Essays in Development and Health Economics

Dissertation

Submitted to acquire the doctoral degree

from the Faculty of Business and Economics,

at the Georg-August-Universität Göttingen

Submitted by

Lisa Bogler

Göttingen, March 2023

Thesis Committee

First academic advisor:	Prof. Dr. Sebastian Vollmer
Second academic advisor:	Prof. Dr. Janina Steinert
Third academic advisor:	Prof. Dr. Krisztina Kis-Katos

Acknowledgments

I would like to thank several people who supported me in the process of my PhD. To start with, I owe particular thanks to my supervisor Sebastian Vollmer who encouraged me to get a PhD in the first place, who continued supporting me through challenging and failing projects, and who showed faith in me by suggesting new projects to me again and again. I also thank Janina Steinert and Krisztina Kis-Katos for agreeing to be part of my thesis committee.

Ann-Charline Weber, a coincidence following unfortunate events led to us working together and I could not be more grateful for this. I learned immensely from you, and you have been the best friend imaginable at work and beyond. We have gone through every state of mind together and my PhD, input and output, would not have been what it is without you.

I want to thank my colleagues at the chair and CeMIS, especially my pre-pandemic office mates, Anna Reuter and Lisa Rogge. I deeply enjoyed the atmosphere of collegiality and benefited greatly from the hints and tips you so willingly shared.

Thank you to my parents and siblings for the support and care you have given me, from the beginning to today and in what may come next. Thank you to Fiston for doing the best you can throughout all struggles at work and at home.

Finally, I would like to say thanks to those who contributed to the realization of specific essays. Data collection is not possible without many hands and minds working together. I am grateful for everyone who helped along the process.

Essays 1 and 2: A big thank you goes to Abhijeet Kumar, who was a reliable and trusted partner in our projects in Madhepura. The implementation of our study in Anganwadi Centres would not have been possible without him. I would also like to thank all data collectors of the observational surveys (Avinash Kumar, Himanshu Raj, Madan Kumar, Prahalad Kumar, Pushkar Anand, Rahul Kumar, Rahul Kumar, Reetesh Kumar Vimal, Sanjiv Kumar, Sarfraz Nayer, Shabaz Khan, Vivekanand Kumar) and phone survey (Anjali Kumari, Anisha Sah, Ashwini Joshi, Rishika Raj, Shivani Kumari, Sulekha Kumari, Tanu Priya, Tejashwini Joshi) for their precise and reliable work during the data collection, all workshop facilitators (Anjali Kumari, Anshu Kumari, Parvati Kumari, Rajni Sharma, Rinku Kumari, Rishika Raj Samar, Shivani Kumari, Shweta Bharti, Sulekha Kumari, Tanu Priya) for their willingness to engage with us during their own training and deliver the workshops to the AWWs, and Harshpreet Kaur for excellent research assistance. I would like to thank the ICDS Directorate in Patna, the District Programme Officer in Madhepura, and Toni for their support, the CDPOs, data entry officers, and LS in Madhepura district for their cooperation, and all AWWs and AWHs in Madhepura district for their participation in this research. I am very grateful to Alastair

Summerlee for the price deal on our large order of Lucky Iron Leaves and for his support in determining the number of Lucky Iron Leaves and boiling time required to achieve the correct dose of iron in this setting.

Essay 3: Thank you to Kehinde Elijah Owolabi for being our partner in Sokoto, in hours of laminating and more. Thank you to Aminu, not only for your interpretations but also for your kindness and care. Thank you to Joel Ade and thank you to Hamza Kabir and Olujonrin Segun from NOI Polls and all supervisors and enumerators for their hard work during data collection.

Essay 4: I am very grateful to Graeme Blair for his support. When I approached him with a request for help after stumbling across inconsistencies in the results, he adjusted the *predict* function in the *List* package to address my specific need.

Abstract

This dissertation consists of four essays, which are all based on primary data. In the first essay, we describe and attempt to explain the state and functionality of pre-schools (Anganwadi Centres) in Bihar, India. We find an overall very low level of functionality and considerable heterogeneity. The second essay is based in the same setting as the first. We introduced a new method of iron supplementation to a sub-set of functional Anganwadi Centres and evaluated this intervention with a randomized controlled trial with two treatment arms. We measure the success of the intervention with several outcomes proxying knowledge and use of the new iron supplementation method. We find that even fourteen months after implementation, a large share of workers in Anganwadi Centres remembered the method and could explain how it is used. However, indication of actual usage of the method was very low. In the third essay, we use list experiments to measure sensitivity bias in questions on health behaviour and health among an adult population in Dar es Salaam, Tanzania. Additionally, we analyse heterogeneities in sensitivity bias across demographic characteristics of respondents. We find that sensitivity bias is present in some of the outcomes, but not all. The bias also varies across subgroups, especially between men and women. In the last essay, we measure the skill gap between formal and informal mathematics among primary school pupils in Sokoto, Nigeria. We find that a considerable share of children is not able to solve tasks of addition and subtraction when presented in a standard formal way, but able to solve similar and even more complex tasks when presented in an informal way, as a market transaction. This skill gap is partly explained by children engaging in market activities.

Kurzzusammenfassung

Die Doktorarbeit besteht aus vier Aufsätzen, die alle auf Primärdaten beruhen. Im ersten Aufsatz beschreiben wir die Funktionalität von Vorschulen (Anganwadi Centres) in Bihar, Indien, und versuchen diese zu erklären. Wir zeigen, dass die Funktionalität innerhalb der Region stark variiert, aber insgesamt sehr niedrig ist. Der zweite Aufsatz ist in derselben Studienregion situiert. In Anganwadi Centres, die ein minimales Kriterium an Funktionalität erfüllten, führten wir eine Methode zur Anreicherung von Trinkwasser mit Eisen ein. Diese Intervention wurde mit einer randomisierten kontrollierten Studie evaluiert und ihr Erfolg anhand von Indikatoren für das Wissen über die Methode und ihre Anwendung gemessen. Wir zeigen, dass ein großer Anteil der Arbeiterinnen in Angwandi Centres sich auch vierzehn Monate nach der Intervention noch an die Methode erinnern und das Vorgehen bei ihrer Anwendung erklären können. Die Indikation der tatsächlichen Anwendung der Methode ist allerdings niedrig. Im dritten Aufsatz verwenden wir List Experimente, um Bias in sensitiven

Fragen zu Gesundheit und Gesundheitsverhalten in Erwachsenen in Dar es Salaam, Tansania, zu messen. Außerdem untersuchen wir Heterogenitäten im Bias zwischen Bevölkerungsgruppen. Wir zeigen, dass der Bias für mehrere sensitive Fragen existiert aber nicht für alle, und dass er besonders zwischen Männern und Frauen variiert. Im letzten Aufsatz messen wir den Unterschied in Kenntnissen der formalen und der informellen Mathematik in Schulkindern in Sokoto, Nigeria. Wir zeigen, dass ein sehr großer Teil der Kinder Additions- und Subtraktionsaufgaben nicht lösen kann, wenn sie standardmäßig als formale Mathematik präsentiert werden, aber vergleichbare und durchaus komplexere Aufgaben lösen können, wenn diese informell als Marktspiel präsentiert werden. Diese Diskrepanz in den Mathematikkenntnissen kann zu einem Teil dadurch erklärt werden, dass die Kinder in ihrem Alltag im Markt tätig sind.

Contents

Index of figures in main text

Index of tables in main text

Index of figures in appendices

Index of tables in appendices

Chapter 1

Introduction

1.1 Background and Motivation

Self-reported survey data are a key data source for empirical social science, including development economics. Unfortunately, such data may be inaccurate for multiple reasons (Weisberg, 2009). Inaccurate data, in turn, can influence analyses and conclusions drawn from survey data, possibly leading to ineffective or counter-productive policies (Bruckmeier et al., 2021; Meyer & Mittag, 2019). It is therefore important to understand sources of inaccuracies and apply methods to mitigate or avoid them.

One source of inaccuracy is measurement error, meaning that the measure does not capture accurately what was supposed to be measured (Weisberg, 2009). Measurement error can originate from the survey respondents. To start with, survey data may be inaccurate because the respondents do not know how to answer accurately. Demographic Health Surveys and Multiple Indicator Cluster Surveys, for example, measure the prevalence of acute respiratory infections in children as a proxy for pneumonia by asking mothers about their children's symptoms such as cough and difficulty breathing. However, mothers' responses may be too un-specific for identifying the true prevalence of acute respiratory infections or pneumonia (Campbell et al., 2013; Harrison et al., 1995; Hazir et al., 2013). Literacy may also play a role (Lupu & Michelitch, 2018). About half of the respondents in previous Afrobarometer surveys, where average literacy is relatively low, had difficulties answering questions, even more than half in the poorest countries such as Guinea and Mali (Lupu & Michelitch, 2018). Recall bias is another reason for measurement error in self-reported data and occurs when questions relate to previous time periods that are recalled less accurately than the present. Longer recall periods lead to larger errors. This has been shown for questions on food consumption (Zezza et al., 2017), income and assets (de Nicola & Giné, 2014), agricultural variables (Wollburg et al., 2021), as well as vaccination status (Czaja et al., 2005; Liu et al., 2017; Miles et al., 2013; Ramakrishnan et al., 1999) and health care visits (Brusco & Watts, 2015).

While the lack of understanding of the question or answer options and recall bias lead to unintentional inaccuracy in reporting, survey data may also be affected by intentional misreporting by respondents. If respondents expect to benefit or avoid negative consequences from giving specific responses, they may misreport. For example, in a poverty alleviation programme, underreporting of goods and desirable home characteristics has been shown to be widespread, presumably because respondents expect to benefit from the programme if they report to be poorer than they actually are (Martinelli & Parker, 2009). Underreporting of income has been identified in various settings (Brewer et al., 2017; Carletto et al., 2022; Hurst et al., 2014). Respondents may also misreport if the beliefs or behaviours at question are affected by social stigmata. The so-called social desirability bias has been observed in a range of topics and settings, including racism (Gilens et al., 1998), voting (Gonzalez-Ocantos et al., 2012), condom use (LaBrie & Earleywine, 2000), intimate partner violence (Cullen, 2020), and abortion (Moseson et al., 2017).

Furthermore, measurement error can stem from the design of the survey or question. Different designs of a measure can result in different accuracies (Beegle et al., 2012), which may not be known a priori. Especially complex measures, such as labour market indicators, are captured differently across surveys. Depending on the survey instrument used, the outcome measure, such as the employment rate, differs (Desiere & Costa, 2019). To highlight one more survey feature, longer surveys lead to response fatigue. This affects answers to various topics and question types. In a study in Malawi, survey length affected predicted poverty rates (Kilic & Sohnesen, 2019), and in Ethiopia, placing a food consumption module closer to the end of the survey affected the calculated dietary diversity score (Abay et al., 2021).

Measurement error can be systematic or random. Random measurement error makes a measure less precise but does not necessarily bias the measure up- or downwards. It increases the variance of a measure but does not affect its mean. A larger variance leads to lower reliability of the measure, attenuated correlations with other variables, and less statistical power for estimations (Weisberg, 2009). Systematic measurement error, in contrast, creates an up- or downward bias in a measure. This effect on the mean value of the measure reduces its validity (Weisberg, 2009). Both types of measurement error are a concern because they might affect the conclusions drawn from analyses based on survey data. For example, measurement error can lead to an underestimation of the prevalence of child undernutrition, such that resources for research and policies against child undernutrition are shifted to other health concerns, leading to a suboptimal resource allocation. Imprecisely measuring the association between individual characteristics, for example age and undernutrition, due to attenuated correlations, may mean that the population group with these characteristics, i.e. those in a specific age group, are disregarded in a policy response. Low reliability of key measures also affects progress tracking with indicators for the Sustainable Development Goals (Wollburg et al., 2021). Systematic measurement error generally causes bias in causal estimates, for example in the evaluation of an intervention. The effectiveness of a programme may therefore be underestimated, leading to its abolishment, or vice versa, a programme with overestimated impact may be scaled-up despite an actually low impact. For example, anti-poverty programmes in the US seem to be more effective than previously reported due to underreporting among beneficiaries about transfer dollars received (Meyer & Mittag, 2019). Similarly, survey data of take-up behaviour of unemployment benefits in Germany seem to be biased, leading to the inaccurate evaluation of their effectiveness (Bruckmeier et al., 2021).

