on Interviewer-Respondent Interaction. *Survey Practice*, *9*(2), 1–3. https://doi.org/10.29115/SP-2016-0007

Schaeffer, N. C., & Maynard, D. W. (2005). From paradigm to prototype and back again: Interactive aspects of 'cognitive processing' in standardized survey interviews. In H. te Molder & J. Potter (Eds.), *Conversation and Cognition* (pp. 114–133). Cambridge University Press.

Schaeffer, N. C., & Presser, S. (2003). The Science of Asking Questions. *Annual Review of Sociology*, *29*(Volume 29, 2003), 65–88. https://doi.org/10.1146/ANNUREV.SOC.29.110702.110112/CITE/REFWORKS

Schnell, R., & Kreuter, F. (2005). Separating Interviewer and Sampling-Point Effects. *Journal of Official Statistics*, *21*(3).

Schober, M. F., & Bloom, J. E. (2004). Discourse Cues That Respondents Have Misunderstood Survey Questions. *Discourse Processes*, *38*(3), 287–308. https://doi.org/10.1207/S15326950DP3803_1

Scott, C., Vaessen, M., Coulibaly, S., & Verrall, J. (1988). Verbatim Questionnaires versus Field Translation or Schedules: An Experimental Study. *International Statistical Review / Revue Internationale de Statistique*, *56*(3), 259. https://doi.org/10.2307/1403353

Sha, M., & Gabel, T. (2020). *The essential role of language in survey research*. RTI Press.

Sharma, R., Dwivedi, L. K., Jana, S., Banerjee, K., Mishra, R., Mahapatra, B., Sahu, D., & Singh, S. K. (2022). Survey implementation process and interviewer effects on skipping sequence of maternal and child health indicators from National Family Health Survey: An application of cross-classified multilevel model. *SSM - Population Health*, *19*, 101252. https://doi.org/10.1016/J.SSMPH.2022.101252

Singh, A., Kumar, K., & Arnold, F. (2022). How Interviewers Affect Responses to Sensitive Questions on the Justification for Wife Beating, the Refusal to have Conjugal Sex, and Domestic Violence in India. *Studies in Family Planning*, *53*(2), 259–279. https://doi.org/10.1111/SIFP.12193

Singh, S., Dwivedi, L. K., & Jana, S. (2024). Fieldworker effects on quality of data collected on sensitive questions in National Family Health Surveys: An analysis based on intimate partner violence in India. *SSM - Population Health*, *25*, 101557. https://doi.org/10.1016/J.SSMPH.2023.101557

Stecklov, G., & Weinreb, A. (2010). *Improving the Quality of Data and Impact-Evaluation Studies in Developing Countries*, Impact-Evaluation Guidelines, Technical Notes (No. IDB-TN-123), most recently retrieved from https://publications.iadb.org/publications/english/document/Improvingthe-Quality-of-Data-and-Impact-Evaluation-Studies-in-Developing-Countries.pdf on 6 January 2026.

Strauss, J., & Thomas, D. (1996). Measurement and Mismeasurement of Social Indicators. *The American Economic Review*, *86*(2), 30–34. http://www.jstor.org/stable/2118091

Sundström, A., & Stockemer, D. (2022). Measuring support for women's political leadership: Gender of interviewer effects among African survey respondents. *Public Opinion Quarterly*, *86*(3), 668–696. https://doi.org/10.1093/poq/nfac031

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Tu, S. H., & Liao, P. S. (2007). Social distance, respondent cooperation and item nonresponse in sex survey. *Quality and Quantity*, *41*(2), 177–199. https://doi.org/10.1007/S11135-007-9088-0/METRICS

Weinreb, A. A., & Sana, M. (2009). The effects of questionnaire translation on demographic data and analysis. *Population Research and Policy Review*, *28*(4), 429–454. https://doi.org/10.1007/S11113-008-9106-5/TABLES/5

Wenz, A., Al Baghal, T., & Gaia, A. (2021). Language Proficiency Among Respondents: Implications for Data Quality in a Longitudinal Face-To-Face Survey. *Journal of Survey Statistics and Methodology*, *9*(1), 73–93. https://doi.org/10.1093/JSSAM/SMZ045

West, B. T., & Blom, A. G. (2017). Explaining Interviewer Effects: A Research Synthesis. *Journal of Survey Statistics and Methodology*, *5*(2), 175–211. https://doi.org/10.1093/JSSAM/SMW024