There are various ways of addressing measurement error in survey data. Firstly, some measures can be physically verified through enumerator observation. In an evaluation of the Mexican program *Oportunidades*, for example, households were visited by administrators to verify self-reported asset information (Martinelli & Parker, 2009). Secondly, qualitative validation techniques can be used to verify data on behaviour. To measure misreporting on self-reported engagement in crime, drug use, gambling, discretionary spending, and homelessness, a study in Monrovia, Liberia, employed intensive qualitative approaches (Blattman et al., 2016). Research staff visited survey respondents for several hours at a time, building trust, engaging in in-depth conversations and observation. However, physical verification and qualitative validation may both be very costly or impractical for some settings and measures. Indirect questioning, a third approach, is increasingly used to reduce social desirability bias in self-reported survey data (Rosenfeld et al., 2016). Techniques include list experiments (Corstange, 2009; Glynn, 2013; Imai, 2011), random response techniques (Blair et al., 2015; Gingerich, 2010), and endorsement experiments (Bullock et al., 2017; Lyall et al., 2013). In general, these techniques require larger samples compared with direct questioning (Corstange, 2009) and come with additional limitations. The unusual question format, for example, is more complex and difficult to understand than direct questions (Kramon & Weghorst, 2019) and can lead to non-compliance with the instructions on how to answer the question (Chuang et al., 2021). Finally, administrative data can be used together with or instead of self-reported information. Information on health indicators, for example, can be collected through self-reports of households or through health service providers. A study in Mali compared three coverage rates, contraceptive prevalence, institutional delivery, and vaccination with the diphtheria, pertussis, and tetanus vaccine, estimated from self-reported survey data and routine health system data (Sawadogo-Lewis et al., 2021). They found that, at least on national level, routine data seemed sufficient to inform program planning and prioritization. However, administrative data, such as vital statistics, may often not be available, especially in low-income settings or crisis situations, or may be incomplete (Mikkelsen et al., 2015).

Measurement error clearly is a concern in survey data. However, there is no panacea to avoid it and each data collection method comes with its own challenges, limitations, and sources of measurement error. In many settings and for many topics, the choices of data sources are limited, and self-reported survey data may still be the best source of information available. Respective data limitations should be reflected on and reported. Assumptions about sources of measurement error, their direction and impact can be incorrect even in presumably well-researched settings (Blattman et al., 2016). In the best-case scenario, the presence and magnitude of measurement error should therefore be tested so that it can be taken into account when conclusions are drawn.

1.2 Chapter Overview

The essays included in this dissertation address the topic of measurement error in different ways. All essays are based on survey data. The first essay uses a combination of observational data and individual phone survey data to capture the functionality of pre-schools. Observational data, a form of physical verification, is used to address the problem of inflating attendance rates in administrative records of pre-schools in this setting (Deshpande, 2019). The second essay uses a sub-sample of the same data, with a focus on the phone survey data, to evaluate a new intervention. Measurement error is addressed by choosing those measures as outcomes that are least affected by intentional misreporting. The third essay is based on individual survey data from list experiments. Answers to the indirect questioning technique are compared with answers to direct questions to measure social desirability bias, one important aspect of inaccuracy in survey data. The fourth and last essay uses individual data of a learning assessment among primary school pupils. The learning assessment is specifically designed to measure skills in numeracy accurately, capturing aspects not considered in standard numeracy tests.

The essays are placed in three different settings. The first and second essays are based on data collected in pre-schools in rural Madhepura, Bihar, India. The list experiments used in the third essay were conducted with adults older than 40 years in urban Dar es Salaam, Tanzania. The fourth essay is set in primary schools in rural Sokoto, North-West Nigeria. While these settings would all be characterised as low- or middle-income contexts, they vary considerably in terms of social structure, culture, and political environment. The essays also differ methodologically. The first and fourth essay rely on in-depth descriptive analyses. In the second essay, a field experiment is used to evaluate the effectiveness of a new intervention. The third essay analyses list experiments.

Essay 1: Functionality of Anganwadi Centres in Madhepura district, Bihar, India

The first essay describes and attempts to explain the functionality of Anganwadi Centres (AWCs) in Madhepura district, Bihar, India. AWCs are pre-schools providing education and daily meals to children aged three to six years, among other services, and are an integral part of the Integrated Child Development Services (ICDS) scheme in India. Previous evidence indicates a lack of functionality among AWCs (Asha, 2014). Specifically in Bihar, fund leakage is a major concern and AWCs seem to be underperforming in terms of opening hours and meal provision (Fraker et al., 2013). The goal of the study is to evaluate the potential of AWCs in Madhepura for delivering nutrition interventions.

We implemented an observational survey of all AWCs in ten out of thirteen blocks, the administrative regions of Madhepura district. The observational survey captured the presence of caretakers, called Anganwadi workers (AWWs), their helpers (AWH), and children at the AWC, their activities at the time of the visit, and the physical structure of the building. We chose to implement an observational survey because previous visits to AWCs and interactions with AWWs as well as evidence from other studies indicated a severe problem of misreporting on child attendance and activities (Deshpande, 2019; Fraker et al., 2013). The observational survey data was subjected to strict quality control measures. Enumerators took pictures of the AWCs which we used to validate survey responses daily. This observational data was combined with data from a phone survey with AWWs to identify factors explaining functionality of the AWCs. As functionality is a complex concept, we used five different measures to capture its multiple aspects: AWC being open, AWW being present, 20 or more children being present, learning ongoing, indication of food being served. All five measures were directly observable, removing the possibility of misreporting by AWWs. We used simple linear regressions to identify associations between characteristics of the AWW and the AWC and our measures of functionality.

Overall, we found a low level of functionality among AWCs and considerable heterogeneity across blocks in Madhepura district. The potential of AWCs in Madhepura for delivering nutrition interventions was therefore limited at this stage. The variables captured by our surveys were not able to explain the variation in functionality. This may indicate that functionality is driven by the intrinsic motivation of individual caretakers, which was not measured in our surveys.

The main contribution of this essay is to capture the functionality of all AWCs in ten blocks of Madhepura using objective, observable measures, thereby reducing the possibility of measurement error through misreporting of AWWs. This objective data provides valuable information for the ICDS, which usually relies on often tampered monthly progress reports completed by AWWs and their superiors (Deshpande, 2019; Fraker et al., 2013). Since the survey, AWCs were closed for over a year due to the COVID-19 pandemic. Upon the re-opening of AWCs, the ICDS set out to implement a range of new initiatives aimed to strengthen the infrastructure of AWCs and with a focus on battling anaemia. Our survey provides a valuable baseline for comparison to evaluate these new initiatives and to track changes in functionality. The method used for the survey can easily be replicated.

Essay 2: Introducing a new method of iron supplementation into Anganwadi Centres in rural Bihar, India: a randomized controlled trial

The second essay is based on the same data as the first. We used a sub-sample of AWCs fulfilling a minimum level of functionality to evaluate the effectiveness of a new intervention. The goal of the intervention was to introduce a new method of iron supplementation to AWCs to combat the high prevalence of anaemia among children in this setting (Bogler et al., 2021; International Institute for Population Sciences (IIPS), 2021). The effectiveness of the intervention was causally identified through its design as a randomized controlled trial with two treatment arms. AWWs in the light treatment arm were invited to a workshop where a new method of iron supplementation was introduced, the Lucky Iron Leaf. AWWs in the intensive treatment arm were invited to a longer workshop that introduced the same method and additionally contained elements of empowerment. They received a regular delivery of ingredients for the daily use of the new method and a phone call to remind them of the method and to clarify potential questions. AWWs in the control arm did not receive anything. While the iron supplementation method had been shown to be successful in increasing haemoglobin levels to reduce anaemia rates in Cambodia (Charles et al., 2011, 2015), it was not known whether this method is feasible for implementation in AWCs.

We originally planned to evaluate the intervention by measuring haemoglobin levels of children attending AWCs as an objective measure of anaemia. Due to the COVID-19 pandemic and following AWC closure, collecting these data was not possible. We therefore had to adjust and conducted a phone survey with AWWs instead. As outcome measures for the evaluation, we chose indicators for the knowledge and usage of the method. We specifically chose such measures that presumably were least affected by intentional misreporting. First, we asked which methods of iron supplementation the AWW knows. Mentioning the Lucky Iron Leaf as a response to this question, without being prompted about it, was chosen as a first indicator that the AWW remembers the workshop. We further asked how the Lucky Iron Leaf is used. Mentioning key steps of its use was chosen as an indicator that the AWW remembers the content of the workshop. Lastly, we used data from the observational survey that was conducted directly after the implementation of workshops. We chose as indicator of potential usage the observation of the Lucky Iron Leaf at the AWC.

We found that even fourteen months after implementation of the intervention, a large share of AWWs remembered the new method and could explain how it is used, especially among those in the intensive treatment arm. However, observed usage of the new method was very low. This indicates that substantial barriers stopped the AWWs from implementing a method which they understood and reportedly found important.

The key contribution of this essay is the assessment of a potentially powerful distribution channel for iron supplementation as the AWCs reach the most vulnerable population group even in remote areas of India. While it was not possible to use haemoglobin levels, a biomarker and arguably objective outcome, to evaluate the intervention, we chose those outcomes presumably least prone to misreporting. Doing so avoided intentional misreporting as source of measurement error. We showed that, using our interactive workshops, teaching AWWs about the new method was possible and relevant information was retained. For ensuring that the new method is embedded in the daily routine in AWCs and health impacts could be realised, institutional support and an improvement of overall functionality of AWCs would be necessary.

Essay 3: Sensitivity bias in questions on sexual behaviour, sexual health, drug consumption, and intimate partner violence in an adult population in urban Tanzania

In the third essay, we use list experiments to measure sensitivity bias, also called social desirability bias, in questions on health behaviour and health among an adult population in Dar es Salaam, Tanzania. List experiments are an indirect questioning technique that have been used to address social desirability bias in self-reported surveys on a range of topics, including racism (Gilens et al., 1998), voting (Gonzalez-Ocantos et al., 2012), condom use (LaBrie & Earleywine, 2000), intimate partner violence (Cullen, 2020), and abortion (Moseson et al., 2017).

In a list experiment, survey respondents are split into a treatment and a control group. The control group receives a list of *J* statements, the control statements, while the treatment group receives the same list of *J* statements and an additional *J+1th* statement about a sensitive topic. Respondents are asked to report how many of the statements are true for them. As this method makes it impossible to infer the individual response to the sensitive item, an increased sense of privacy is created. Respondents are therefore more likely to answer the sensitive item, e.g., whether or not they engage in a risky behaviour, truthfully.

While this technique is useful to reduce sensitivity bias from a prevalence estimate, it is also informative to measure the magnitude of sensitivity bias itself in different outcomes. This is possible by comparing answers given in list experiments with answers given to direct questions on the same topic. By doing this we explicitly measure the presence and magnitude of sensitivity bias in this setting. In addition, we analyse how sensitivity bias varies with demographic characteristics of respondents. As list experiments underly a set of assumptions (Imai, 2011), we also intensively test for these to assess the validity of the list experiments.

Our findings indicate the presence of sensitivity bias in the responses on some health behaviour and health outcomes but not all. The findings for alcohol consumption and transactional sex are different from common assumptions. We do not find any sensitivity bias in the question on having consumed any alcohol in the past month. We do find sensitivity bias in the question on having engaged in transactional sex, but this activity is overreported. The analysis of heterogeneity by demographic characteristics shows differences in the presence and magnitude of sensitivity bias between men and women. For example, while men show a high sensitivity bias in the question on physical intimate partner violence, this is not the case for women.

The contribution of this essay is the estimation of sensitivity bias, one important source of measurement error on sensitive topics, on a set of ten outcomes related to health in the context of urban Tanzania. Our findings show that list experiments are valuable in at least two ways. They can reduce sensitivity bias for sensitive questions that are affected by it. In addition, in combination with direct questions, they can help to measure the size of the measurement error due to intentional misreporting and correct common assumptions.

Essay 4: Formal vs. informal mathematics: Assessing numeracy with school and market items in 5,997 school children in North-West Nigeria

In the fourth essay, we measure the skill gap between formal and informal mathematics among primary school pupils in Sokoto, Nigeria. During the piloting of a learning assessment for a school programme evaluation, we observed the pupils' low performance in numeracy exercises. At the same time, school-aged children could be seen engaging in market transactions, buying and selling goods on their own. This suggested a gap between numeracy skills as measured by formal assessments in school and numeracy skills existing in the daily life of the children. The standard learning assessment did not capture the latter skills. We therefore designed a market simulation game as one part of the learning assessment to measure this aspect of numeracy skills among the pupils.

We define the skill gap as the share of pupils who were able to solve tasks of the market simulation game but were not able to solve similar formal mathematics tasks. We describe the magnitude of the skill gap and attempt to explain it with the pupils' engagement in market activities using simple linear regressions. In addition, we explore whether differences in the design of the formal tasks and the market simulation game can explain the skill gap.

We find that the skill gap is considerable and exists across all schools in the sample, ages, grades, and gender. It is robust to alternative definitions. Engagement in market activities seems to explain part of the skill gap. The association between the skill gap and our measure of market engagement is also robust to alternative specifications and sample restrictions to address concerns with data quality. In contrast, design effects seem insufficient to explain the skill gap.

This essay is the first study to measure the skill gap between formal and informal mathematics in a large sample of pupils in a school setting. We developed a new learning assessment that captures two separate aspects of numeracy skills as a more accurate measure compared to standardized tests. This extended learning assessment reduces the measurement error arising due to the incompleteness of standard tools assessing numeracy skills. This learning assessment, if validated in other settings, could help to improve the measurement of numeracy skills that are relevant for daily life and employment, outside the performance in standardized, formal school tests.

1.3 Summary and Conclusion

The four essays all focus on different topics and use different methodologies. Nevertheless, being based on survey data, measurement error is a concern in each. We use different ways to address, reduce, and analyse the measurement errors we identify or assume to be present in our data. We choose specific outcome and explanatory variables to avoid sources of measurement error, we explicitly measure sensitivity bias as one aspect of measurement error, and we design a measurement tool to create a more accurate measure than standard tools capture. We acknowledge that the possibility of measurement error is not removed entirely from any of the analyses as there are multiple sources of measurement error that may be unknown a priori. Nevertheless, these essays present examples of how measurement error can be reflected on and addressed in studies using survey data from different contexts.

Chapter 2

Functionality of Anganwadi Centres in Madhepura district, Bihar, India

with: Ann-Charline Weber, Abhijeet Kumar, and Sebastian Vollmer

Abstract

Anganwadi Centres (AWCs) are an integral part of the Indian Integrated Child Development Services scheme. They are providing supplementary nutrition, pre-school non-formal education, as well as nutrition and health education across India. Reaching even remote areas and vulnerable populations, AWCs are a potentially powerful platform for combating malnutrition. However, there is evidence of low functionality. We use data collected through observational visits and a phone survey with Anganwadi Workers (AWWs) to describe the functionality of AWCs in Madhepura district in terms of services provided on site between November 2019 and February 2020. Functionality is measured with five observable indicators. In addition, we identify factors associated with functionality. We find that functionality of AWCs in Madhepura is overall low but very heterogeneous. We conclude that AWCs in this state are not able to effectively deliver intended services. Structural reforms reducing the workload of AWWs and introducing regular renumeration would be crucial steps to improve service delivery through AWCs.

2.1 Introduction

The Integrated Child Development Services (ICDS) scheme in India was launched in 1975 to address key aspects of child development, including nutrition, education, and health. It has since become one of the largest early childhood development schemes in terms of beneficiaries globally (Fraker et al., 2013). Anganwadi Centres (AWCs) are an integral part of the ICDS. They are responsible for implementing the supplementary nutrition programme (SNP), as well as providing pre-school non-formal education, and nutrition and health education. The targeted beneficiaries of services delivered through AWCs include children below six years, pregnant and lactating women, and adolescent girls.

AWCs are widely distributed across India, reach even remote areas, and specifically target vulnerable populations, making them a potentially powerful platform for the delivery of specific nutrition interventions. However, there is evidence of low functionality even in terms of routine services, poor infrastructure, and reports of fund leakage (Asha, 2014; Dasgupta et al., 2012; Fraker et al., 2013; Maity, 2016). It is therefore questionable how effective AWCs can be in combating malnutrition and delivering additional programs.

Improving the functionality of AWCs would clearly be beneficial for the intended beneficiaries. In a descriptive study, Fraker et al. show that better service provision is positively associated with the nutritional status of children (Fraker et al., 2013). Using a causal identification, another study finds that children benefiting from cooked meals at the AWCs have better nutritional status than those who do not, meaning that there is no substitution away from meals cooked at home (Mittal & Meenakshi, 2019). A comprehensive overview of the state of AWCs, understanding potential problems and factors contributing to deficiencies in functionality are important first steps towards improvement of service provision and ultimately nutritional status of beneficiaries.

We contribute to this understanding. In preparation of an iron supplementation intervention in AWCs in Madhepura district, Bihar, we assess the functionality of AWCs in this district. We specifically focus on those AWC services that are provided on site on six days a week, namely pre-school education and daily meal preparation for children aged three to six years. We captured the state of AWCs in Madhepura district before their closure due to the COVID-19 pandemic, between November 2019 and February 2020. In addition to describing the functionality of these AWCs, we identify factors associated with their functionality. We use data collected through two rounds of unannounced visits to AWCs and a phone survey with Anganwadi Workers (AWWs) of all AWCs in ten blocks of Madhepura district.

Other studies have explored the functionality of AWCs in different Indian states before. One study measured the efficiency of AWCs in service provision in a sample of 200 AWCs in one district of Kerala (Asha, 2014). While food supply was reported to be fully functional in this setting, infrastructure was inadequate in about 74 percent of AWCs. Ethnographic work in Maharashtra in 2014 to 2015 found that less food and food of lower quality than determined by the ICDS menu was provided to beneficiaries and that reported attendance of children was inflated while a system of informal payments between ICDS functionaries was sustained (Deshpande, 2019). Similarly, a quantitative assessment of SNP, conducted 2013 in Bihar, reported that meals were not served every day when the AWC was open, that meals were reduced in terms of calories and proteins, and that fewer than the scheduled number of children were attending (Fraker et al., 2013). These factors resulted in 53 percent of the budget going missing due to leakage (Fraker et al., 2013). Take-home rations, provided to beneficiary households monthly, were similarly affected, with beneficiaries receiving less than the intended quantity and some intended beneficiaries not receiving their rations. All this was linked to low levels of nutrition (Fraker et al., 2013).

Our study is different from previous studies as it covers the entirety of AWCs in a specific region instead of a selected sample and combines observations of enumerators survey data with self-reported data. To measure various aspects of functionality, we use objective, observable indicators instead of constructed indices.

2.2 Setting

This study is set in Madhepura district, Bihar. The district is mostly rural and most key development indicators lag behind the national average (International Institute for Population Sciences (IIPS), 2021; International Institute for Population Sciences (IIPS) and ICF, 2021a, 2021b). Most recent data from the National Family Health Surveys 5 in 2019-2020 (NFHS-5) report low nutrition levels among children under 5 years. According to this data, 46.3 percent of children under 5 years are stunted (short for their age) compared to a national average of 36 percent, 20.6 percent are wasted (thin for height), and 41.0 percent are underweight (thin for age). The share of children under 5 years with anaemia is 67.7 percent. While rates of undernutrition reduced since the previous round of NFHS in 2015-2016, the improvement has been minimal (International Institute for Population Sciences (IIPS), 2021). The share of anaemic children even increased since 2015 (International Institute for Population Sciences (IIPS), 2021).

The ICDS builds on the structure of the administrative regions. At the block level, the highest ICDS functionary is the Child Development Project Officer (CDPO) who oversees all AWCs in the respective block. Lady Supervisors, the next-lower functionaries, oversee a group of usually 25 AWWs. Each AWW is in charge of one AWC providing, among other services, pre-school education and daily meal for up to 40 children and supported by one Anganwadi Helper (AWH). Mini AWCs can register only up to 20 children for on-site services and do not have any AWHs. Madhepura district consists of thirteen administrative blocks. At the time of the survey, one of the CDPOs in Madhepura district was in charge of multiple blocks. Lady Supervisors often oversaw larger groups of up to 50 AWWs. Seven AWWs in our sample were in charge of two AWCs as they were additionally substituting in an AWC where no AWW was assigned at that time. In Madhepura district, the daily meal provided to attending children is prepared by the AWH and/or AWW either directly at the AWC or at the home of the AWW. According to monthly progress reports for November 2019, 91.5 percent of AWCs that reportedly had any children registered, had 40 children registered officially. Among Mini AWCs, 82.4 percent of those with any children had 20 or more children registered.

2.3 Data

Data on functionality was collected in observational surveys. Enumerators made two unannounced visits to AWCs and recorded the presence of AWW, AWH, children, and any other adult, and their respective current activity. AWW and AWH were usually recognizable through their uniforms. The first observational survey took place between November 20 to December 16, 2019 and captured all AWCs in ten out of thirteen blocks of Madhepura district. AWCs were observed Mondays to Saturdays between 10:15 am and 1:45 pm as operating times were 10 am to 2 pm and this allowed a margin of 15 minutes for coming late or closing early. In addition to ongoing activities, the survey recorded physical characteristics of the AWC building. The reported characteristics were cross-checked daily with pictures taken of each AWC and field reports, corrections were implemented after clarifying with enumerators. In addition, the geo-location of all AWCs was recorded. The second observational survey captured the subset of AWCs that were defined as running at the first visit. These are centres in which five or more children together with the AWW, AWH or another involved person, or at least ten children were observed. The survey took place between January 9 and February 4, 2020. At the time of this survey, opening hours were shortened due to winter cold. Observations of AWCs were therefore made Mondays to Saturdays, between 12:00 pm and 2:00 pm. All enumerators of the observational surveys were male because they had to travel long distances on motorbike, which was not considered safe for female enumerators.

All AWCs were closed in March 2020 due to the COVID-19 pandemic. The planned face-to-face survey with AWWs was therefore replaced by a phone survey with all AWWs and Lady Supervisors of the same ten blocks in Madhepura district. The phone survey took place between March 22 to May 31, 2021. AWWs were asked about personal demographic characteristics, their work as AWW, and their perception of the AWC system. Supervisors received very similar questions. All enumerators of the phone survey were female because we expected AWWs to be more comfortable with female interviewers.

All enumerators reported their observations from the surveys daily. This way, we collected anecdotal information that was not captured directly in the quantitative surveys.

2.3.1 Ethical considerations

The study design was reviewed by the ethics committee of the University of Göttingen, Germany, with no objections. We received approval for the data collection from the ICDS Directorate in Patna, Bihar, responsible for all AWCs in Bihar, as well as the District Programme Officer in Madhepura district. We took informed consent from each AWW and supervisor personally for the phone survey. Enumerators of observational visits and the phone survey were blinded to treatment assignment. The study was registered at the AEA RCT Registry on February 11, 2020 under AEARCTR-0005449 (https://www.socialscienceregistry.org/trials/5449).