West, B. T., Conrad, F. G., Kreuter, F., & Mittereder, F. (2018). Can Conversational Interviewing Improve Survey Response Quality Without Increasing Interviewer Effects? *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*(1), 181–203. https://doi.org/10.1111/RSSA.12255

Woolfrey, L. (2009). Knowledge Utilization for Governance in Africa: evidence-based decision-making and the role of survey data archives in the region. *Information Development*, *25*(1), 22–32. https://doi.org/10.1177/0266666908101261

Wu, Q., & Xie, Y. (2024). Interviewer Ratings of Physical Appearance in a Large-Scale Survey in China. *Journal of Survey Statistics and Methodology*, *12*(4), 987–1010. https://doi.org/10.1093/JSSAM/SMAD046

Yan, T. (2021). Consequences of asking sensitive questions in surveys. *Annual Review of Statistics and Its Application*, *8*(Volume 8, 2021), 109–127. https://doi.org/10.1146/ANNUREV-STATISTICS-040720-033353/CITE/REFWORKS

Appendices

Table A1. Two examples for language switching from the transcript of one interview

Onset time	Offset time	* { Chewa transcript }* and ** { Nsenga transcript } **
00:34	00:35	* { I: Ok muli nadzaka zingati? } * [I: Ok how old are you?]
00:36	00:37	** { R: Um kaya naluwa newo. } ** [R: Um I don't know.]
00:38	00:39	* { I: Uh? } *
00:39	00:40	** { R: Naluwa. Haha. } ** [R: I don't know. Haha.]
00:40	00:41	** { I: Dzaka zanu zomwe mwaluwa? } ** [I: You don't know how old you are?]
00:41	00:42	** { R: Ee naluwa. } ** [R: Yes, I don't know.]
04:12	04:20	* { I: So tingati mumawakhulupilila bwanji a Zanaco. Tingati koteratu, kwambiri, pang'ono kapena simumawakhulupilila? } * [I: So how much do you trust Zanaco? Can we say completely, much, a bit, not at all?]
04:22	04:25	* { R: Ati a Zanaco timawakhulupilila bwanji? } * [R: You mean how much we trust Zanaco?]
04:26	04:34	* [I: Ee, a Zanaco Bank. Kwainu mweka osati tima. Mumawakhulupilila bwanji? } *

		[I: Yes, how much you trust Zanaco personally not collectively. How much do you trust it?]
04:35	04:37	** { R: haha. Owo. Uh kambani soti. } ** [R: haha. Ok. Uh can you say it again.]
04:38	04:38	* { I: Uh?} *
04:39	04:40	** { R: Kambani soti? } ** [R: Can you say it again?]
04:41	04:44	* { I: Kodi Zanaco mumaikhulupilila motani? } * [I: How much you trust them?]
04:46	04:50	** { R: Chotiikhulupilila chifukwa asunga ndalama. } ** [R: We trust them because they keep our money.]
04:51	04:56	* { I: So muwakhupilila motani? Koteratu, kwambiri, pang'ono kapena simuwakhulupilila? } * [I: So how much do you trust them? Completely, much, a bit or not at all?]
04:57	05:00	* { R: Tikhulupilila kwambiri. } * [R: We trust them much.]
05:01	05:02	* { I: Kwambiri? } * [I: Much?]
05:03	05:03	* { R: Ee. } * [R: Yes.]

Table A17. Operationalisation of question characteristics

Q. Id.	Characteristic	English	Bemba	Chewa
K27	Wording	"I feel like what happens in my life is determined by others.	Ngufwa kwati abantu bambi e bapima ifili no kucitika mu bumi bwandi ne mikalile.	Ndiwona monga kuti zimene zimacitika pa umoyo wanga zimacitidwa ndi anthu ena.
	Question type	Attitudinal		
	Response option format	Ordinal closed		
	Number of response options read out	5		
	Sequence no.	24		
	Word count	12	14	12
D05	Wording	"In the current cycle until today, how much have you saved individually with the savings group?"	Bushe pa ndalama mwaletako ukufika na buno bushiku, ni shinga isho mwasungapo mu kabungwe kenu akasunga no kukongwesha indalama?	Kodi inu mwakwanisa kusunga ndalama zigati mubungwe lanu pozafika pano?