2.4 Description of state of AWCs and AWW working conditions

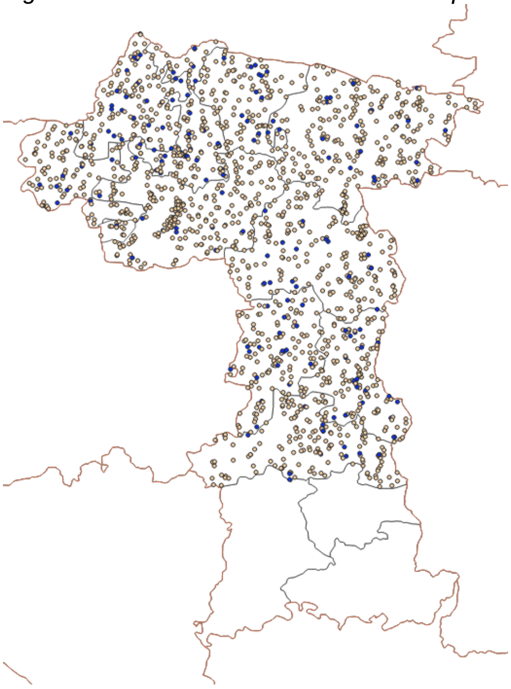
2.4.1 State of AWCs

In the first observational visit, 1719 AWCs were surveyed. Of those, 819 were observed as running and re-visited in the second observational survey. The distribution of AWCs across the ten blocks of Madhepura can be seen in Figure 2.1 and Table A 2.1. We find on average substantial deficiencies and large variations in the characteristics of buildings as well as attendance and provision of services.

Characteristics of AWC buildings

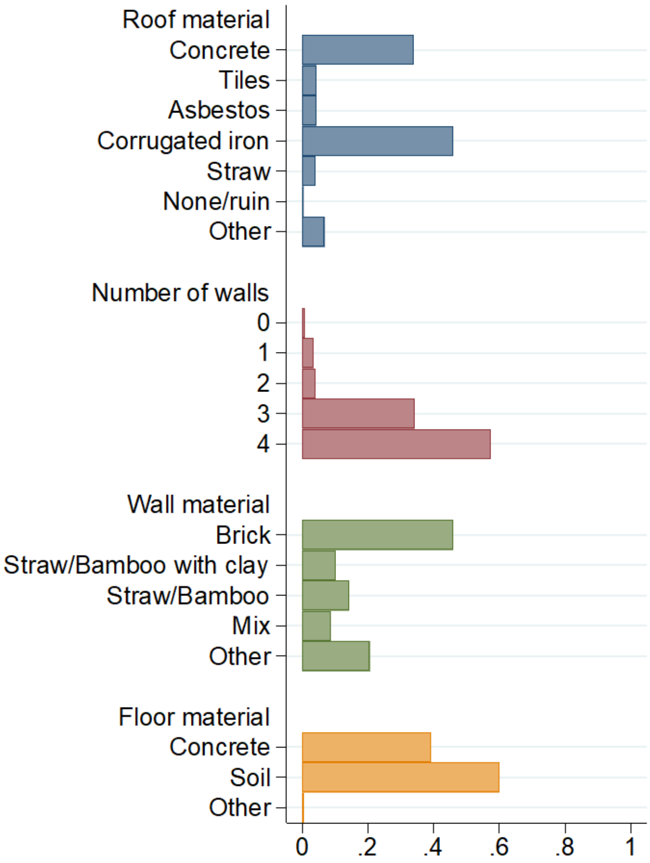
The characteristics of the AWC building do not necessarily determine service provision by the AWC, but they are likely relevant in multiple ways. Stronger, more robust buildings provide better protection from weather, i.e. rain, heat, or cold, to children attending the AWC. More children might therefore attend the AWC if they are more comfortable there. More robust buildings that can be locked also provide better protection for food and equipment being stored there for the daily meal or teaching. AWCs with sufficient space to play outside and inside might make it easier for AWWs to engage the children in structured playing. It is therefore relevant to take stock of available infrastructure.

Figure 2.1: Distribution of AWCs in Madhepura



Note: AWCs in Madhepura district, blue: Mini AWCs, orange: regular AWCs.

Figure 2.2: Characteristics of AWC buildings



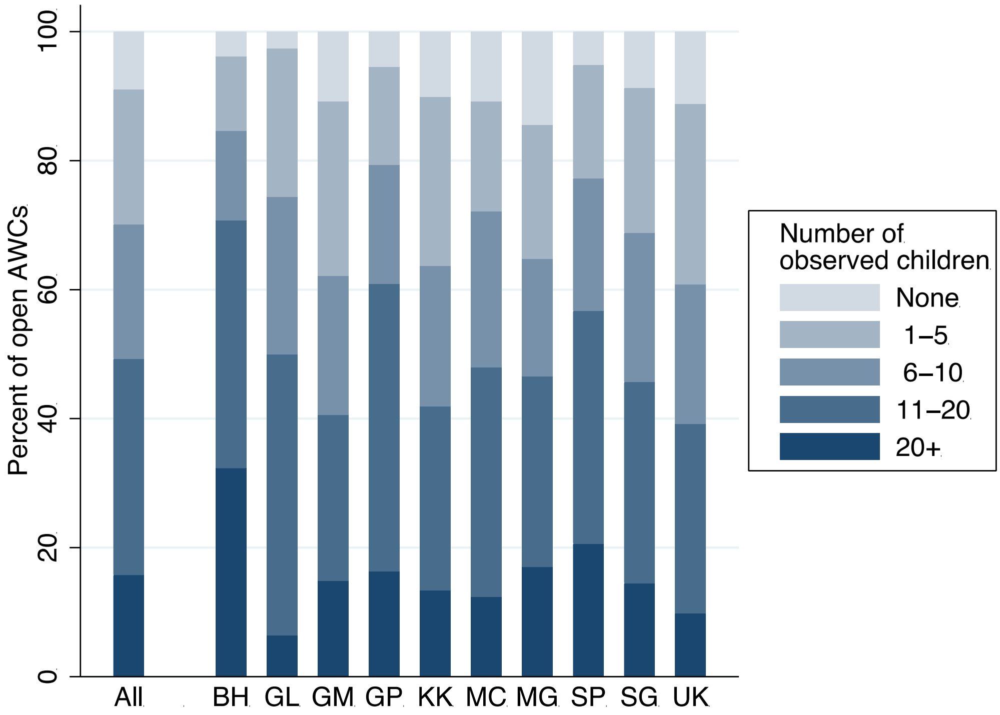
Note: Shares of AWCs with the respective characteristics at the first observational visit, full sample.

The types of buildings in which AWC are hosted varied widely, ranging from full brick and cement buildings to instable constructions with only two walls and a leaking roof or a roof on four poles, offering little protection from rain, heat, or cold. Details are presented in Figure 2.2. Most roofs of AWC buildings were made of corrugated iron (36.0 percent) or concrete (34.0 percent). More than half of AWC buildings had four walls (57.4 percent), but there were also many with only three walls (34.2 percent) or less (4.1 percent with two walls, 3.5 percent with one wall). The material of the walls was mixed, sometimes even within one building. Walls were often made of brick (46.1 percent), followed by straw and/or bamboo (14.4 percent) and straw and/or bamboo covered with clay (10.2 percent). The floor was mainly constructed out of soil (60.1 percent), followed by concrete (39.3 percent). The floor was elevated in 71.7 percent of AWCs, meaning that in 28.3 percent of AWCs, rain could easily flow into the AWC, which is problematic especially when the floor is made of soil. Out of all AWCs observed, 27.6 percent had a roof out of concrete or tiles as well as four walls, all made of either bricks or straw/bamboo covered with clay, constituting a building that provides proper protection to children attending the AWC from weather and other disruptions. Overall, the quality of AWC buildings was better among AWCs that were owned by the ICDS compared to AWCs that were rented. Among the ICDS owned buildings, the shares of buildings with concrete roof, four walls, brick walls, and a concrete floor was considerably higher. Not only the type of construction, also the centres' condition, equipment, facilities, and child-friendliness of the location varied greatly. Enumerators reported neatly kept AWCs, AWCs with educational wall paintings and posters, separate kitchen, storage room, toilet and water pump as well as deteriorating buildings without any facilities, AWCs with a large playground as well as centres located right next to a big road or a pond.

Attendance and activities

In all indicators capturing the attendance and activities of AWW and children, considerable variation across blocks can be observed (see Figure 2.3 and Figure 2.4). 74.7 percent of AWCs were open at the time of the first visit, ranging from 64.9 percent in Kumarkhand to 85.2 percent in Singheshwar. We defined an AWC as open if any person was present at the time of the visit, including children, AWW or AWH. There is similarly large variation for the presence of the AWW. While the total share of AWWs found present at the AWC was 41.4 percent, the lowest shares were 16.8 percent in Gwalpara, 32.4 percent in Murliganj and 32.6 in Kumarkhand, the highest 60.6 and 60.9 percent in Bihariganj and Shankarpur. The very low share in Gwalpara might capture that some survey days in Gwalpara coincided with unannounced training or meeting days for AWWs in this block.

Figure 2.3: Attendance of children



Note: Shares of open AWCs with specific number of children present at the time of the first observational visit in the full sample, and by block; BH = Bihariganj, GL = Gailadh, GM = Gamharia, GP = Gwalpara, KK = Kumarkhand, MC = Madhepura City, MG = Murliganj, SP = Shankarpur, SG = Singheshwar, UK = Udakishunganj.

The second observational visit was only conducted among the 819 AWCs (47.6 percent of the full sample) that we defined as running during the first visit. At the second visit only 66.2 percent of this subsample were again observed as running. 81.7 percent were classified as open, ranging from 58.0 percent in Madhepura City to 96.7 percent in Gwalpara.

The AWW was present in 50.7 percent of AWCs, slightly more often than in the sample of all AWCs at the first visit (41.4 percent), but less often than in this sub-sample of AWCs at the first visit (65.7 percent). Again, variation was large with AWWs present in 76.2 percent of AWCs in Bihariganj and 20.8 percent in Gailadh. Gwalpara does not stand out with a low share of AWWs present at the second visit (49.2 percent), indicating that a meeting day might have affected the first observation.

The ICDS intends that each AWC caters to 40 children on site, while Mini AWCs cater to up to 20 children. We observed that 20 or more children were present in only 15.7 percent of open AWCs at the time of the visit and 5 or fewer children in 29.9 percent (see Figure 2.3). In 34.9 percent of open Mini AWCs 10 or more children were present compared to 50.9 percent of open regular AWCs and 17.3 percent of open regular AWCs having more than 20. The presence of children varied greatly across blocks. Most AWCs with 20 or more children present were observed in Bihariganj (32.3 percent) and least in Gailadh (6.4 percent). The shares of open AWCs with five or fewer children present were highest in Udakishunganj (39.2 percent), Gamharia (37.8 percent), and Kumarkhand (36.3 percent).