	Question type	Behavioural		
	Response option format	Open numeric		
	Number of response options read out	0		
	Sequence no.	47		
	Word count	16	19	10
D09b	Wording	When did you receive the last share out?	Bushe ni lilali mwalekelesheko ukwakana indalama shonse nangu ukupasa mu kabungwe kenu akasunga no kukongwesha indalama?	Kodi nthawi yothela kugawana ndalama inali liti?
	Question type	Behavioural	Bushe mwalicetekela abena mupalamo?	Kodi mumawakulupilila kufikila pati anansi anu?
	Response option format	Open numeric (date)		

	Number of response options read out	0		
	Sequence no.	66		
	Word count	8	16	7
X02	Wording	“To what degree do you trust your neighbours?”	Bushe mwalicetekela abena mupalamo?	Kodi mumakulupila kufikila pati anansi anu?
	Question type	Attitudinal		
	Response option format	Ordinal closed		
	Number of response options read out	6 (including DK-option which was supposed to be read out)		
	Sequence no.	66		
	Word count	8	4	6
X23	Wording	“What happens in my life is determined by others.”	Abantu bambi e bapima ifikancitila mu bumi bwandi ne mikalile.	Zimene zimacitika pa umoyo wanga zimacitidwa ndi anthu ena.
	Question type	attitudinal		
	Response option format	Ordinal closed		

	Number of response options	7 (including DK-option which was supposed to be read out)		
	Sequence no.	69		
	Word count	9	10	9
X06	Wording	“To what degree do you trust your neighbours?”	Bushe mwalicetekela abena mupalamo?	Kodi mumakulupila kufikila pati anansi anu?
	Question type	Attitudinal		
	Response option format	Ordinal closed		
	Number of response options	11 but only 2 response options in the extremes labelled and read out		
	Sequence no.	66		
	Word count	8	4	6
X26	Wording	“I would like to ask to what extent what happens in your life is determined by others.”	Ndefwaya mipushe ifyo abantu bambi bapima ifimicitila mu bumi bwenu ne mikalile.	Ndifuna kudziwa kuti nanga zinthu zimene zimacitika pa umoyo wanu zimacitidwa ndi anthu ena pamulingo wotani.
	Question type	Attitudinal		
	Response option format	Ordinal closed		

	Number of response options	5		
	Sequence no.	69		
	Word count	17	12	16
L05	Wording	Imagine a scheme which requires you to pay a regular amount for the risk of an event happening (e.g., harvest or business failure, illness, etc.). Others who also subscribed to this scheme pay as well in this common fund. In case an agreed bad event like crop loss or illness occurs, you will receive the agreed amount of money out of the common scheme. In case there is no bad	Tontonkanyeni ukwa kubikisha indalama sha kubomfya nga kwacitika ubusanso (pamo nga ifya kulya mwalimine fyaonaika nangu ubukwebo bwawa, kwaba ukulwala, na fimbipo). Bambi nabo balabikishako indalama. Nga ca kutila ifintu fimo fyacitika ica kuti ifya kulya ifyo mwalimine fyaonaika nangu kwaba ukulwala, kuti mwapoka indalama kuntu mubikisha pamo ukulingana ne	Yelekezani za makonzedwe akuti muzilipila ndalama zimene zidzathandiza ngati pangacitike zinazake (monga kusakolola bwino kapena malonda kugwa, kudwala. Ena amene analembesa nawonso angazilipila mu tumba la ndalama limodzi-modzi. Ngati zimene munalembesa zacitika monga kusakolola bwino kapena kudwala, muzalandila ndalama zimene munavomelezana

		event, there is no pay-out. "Would you be willing to participate in this scheme?"	fyo mwapangene. Nomba nga takucitike ubusanso, ninshi teti kube ukupoka indalama. Bushe kuti mwatemwa ukubomfya iyi nshila iya kubikishamo indalama?	kucokela mu thumba limeneli. Koma bwanji ngati zinthu zimene munalembelana sizinacitike, ndiye kuti simuzapatsidwa ndalama. Kodi mungatengeko mbali mumakonzedwe amenewa?
	Question type	Behavioural (hypothetical)		
	Response option format	Yes/no		
	Number of response options	2		
	Sequence no.	73		
	Word count	84	70	61