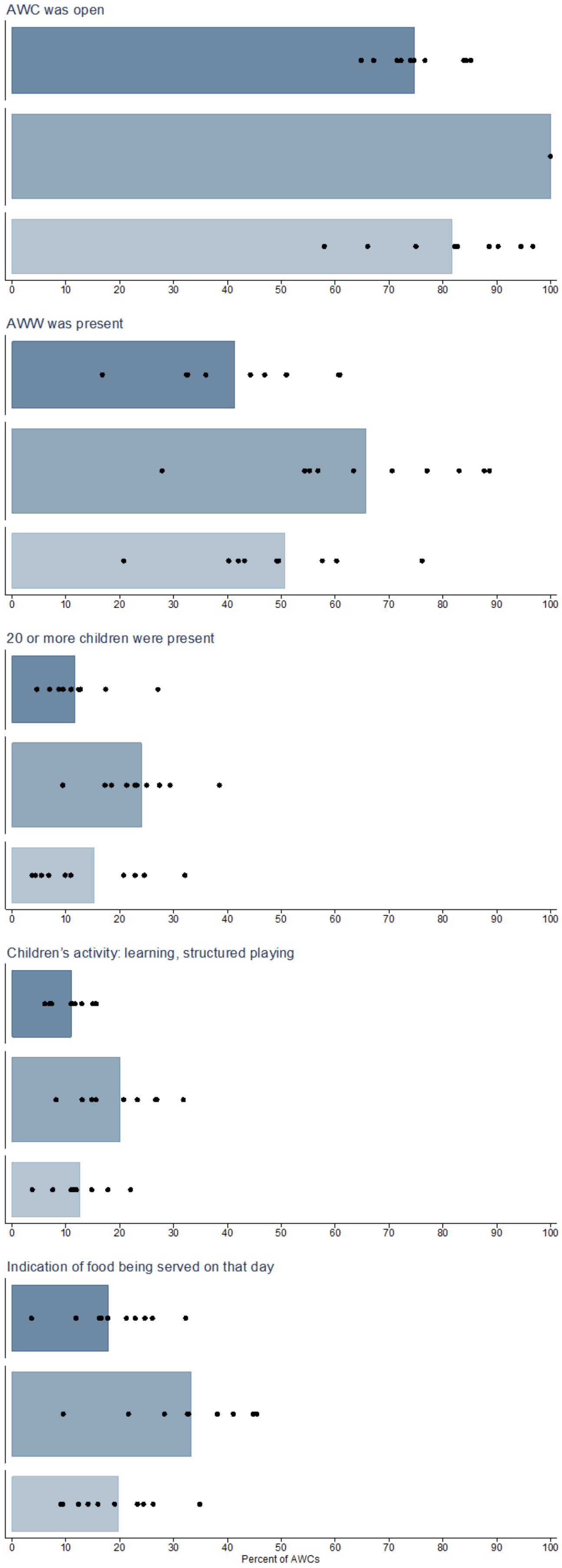
In the subsample of AWCs visited twice the share of AWCs with 20 or more children attending was 24.1 percent at the first visit and 18.7 percent at the second visit, closer to the share in the full sample at the first visit.

According to the ICDS, key daily activities at the AWCs are informal pre-school education and food provision (morning snack and lunch) for attending children. However, these activities were not regularly observed (see Figure 2.4). Learning or structured playing were going on in 11.1 percent of AWCs (6.1 percent in Murliganj, 15.7 percent Shankarpur). Any indication of food served on that day was observed in 14.4 percent of AWCs (2.9 percent in Kumarkhand, 25.8 percent in Bihariganj). As indication of food being served we considered food being prepared or served, food leftovers in the AWC, and children or workers doing dishes. The share of AWCs serving food on the day of the visit may be underestimated as the survey captured only one moment during the opening hours of the AWCs, which might not coincide with the time the morning snack or lunch was served or prepared. Moreover, the food might not be prepared directly at the AWC, e.g. if there is no kitchen available or the AWW prefers to prepare the food at her home or elsewhere.

In terms of the activities observed, AWCs selected for the second visit also seemed more functional. Children were observed to be learning and any indication of food was observed in 20.2 percent and in 26.6 percent of AWCs, respectively, at the first visit. However, at the second visit, these shares dropped to 12.7 percent and 19.8 percent, again closer to the observation in the full sample at the first visit.

The overall picture of attendance and activities at AWCs in Madhepura was bleak. We even observed AWC buildings used for storing harvest or other things and that appeared not to be used as AWCs for at least some time. However, we also observed positive examples of highly functional AWCs. Enumerators reported several exceptional AWCs during the surveys. In addition, we also identified highly functional AWCs while visiting selected AWCs multiple times in short succession to select AWCs that could serve as benchmark centres for the evaluation of a nutrition intervention. At these AWCs the AWWs and about 40 children were present every time we passed, and meaningful activities were going on. We observed AWCs attended by more than 40 children, eating on proper plates and sitting on mats, or the AWH preparing food and the AWW teaching. In one AWC, children were even served additional fruit. In another AWC, the observed activities included learning about body parts and singing English songs.

Figure 2.4: Indicators of functionality



Note: Share of AWCs (1) that were observed to be open, (2) where AWW was present, (3) where 20 or more children were present, (4) where children were observed to be learning or engaged in structured playing, (5) where any indication of food being served on that day was observed, among (a) full sample of AWCs at first visit (dark blue), (b) sample of running AWCs at first visit (medium blue), and (c) sample of running AWCs at second visit (light blue).

2.4.2 AWW characteristics and working conditions

In the phone survey, 1586 AWWs were reached, of which 1580 (99.6 percent) gave consent to participate in the survey. Incorrect phone numbers were the main reason AWWs could not be reached. The average age of AWWs in our sample was 38 years. Most were between 25 and 44 years (66.0 percent), while 29.3 percent were older than 44 years and 4.7 percent were younger than 25 years (see Table A 2.1). The majority of AWWs completed secondary education (53.2 percent) or graduated (29.3 percent), highlighting their qualification.

AWWs seemed to be burdened with an immense workload. On average, AWWs reported 6.3 different task categories without being prompted. Tasks mentioned included counselling pregnant women (mentioned by 82.2 percent of AWWs), supporting the distribution of vaccinations (77.6 percent), completing surveys (65.5 percent), teaching and playing with the children (63.4 percent), non-teaching activities with the children (61.9 percent), tasks related to child health (51.3 percent), organising or handing out take-home rations (50.1 percent), completing records or other administrative tasks (40.7 percent), counselling adolescent girls (32.8 percent), providing food (31.5 percent), polio vaccination and ORS distribution campaigns (24.0 percent), participating in meetings and trainings (21.8 percent), and COVID-19-related tasks (18.7 percent). It is noteworthy that tasks not intended as core activities of AWCs are mentioned by many AWWs. Teaching and playing with children and providing food were not mentioned by all AWWs although these constitute two of the main purposes of AWCs. It is possible that the provision and preparation of food was understood to be primarily the tasks of AWHs.

A further burden on AWWs seemed to be problems with funds for their AWCs and their own stipend[1]. The AWW requires AWC funds for buying ingredients for the daily lunch, among other things. 80.3 percent of AWWs mentioned that they had some problem with these funds, either they had not received them for each month, they had not received the full amount, or the funds had not arrived on time. While the largest problem seemed to be not receiving the funds for every month (70.9 percent), 28.2 percent reported that they had not received the funds on time and 11.9 percent reported not receiving the full amount. These problems seemed even worse

[1] AWWs receive little financial compensation in form of a stipend, as they are seen as “volunteers” rather than regular workers earning a salary (Verma et al., 2018).

for the stipend of the AWWs, where 96.8 percent mentioned any problem with their stipend. 65.2 percent reported to have received their stipend only for some or very few months, 40 percent reported not to have received the full amount even when they get it, and 94.9 percent had not received it on time. While irregularities in the flow of AWC funds can hinder the smooth provision of services, irregularities in the stipend of AWWs can greatly impede their motivation to open the AWC every day and perform their multiple tasks.

2.5 Explaining functionality

The overall picture of AWC functionality seemed poor in Madhepura district before the closure due to the COVID-19 pandemic in 2020 with stark variation across blocks. Moreover, working conditions of AWWs were difficult throughout the district. At the same time, we observed several highly functional AWCs. This leads to the question what could explain this variation in functionality. In the following, we try to identify factors associated with functionality. We use simple linear regressions with robust standard errors to regress several indicators of functionality on a range of potential explanatory variables. The sample used in this analysis is smaller than the full sample due to missing information in explanatory variables. We restrict the sample to those observations for which all explanatory and all outcome variables are available. Furthermore, we include only those AWWs who were in charge during the observational survey in 2019 and the phone survey in 2021. Characteristics of the analysis sample are compared to the full sample in Table A 2.1.

2.5.1 Definition of functionality outcomes

We capture functionality of AWCs by looking at five different aspects. Due to evidence of incorrect reporting in administrative records (Deshpande, 2019; Fraker et al., 2013) and the tendency of functionaries to give ready-made answers (Verma et al., 2018), we rely only on indicators that could be directly observed. As a minimum requirement, the AWC needed to be open at the time of the unannounced visit, meaning that any person, either children, the AWW or AWH, was present at the AWC (outcome *open*). A second aspect of functionality is the presence of the AWW (outcome *AWW present*). An AWW could be absent for justified reasons such as a monthly meeting of AWWs or ICDS training. The AWC could still function in that case, if the AWH or often also a relative carries out her tasks during her absence. In fact, enumerators observed several instances where a relative was helping at the AWC and a qualitative study in Bihar reported that most AWWs were supported by family members in their daily tasks (John et al., 2020). However, as the person carrying the main responsibility for the AWC, frequent absence of the AWW is an indicator of low functionality. The attendance of children is another aspect of functionality. Each AWC is supposed to provide on-site services daily to the registered children, usually 40 children for regular AWCs and 20 children for Mini

AWCs. Demand for AWC services, and therefore the number of attending children, could be low for various reasons irrespective of the quality of service provision. However, we assume that poor provision of AWC services, such as irregular meals or frequent absenteeism of AWW and AWH, is a major factor for low attendance. In fact, meal provision seems to be positively associated with the number of children present at an AWC (Fraker et al., 2013). We define one outcome of functionality as a binary variable equal to one if 20 or more children were observed at the AWC at the time of the unannounced visit (outcome *20 or more children*). Two further outcomes relate to key tasks of the AWCs, the provision of pre-school education and nutrition. One indicator captures whether any children at the AWC were observed to be learning or participating in structured play (outcome *learning*) and the other if any indication of food served on this day was observed (outcome *food indication*). Indication of food includes children eating, the AWW or AWH preparing or serving food, or washing dishes, or cooked food or leftovers around the AWC.

2.5.2 Definition of explanatory variables

The choice of potential explanatory variables is based on theoretical reasoning and relevant literature. Firstly, we include several sociodemographic characteristics of AWWs as explanatory variables. AWWs' age is closely linked to the duration of her work experience as AWW. Moreover, younger AWWs were recruited following different rules compared with older AWWs. Age might also capture other life experiences, such as own experience of raising children, or confidence and independence that comes with increasing age. We code the age of AWW in four categories, younger than 25 years, 25 to 34 years, 35 to 44 years, and 45 years or older. Education of AWWs is included as more educated AWWs might have a better understanding of the ICDS programme and find it easier to fulfil their tasks, especially the counselling of pregnant women and adolescent girls on nutrition or filling growth charts. While the level of knowledge of nutrition requirements among AWWs has been reported to be relatively high in Bihar due to investments in training centres (Maity, 2016), awareness of ICDS programmes has been identified as a weakness in other settings (Parmar et al., 2015). Education is categorised into completion of primary or lower secondary school, higher secondary school, and graduation. Ownership of land captures whether the AWW and her family owned land and is a proxy for the wealth or socioeconomic standing of the AWW and her family. A wealthier AWW might be less dependent on the income generated through her work as AWW, because other income sources can sustain her family and herself. This might change the role of financial incentives for managing a functional AWC. Ownership of a scooter captures whether the AWW or her family owned a motorbike or scooter and proxies for both her wealth and mobility. Some AWWs might require transportation for reaching their AWCs and owning a scooter could facilitate this. Even if the AWC is in walking distance, attending regular meetings that take place at the ICDS block office or elsewhere could be simplified if the AWW or a family member owned a scooter. Transport duration between AWW's home and AWC, similarly, captures the ease of the AWW to reach her AWC. It is measured in minutes. We assume that AWWs living more closely to their AWCs are more likely to be there, either because it is easier to reach the centre or because they might be more embedded in the social structure of the village. We furthermore include an indicator that captures whether the AWW identified herself as being from the same caste as the majority of people in her village or from the same religion in case of a muslim AWW. This could proxy for the sense of belonging and responsibility the AWW feels towards the beneficiaries of her AWC. Caste dynamics have been identified as drivers of performance as tensions can arise if the AWW and the majority of villagers belong to different caste groups (John et al., 2020). We also incorporate two binary variables denoting if the AWW had contact with other AWWs at least every week and if the AWW had contact with her Lady Supervisor at least every week. With these indicators we want to capture the integration of the AWW into the ICDS system of functionaries, the strength of her network. A priori, it is not clear whether more integrated AWWs would perform better or worse. They could perform better if the frequent exchange helps to perform the daily duties and overcome challenges. On the other hand, they could perform worse if the strong integration and potentially informal systems of payments between functionaries (Deshpande, 2019) mean that the AWWs do not feel the need to perform their duties well. It is also possible that Lady Supervisors visit less functional AWCs more often to support the AWW. As mentioned above, financial incentives have been identified as drivers of performance (John et al., 2020). We therefore include a binary variable indicating whether the AWW reported not to have received her full stipend. As a further explanatory variable proxying the characteristics of the AWC building, we add a binary variable capturing whether the building had any walls made of concrete. More solid walls could provide better protection from weather to the children, making more children attend the AWC, and allowing the AWC to be open on more days. This indicator could also capture whether the AWC is located in a slightly better-off area. We also want to consider structural differences. We add a binary variable for being a Mini AWC. Mini AWCs cater to up to 20 instead of 40 children on site and they do not have an AWH. Thus, we would expect fewer children to be present in a Mini AWC on any given day. Moreover, as the AWW has to fulfil the tasks of the AWH in addition to her own, she could be absent more often while she is preparing food, food could be served less frequently, or learning could be going on less often. We also include block fixed effects as blocks seemed to differ fundamentally in their organisational management. To control for survey artefacts, we add a variable denoting if the day of the observational visit to the AWC coincided with a training day. From the district level, we received information about planned training days. In addition, we identify training or meeting days through the daily reports of enumerators. On a training or meeting day, the AWW is less likely to be present at the AWC, the AWC might be closed completely, less learning may take place, food might be less likely to be served. Lastly, we include enumerator fixed effects based on enumerators conducting the phone survey as their personalities might have led to different answer patterns across AWWs. As a robustness check, we use observational survey day fixed effects instead of block fixed effects, dropping the indicator for training day from the list of explanatory variables.

2.5.3 Results of regression analysis

Results of the regression analysis are presented in Table 2.1. Most notable in the regression results is the overall lack of robust associations.

Two individual characteristics show associations. Longer duration of transport to the AWC slightly is associated with a lower probability of observing 20 or more children in the AWC and any indication of food. However, these coefficients are small. A longer duration by 10 minutes would is associated with a 0.9 percentage point lower probability of observing 20 or more children at the AWC and a 1.0 percentage point lower probability of observing any indication of food. The average duration between AWW residence and AWC is 14 minutes (see Table A 2.2 in the appendix). While the age of AWWs seems to be negatively correlated with the AWC being open, and fewer children were observed in AWCs of AWWs in the oldest age group, age was not associated with the other aspects of functionality, AWW present, learning, and food indication.

As expected, structural differences become apparent. Being a Mini AWC is associated with a 15.8 percentage point lower probability of the AWC being open, a 16.4 percentage point lower probability of 20 or more children being present, and a 7.3 percentage point lower probability of any indication of food being observed at the AWC. These sizeable associations may reflect that AWWs in Mini AWCs have to shoulder a larger number of tasks. Differences between blocks are similarly large. Compared to Kumarkhand, AWCs in Bihariganj are significantly more likely to be observed as open (by 18.4 percentage points), with 20 or more children attending (by 14.3 percentage points) and to serve food (by 21.1 percentage points).

The survey day coinciding with a training or meeting day was associated with a 30.3 percentage point lower probability of the AWW being present at the AWC, as expected.

No association was found for the remaining variables. The education of the AWW was not associated with any indicator of functionality, and neither was the ownership of land or a scooter of the AWW's family. Being of the same caste or religion as the majority of villagers, frequent contact with other AWWs or the Lady Supervisor, the AWW not receiving her full stipend, and the AWC having concrete walls were all not statistically significantly associated with the five indicators of functionality. The adjusted R-squared indicates that even with this wide range of explanatory variables, we can only explain a relatively small share of the variation in functionality.

Table 2.1: Estimation results for functionality outcomes

	(1) Open	(2) AWW present	(3) 20 or more children	(4) Learning	(5) Food indication
Age in years					
25-34	-0.128*	0.00860	-0.0609	0.0416	-0.0800
	(0.0164)	(0.9060)	(0.2978)	(0.3576)	(0.1801)
35-44	-0.171**	0.00183	-0.112	0.00192	-0.0607
	(0.0012)	(0.9800)	(0.0546)	(0.9659)	(0.3084)
45+	-0.181***	-0.0468	-0.124*	0.00266	-0.0614
	(0.0008)	(0.5251)	(0.0347)	(0.9533)	(0.3076)
Education					
Higher secondary	0.0370	0.00318	0.0396	-0.00536	0.0190
	(0.2774)	(0.9292)	(0.0843)	(0.8236)	(0.4821)
Graduation	0.00708	-0.0170	0.0377	0.0122	-0.0215
	(0.8526)	(0.6679)	(0.1381)	(0.6456)	(0.4519)
Owns land	-0.0344	-0.00897	-0.0230	0.0259	0.0105
	(0.2515)	(0.7886)	(0.3295)	(0.2326)	(0.6590)
Owns scooter	0.0270	0.0288	0.0144	-0.0131	0.0364
	(0.2975)	(0.3204)	(0.4466)	(0.4896)	(0.0806)
Transport duration	-0.00126	-0.000914	-0.000928*	-0.000757	-0.000991*
	(0.1434)	(0.2673)	(0.0440)	(0.0936)	(0.0262)
Same caste or religion	0.0264	-0.0228	0.00825	0.0223	-0.000528
	(0.2969)	(0.4165)	(0.6551)	(0.2191)	(0.9792)
Contact with AWWs	-0.0221	0.00553	0.00959	0.00802	0.00684
	(0.4143)	(0.8504)	(0.6394)	(0.6904)	(0.7565)
Contact with supervisor	-0.0334	-0.0396	-0.00861	0.000953	-0.0744**
	(0.3973)	(0.3474)	(0.7759)	(0.9742)	(0.0077)
Not full stipend	0.0135	0.0408	0.00756	0.0302	0.00320
	(0.6620)	(0.2302)	(0.7472)	(0.2027)	(0.8898)
Concrete walls	0.0102	0.0482	-0.00333	-0.00814	0.00752
	(0.6732)	(0.0786)	(0.8622)	(0.6738)	(0.7024)
Training day	-0.0576	-0.303***	-0.0469	-0.0120	-0.0403
	(0.1732)	(0.0000)	(0.0558)	(0.6676)	(0.0891)
Mini AWC	-0.158***	0.0617	-0.164***	0.0125	-0.0733**
	(0.0003)	(0.1859)	(0.0000)	(0.7027)	(0.0094)
Blocks					
Bihariganj	0.184***	0.113	0.143**	0.00365	0.211***
	(0.0009)	(0.0742)	(0.0046)	(0.9337)	(0.0000)
Gailadh	0.0591	-0.00706	-0.0934*	0.0359	0.0224
	(0.3673)	(0.9203)	(0.0158)	(0.4749)	(0.5460)
Gamharia	0.0602	-0.0323	-0.00719	0.0398	0.102*
	(0.3708)	(0.6552)	(0.8735)	(0.4386)	(0.0212)
Gwalpara	0.0500	-0.0942	0.0465	-0.0527	0.104**
	(0.3884)	(0.0660)	(0.2279)	(0.1164)	(0.0038)
Madhepura City	0.110*	0.0135	-0.0268	0.0283	0.0810**
	(0.0339)	(0.7952)	(0.4262)	(0.4590)	(0.0050)
Murliganj	0.0949	-0.148*	-0.00220	-0.0448	0.120***
	(0.0735)	(0.0112)	(0.9536)	(0.2225)	(0.0006)
Shankarpur	0.185**	0.0813	0.0531	0.0616	0.162**
	(0.0013)	(0.2453)	(0.3071)	(0.2362)	(0.0011)
Singeshwar	0.183***	-0.0605	0.00587	-0.00230	0.159***
	(0.0006)	(0.3326)	(0.8871)	(0.9568)	(0.0001)
Udakishunganj	0.0642	-0.0376	-0.0431	-0.0290	0.0616*
	(0.2060)	(0.4409)	(0.1613)	(0.3511)	(0.0244)
Mean dep. var.	0.76	0.43	0.12	0.11	0.14
Adj. R^2	0.03	0.10	0.04	0.00	0.04

Note: Estimation results of a linear probability model with robust standard errors, N = 1,309. All specifications include phone survey enumerator fixed effects. p-values in parentheses. * *$p < 0.05$,* ** *$p < 0.01$,* *** *$p < 0.001$*

Results of the robustness check using survey day fixed effects are very similar (see Table A 2.3).

2.6 Discussion

This comprehensive study of the state of AWCs in ten blocks of Madhepura district combines data from unannounced observational visits and phone interviews with AWWs. We observed overall poor state and low functionality with large variations and AWWs reported challenging working conditions. We could not identify individual AWW and AWC characteristics that have a statistically significant and meaningful association with functionality but observed structural aspects seem to play a large role.

Our findings of poor infrastructure, low attendance of children, substantial rates of absenteeism of AWWs, and low functionality in terms of provision of lunch and pre-school education in AWCs in Madhepura is similar to that observed in other studies (Asha, 2014; Fraker et al., 2013; Maity, 2016). Noteworthy is the variation in infrastructure and functionality across and also within blocks.

AWWs reported a high workload and problems with receiving their stipend and AWC funds. This is also in line with other studies. In a qualitative study of 30 AWWs in Bihar, the majority of AWWs reported feelings of being overburdened and demotivated by low and delayed stipends (John et al., 2020). AWCs are supposed to be open for four hours on six days per week. AWWs, with the help of the AWH, have to clean and prepare the AWC, complete attendance records, prepare and serve food to the children, and engage the children in teaching and playing activities. In addition, they have to buy food supplies for the daily meal, conduct household visits, organise take home rations, participate in monthly meetings and trainings, organise educational events for pregnant and lactating women and adolescent girls, and support the implementation of vaccination days, among other things. Anecdotal evidence provided by enumerators indicated that AWWs were also asked to supervise central exams in schools and organise and participate in rallies. This is an extremely large set of tasks for which the AWWs receive little financial compensation in form of a stipend.

Heavy workloads and weak incentives are likely to impede the effective implementation of AWC services (John et al., 2020; Verma et al., 2018). Easing the work burden could allow AWWs to fulfil their core tasks better. An experimental study in AWCs in Tamil Nadu showed that adding an extra worker for pre-school education to the AWC increased the functionality of AWCs in multiple ways (Ganimian et al., 2021). AWCs that could hire the extra worker were open more often, the absence of the AWWs reduced, instructional time received by children doubled, and rates of stunting and severe malnutrition were lower 16 months after programme rollout. While heavy workloads and weak incentives might explain the overall deficient picture, they cannot explain the large variation in functionality as these challenges are universal to AWWs.

Identifying factors explaining the variation in functionality of AWCs would be highly informative. In our analysis no personal characteristic was robustly associated with different indicators of functionality. We found that some structural aspects showed stronger associations, specifically being a Mini AWC and block indicators. Nevertheless, the combination of all factors included in the estimation models did not explain a substantial fraction of variation in the indicators of functionality.

One study of 200 AWCs in Kerala identified educational status, job status, infrastructure of the AWC, supervision, coordination, and community participation to be associated with AWC efficiency (Asha, 2014). Besides the different setting, the main difference to our study is the use of a score based on six services provided at the AWC to measure efficiency instead of using directly observable indicators. A qualitative study of 30 AWWs in Bihar identified a range of factors influencing performance, defined as provision of services with required quality (John et al., 2020). These included financial motives and family support, service preferences of beneficiaries and AWWs, work environment including workload and stipend, caste dynamics, and corruption. While we included indicators of financial incentives and caste dynamics in our regression models, these did not appear to be statistically significantly associated with measures of AWC functionality.

We believe the major reason for a lack of robust associations is that the easily observable individual characteristics are not the most relevant factors explaining functionality. Rather, structural factors and more private, individual characteristics might be more important.

While our study showed differences across blocks descriptively, it is not able to assess the structural aspects systematically. Anecdotal evidence suggests differences in managerial structures, working relationships, and communication between CDPOs, Lady Supervisors, and AWWs between blocks. Lists of mobile phone numbers of AWWs, for example, were more or less up-to-date and accurate across supervisors, suggesting that some supervisors were more frequently in contact with their AWW via phone than others. The share of AWWs that could be reached for the phone survey ranged between 78.8 and 97.3 percent across blocks. For an iron supplementation intervention, we invited 556 AWWs to a workshop by communicating the invitation through the supervisors. Attendance rates varied from 78 to 100 percent of invited AWWs across blocks. This also suggests differences in organisation and communication. Similarly, some blocks seemed more organised than others in managing data. Systematically analysing such structural organisational differences could prove very beneficial.

On the level of the individual AWW, we assume motivation and other personality traits to be major drivers. Especially in the challenging working environment, with many factors affecting their work outside their control (John et al., 2020) and their very low financial compensation, the AWWs' intrinsic motivation is key. They have to make the best in a difficult situation, navigating their many tasks, often in a setting of corruption and tensions between functionaries and beneficiaries (Deshpande, 2019). We did not attempt to assess motivation of AWWs as a comprehensive assessment is difficult in a phone survey. We also omitted questions on job satisfaction because of doubts whether AWWs would feel comfortable to report very personal or potentially critical opinions on the phone and instead give ready-made answers (Verma et al., 2018). However, we included a short scale measure of locus of control (Kovaleva et al., 2012). This scale did not seem to work in the specific setting of the study and including results of this scale in the regression analysis did not show any association with the functionality indicators. An in-depth assessment of motivation could provide important insights to understanding the factors hindering optimal service delivery and ultimately improving it.

While our study provides a comprehensive overview of the situation of all AWCs in the 10 blocks of Madhepura, it is not equally suitable to measure individual functionality. Our indicators of functionality are dichotomous definitions based on a one-time observation. This entails several related problems. First of all, functionality and quality of service provision is a continuous concept. Any cut-off and translation into binary variables will lead to a loss of information. Observing functionality at one point in time only, we do not capture frequency and quality of service provision. It further means that for an individual AWC functionality might be overestimated. While non-functional as well as highly functional AWCs will be classified correctly, observations on semi-functional AWCs only describe the average correctly.

Furthermore, we only focus on factors of the supply side of service provision in AWCs and not on the demand side. There is evidence suggesting that differences in the awareness level among intended beneficiaries contribute to the explanation of variation in functionality (Maity, 2016) and that underutilisation of ICDS services may be a concern (Dasgupta et al., 2012).

A further concern is the long time between the observational survey and the phone survey, during which AWCs were closed. One could argue that the AWWs were not fulfilling their regular role as AWWs at the time of the phone survey due to the COVID-19 pandemic, changing their answers as a result, or might not be in contact with other AWWs or their supervisors as usual. We addressed this concern by referring to the time before the pandemic in the questions relating to their work. Also, we only included AWWs that were already in charge of their AWC in 2019 in the regression analysis.

One final limitation of this study relates to its descriptive nature. Our analysis cannot identify causal relationships between functionality and the included AWW and AWC characteristics.

2.7 Conclusion

We found large gaps in the functionality of AWCs in Madhepura district in terms of services provided daily on site, namely pre-school education and daily meal provision for children aged three to six years. Functionality, as measured by five observable indicators, varied considerably within and across blocks. The potential of AWCs to act as platforms for delivering additional programmes appeared limited. There were positive examples of highly functional AWCs. However, we could not robustly identify factors associated with several aspects of functionality and why some AWCs were highly functional and others were not. Monitoring the state of AWCs regularly would be important, especially before new tasks are given to AWWs.

Several reforms have been implemented by the ICDS since the re-opening of AWCs after the COVID-19 pandemic. However, reforms per se do not necessarily lead to improvements if structural problems are not addressed. In fact, the vigilance-focused reforms implemented in Bihar before 2015 did not lead to more effective service provision and instead aggravated problems of corruption (Verma et al., 2018). We believe that AWCs provide a great framework to tackle child malnutrition and support child development in theory. However, the current organisation and infrastructure creates severe barriers to its proper functioning. We therefore reiterate recommendations made elsewhere to support improved service delivery through AWCs. AWWs are the largest cadre of community health workers globally (John et al., 2020) and fulfil a very important role for child development and for promoting healthy lives. It seems likely that AWWs would be enabled to perform their work better if their services were regularized and not seen as “volunteer work”, if they received regular remuneration, and if their workload was rationalized and paperwork reduced.

2.8 Appendix

Table A 2.1: Sample distribution across blocks

	Observational sample		Phone survey sample		Analysis sample	
	N	Share	N	Share	N	Share
Bihariganj	155	0.090	137	0.086	112	0.086
Gailadh	108	0.063	105	0.066	87	0.067
Gamharia	101	0.059	94	0.059	80	0.061
Gwalpara	138	0.080	135	0.085	109	0.084
Kumarkhand	278	0.161	233	0.147	201	0.154
Madhepura City	255	0.148	215	0.136	164	0.126
Murliganj	213	0.123	198	0.125	167	0.128
Shankarpur	115	0.067	117	0.074	83	0.064
Singheswhar	162	0.094	160	0.101	138	0.106
Udakishunganj	201	0.116	192	0.121	163	0.125

Note: Number and share of observations included in the observational survey (columns 1 and 2), phone survey (columns 3 and 4), and regression analysis (columns 5 and 6), across blocks. BH = Bihariganj, GL = Gailadh, GM = Gamharia, GP = Gwalpara, KK = Kumarkhand, MC = Madhepura City, MG = Murliganj, SP = Shankarpur, SG = Singheswar, UK = Udakishunganj.

Table A 2.2: Sample characteristics

	Total Sample		Analysis sample	
	N	Share (95 percent-CI)	N	Share (95 percent-CI)
Age in years	1,563			
<25		0.047 (0.036 - 0.057)	1,304	0.035 (0.025 - 0.045)
25-34		0.336 (0.312 - 0.359)	1,304	0.328 (0.303 - 0.354)
35-44		0.324 (0.301 - 0.348)	1,304	0.340 (0.314 - 0.365)
45+		0.293 (0.270 - 0.316)	1,304	0.297 (0.272 - 0.322)
Education	1,580			
Primary/secondary		0.175 (0.157 - 0.194)	1,304	0.169 (0.149 - 0.190)
Higher secondary		0.532 (0.507 - 0.556)	1,304	0.534 (0.507 - 0.561)
Graduation		0.293 (0.271 - 0.316)	1,304	0.297 (0.272 - 0.322)
Owns land	1,531	0.805 (0.786 - 0.825)	1,304	0.799 (0.777 - 0.821)
Owns scooty	1,573	0.655 (0.631 - 0.678)	1,304	0.660 (0.634 - 0.685)
Transport duration	1,576	13.713 (12.996 - 14.431)	1,304	14.034 (13.214 - 14.853)
Of same caste/religion	1,562	0.576 (0.552 - 0.601)	1,304	0.579 (0.552 - 0.606)
Contact w AWWs every week	1,500	0.310 (0.287 - 0.333)	1,304	0.311 (0.285 - 0.336)
Contact w supervisor every week	1,493	0.115 (0.099 - 0.131)	1,304	0.123 (0.106 - 0.141)
Not full stipend	1,465	0.400 (0.375 - 0.425)	1,304	0.413 (0.386 - 0.439)
Concrete walls	1,548	0.551 (0.526 - 0.576)	1,304	0.554 (0.527 - 0.581)
Mini AWC	1,548	0.107 (0.091 - 0.122)	1,304	0.105 (0.088 - 0.122)
Training day	1,548	0.262 (0.240 - 0.284)	1,304	0.260 (0.236 - 0.284)

Note: Shares of AWWs and AWCs with the respective characteristics; total sample includes all observations available for this characteristic; analysis sample includes sample used in regression analysis.

Table A 2.3: Estimation results for functionality outcomes – robustness: survey day fixed effects

	(1) Open	(2) AWW present	(3) 20 or more children	(4) Learning	(5) Food indication
Age in years					
25-34	-0.137**	-0.0113	-0.0747	0.0399	-0.0779
	(0.0070)	(0.8664)	(0.2039)	(0.3812)	(0.1877)
35-44	-0.183***	-0.0201	-0.134*	-0.00330	-0.0610
	(0.0003)	(0.7666)	(0.0219)	(0.9418)	(0.3004)
45+	-0.189***	-0.0659	-0.138*	0.00194	-0.0630
	(0.0003)	(0.3297)	(0.0195)	(0.9659)	(0.2903)
Education					
Higher secondary	0.0362	-0.00696	0.0335	-0.00627	0.0181
	(0.2917)	(0.8400)	(0.1374)	(0.7976)	(0.5072)
Graduation	0.00577	-0.0362	0.0330	0.0114	-0.0258
	(0.8813)	(0.3455)	(0.1959)	(0.6755)	(0.3762)
Owns land	-0.0295	-0.00357	-0.0212	0.0217	0.0152
	(0.3352)	(0.9134)	(0.3556)	(0.3252)	(0.5237)
Owns scooter	0.0282	0.0446	0.0163	-0.00949	0.0380
	(0.2775)	(0.1115)	(0.3870)	(0.6170)	(0.0693)
Transport duration	-0.00114	-0.000619	-0.000806	-0.000765	-0.000946*
	(0.1633)	(0.4219)	(0.0762)	(0.1055)	(0.0349)
Same caste or religion	0.0319	-0.0210	0.00919	0.0227	-0.00280
	(0.2016)	(0.4400)	(0.6169)	(0.2094)	(0.8899)
Contact with AWWs	-0.0219	0.00349	0.00865	0.00965	0.00694
	(0.4174)	(0.9029)	(0.6737)	(0.6312)	(0.7537)
Contact with supervisor	-0.0221	-0.0196	-0.00540	0.00361	-0.0737**
	(0.5742)	(0.6306)	(0.8560)	(0.9035)	(0.0077)
Not full stipend	0.0106	0.0304	0.00818	0.0281	0.00304
	(0.7297)	(0.3503)	(0.7246)	(0.2326)	(0.8961)
Concrete walls	0.000420	0.0328	-0.00248	-0.00958	0.00597
	(0.9863)	(0.2166)	(0.8966)	(0.6184)	(0.7606)
Mini AWC	-0.173***	0.0379	-0.169***	-0.00127	-0.0720*
	(0.0001)	(0.4246)	(0.0000)	(0.9698)	(0.0128)
Mean dep. var.	0.76	0.43	0.12	0.11	0.14
Adj. R^2	0.05	0.16	0.06	0.01	0.04

Note: Estimation results of a linear probability model with robust standard errors, N = 1,309. All specifications include observational survey day and phone survey enumerator fixed effects. p-values in parentheses. * *$p < 0.05$,* ** *$p < 0.01$,* *** *$p < 0.001$*

Chapter 3

Introducing a new method of iron supplementation into Anganwadi Centres in rural Bihar, India: a randomized controlled trial

with: Ann-Charline Weber, Abhijeet Kumar, and Sebastian Vollmer

Abstract

To reduce anaemia prevalence, we tested the introduction of cast iron ingots to prepare iron-enriched drinking water in pre-schools, so called Anganwadi Centres (AWCs), in India. We implemented a randomized controlled trial among 820 AWCs with two different treatment arms, one receiving more comprehensive training and support than the other. The intended primary outcome, children's anemia status, could not be measured due to COVID-19 and temporary closure of the AWCs. Instead we used observations about the functionality of AWCs and a phone survey with Anganwadi Workers (AWWs) to evaluate the intervention. Outcomes included AWW mentioning the newly introduced method without being prompted, and AWW remembering detailed steps of the method, and observed indications of the method's use in AWCs. While a large share of AWWs in both treatment groups remembered the method and could explain its steps, indication of its usage was generally rare but more common in the intensive treatment arm. Successful implementation of the iron supplementation method in the daily routine of AWCs would require institutional support and substantial improvements in AWC functionality.

Study pre-registration: Bogler, Lisa et al. 2020. "Introducing a method of iron supplementation to Anganwadi Centres in Madhepura, Bihar, India." AEA RCT Registry. August 18. https://doi.org/10.1257/rct.5449

3.1 Introduction

Anaemia is a severe global health concern. In 2019, the global prevalence was 23 percent among all age groups and with 40 percent highest among children under five years (Gardner & Kassebaum, 2020). In India, the prevalence was even higher. Most recent data from the National Family Health Survey 2019-2021 (NFHS-5) reports that 67 percent of children under 5 years had some form of anaemia (International Institute for Population Sciences (IIPS) and ICF, 2021). Iron deficiency is thought to be one of the main reasons for anaemia (World Health Organization, 2016). Iron deficiency anaemia especially is a concern in children due to its association with reduced physical and cognitive performance and development (Bobonis et al., 2006; Halterman et al., 2001). The World Health Organization therefore recommends daily oral iron supplementation to reduce iron deficiency anaemia among children (World Health Organization, 2016). Reducing iron deficiency anaemia has long been emphasised in the Indian public health agenda, starting with the National Nutritional Anaemia Prophylaxis Programme launched in 1970 as the first concerted effort and the National Iron Plus Initiative launched in 2013. India’s National Nutrition Strategy emphasises dietary diversification, distribution of iron and folic acid tablets or syrup to vulnerable groups, and the usage of salt fortified with iron and iodine in meals in pre-schools and primary schools. However, implementation is still deficient. Supplementation with tablets and syrup may be impaired due to the work-intensive administration or insufficient compliance of beneficiaries (Kapil et al., 2019), while the use of double fortified salt is limited by its unavailability in the market in some parts of India.

The Lucky Iron Leaf® (also marketed as Lucky Shakti Leaf®, Lucky Iron Leaf in the following) is a cast iron ingot that enriches water or food with bio-absorbable iron while cooking and may provide a low-cost, low-dosage, sustainable alternative method for regular iron supplementation. It is a design variation of the Lucky Iron Fish®, that was specifically adapted to the Indian context for a different study (Ebert et al., 2020). The potential of iron ingots for iron supplementation has been shown, the Lucky Iron Fish® effectively reduced anaemia among rural Cambodian women (Charles et al., 2011, 2015). The challenge, however, is to ensure its daily use. A study in Bihar, India, showed that households who had received the Lucky Iron Leaf were reluctant to use it, preventing the health benefit from being realised (Ebert et al., 2020). Usage is likely higher in an institutional setting with centrally regulated meals. This is suggested by findings on the use of iron-fortified salt, another method of low-dose iron supplementation. While the distribution of iron-fortified salt to households in Bihar did not lead to a reduction in anaemia (Banerjee et al., 2018), the distribution to primary schools for use in the school kitchens, also in Bihar, resulted in considerable reductions in anaemia even four years after the start of the intervention (Krämer et al., 2021; von Grafenstein et al., 2021).

In this study, we introduced the Lucky Iron Leaf to Anganwadi Centres (AWCs) in Madhepura district, Bihar. AWCs are an integral part of the Integrated Child Development Services (ICDS) scheme, the largest programme of the Indian government to support early childhood development. At the AWC, among other services, the Anganwadi Worker (AWW) and a helper (AWH) provide pre-school education as well as a daily meal to children aged three to six years. The Lucky Iron Leaf was presented as a sustainable method to enrich the drinking water that is given to attending children with iron. We hypothesized that implementation on the institutional level would increase the likelihood of daily use, reducing anaemia among children attending the AWCs.

We aimed to evaluate the introduction of the Lucky Iron Leaf to AWCs using a randomized controlled trial (RCT) in Madhepura district, Bihar. The originally intended outcome measure was the hemoglobin level in the blood of children who were attending the AWCs. As this was not possible due to the COVID-19 pandemic and long closures of the AWCs, we assessed indicators of knowledge and observed indicators of usage among AWWs as endpoints. We argue that knowledge and usage of the Lucky Iron Leaf are the relevant first steps before any impact on health can be realized. Only if the Lucky Iron Leaf is used regularly and correctly, its use can affect haemoglobin levels and anaemia rates.

We find evidence that the workshops introducing the Lucky Iron Leaf were effective in generating knowledge about their use. However, observed usage and general functionality of AWCs was low. While the introduction of the Lucky Iron Leaf to AWCs by training AWWs seems feasible, it remains unclear whether a positive health outcome can be achieved in the current state of AWCs. We discuss several barriers to address this concern.

The rest of the paper is structured as follows. Section 2 gives an overview of the context and describes the intervention in detail. The study design and data are explained in section 3, the method in section 4. Results are presented in section 5 and discussed in section 6. The final section concludes.

3.2 Context and intervention

3.2.1 Background

The study took place in Madhepura district, in Bihar, where anaemia is a major health concern. The prevalence in children between 6 and 59 months was 67.7 percent in 2019-20, which constitutes a 10 percent increase since 2015-16 (International Institute for Population Sciences (IIPS), 2021). To address the high prevalence of anaemia, a study introduced the Lucky Iron Leaf to households in Madhepura in 2016 but did not find any effects on haemoglobin levels, due to low take-up as indicated by low self-reported usage (Ebert et al., 2020). Potential reasons given included the possibilities that beneficiaries either did not understand the process, were sceptical towards the implementors unknown to them who distributed the Lucky Iron Leaf or did not understand the necessity to use it as awareness of anaemia was very low.

Vulnerable population groups might be more successfully reached via established trusted institutional channels. The use of the Lucky Iron Leaf to fortify centrally provided meals could offer a low-cost, sustainable iron supplementation to the beneficiaries, with little continuous administration efforts. We focused on AWCs, which are a central part of the ICDS. The scheme targets children below the age of six years, pregnant women and lactating mothers to address key aspects of child development, including nutrition, education, and health and is the world's largest childhood development scheme with 86 million children below six years benefiting from its supplementary nutrition programme in 2020 (Ministry of Women and Child Development, India, 2020). 1.3 million AWCs exist across India, reaching even remote rural areas. AWCs are staffed with a trained AWW, usually a married woman fulfilling a minimum requirement of education from the area, and an AWH. One AWC usually caters to 500 to 1000 households. Among other services, each AWC serves as pre-school for 40 children between three and six years, targeting children with poor nutritional status. So called Mini AWCs cater to fewer households, offer pre-school to only 20 children and do not have an AWH. In all AWCs, the pre-school is usually open four hours a day on six days a week. Attending children are offered a snack, lunch, and some non-formal pre-school education. In Madhepura district, the lunch is prepared by the AWW and AWH at the AWC or in their private kitchen. The lunch menu is fixed centrally by the ICDS, detailing the food given to the children on each day of the week. AWWs are trained for their tasks, which include the distribution of iron tablets or syrup to adolescent girls and pregnant women, and are therefore more aware of anaemia and the importance of iron supplementation than the general population.

3.2.2 Intervention

We tested two different ways of introducing the Lucky Iron Leaf to AWCs, with a light intervention and an intensive intervention. The light intervention included a short training workshop on the Lucky Iron Leaf for the AWWs at the end of which they each received five Lucky Iron Leaves, which are necessary to prepare enough drinking water for 40 children. The intensive intervention included a longer training workshop for the AWWs at the end of which they also received five Lucky Iron Leaves, a short training workshop for the AWHs, deliveries of ingredients (lemon and sugar) required to prepare the drinking water, and a reminder phone call. The intensive intervention was designed to test whether an introduction of the Lucky Iron Leaf in AWCs in Madhepura is feasible and has the potential to improve haemoglobin levels of beneficiaries. The light intervention was designed to simulate an easy-to-implement and low-cost introduction that can potentially be scaled up by the ICDS or similar structures.

The Lucky Iron Leaf

The Lucky Iron Leaf is a cast iron ingot made of bio-absorbable iron powder (see Figure A 3.1). When it is boiled in water or with food and when fruit acid such as from lemon, tomato or tamarind is added, it releases low doses of iron into the water, which can then be absorbed by the body by drinking the water or eating the food. The quantity of iron released depends on the boiling time and acidity level. The Lucky Iron Leaf can be used for up to 5 years even when used daily.

While originally designed for use in cooking, we adjusted the method for the preparation of iron-enriched drinking water. AWCs are supposed to provide a meal as well as drinking water to the attending children. Drinking water was mostly not boiled, making the preparation of iron-fortified drinking water an extra task for AWWs and AWHs. However, most dishes of the centrally fixed lunch menu did not contain fruit acid and it seemed difficult to incorporate lemon, tomato or tamarind in a sufficient number of dishes. We also observed that many AWCs did not provide food as regular as intended.

In cooperation with the developers of the Lucky Iron Leaf, we calculated that five ingots were required for preparing drinking water for 40 children, so each child could drink 100 ml of water and receive a dose of 5 mg iron per day, equivalent to half of the daily recommended dietary allowance. The required boiling time would be 22 minutes. We experimentally tested the amount of lemon required to reach the correct level of acidity, a ph-value of 4, for iron to be released from the ingot. To cover the slightly sour taste, we decided to add sugar to the water after boiling, and experimentally tested the amount required with children. Taking into account that AWCs were not attended by 40 children each day, we prepared a manual that specified the quantity of water, lemon and sugar required, as well as the boiling time, for different numbers of children (see Figure A 3.2 in the appendix). The manual explained each step of preparation using pictures and contained a table for how to adjust the process depending on the number of children present.

Workshops

AWWs of AWCs in both treatment arms and the AWHs in the intensive treatment arm received an invitation to participate in a workshop. The invitation was delivered through their supervisors who received a list of participants for each training by the research team. The workshops had 10 to 15 participants each and were led by teams of two facilitators. The workshops took place in locations convenient for the respective participants, as suggested by the responsible supervisor, to reduce traveling time for participants. This was either a suitable AWC or the block-level office of the ICDS. The workshops were planned as interactive sessions that would involve AWWs actively. During the workshop, the research project was introduced to the participants. It was explained that this workshop had the permission of the ICDS, but was conducted by a research group. At the end of the workshop, all participants received five Lucky Iron Leaves in a bag for storage together with a manual. A total of 46 workshops, 24 short and 22 long, were conducted.

The short workshop took two hours. After the introduction, the group collected existing knowledge about anaemia and iron supplementation. Facilitators gave additional input on anaemia and explained how to use the Lucky Iron Leaves. Participants practiced the method in a round of mock cooking. One pot of iron-enriched water was prepared by facilitators while participants watched and everyone tasted the water. The session was closed with a discussion of take-home messages, collected by participants on posters.

The intensive workshop took four hours and contained several additional elements. After the introduction of the project, their workshop in addition included a game to make AWWs interact as a group. AWWs were asked to share their visions of a perfect AWC and create posters of their visions in groups of two. Two other game-like activities showed AWWs that they can overcome challenges as a group. The workshop also tried to highlight the importance of the AWWs' work, potentially increasing their motivation for their daily tasks.

The workshop of the AWHs were held immediately after the workshop of the respective AWWs and followed the same structure as the short workshop for AWWs.

Workshop facilitators were women from Madhepura Sadar block, Madhepura district, with two exceptions coming from other states, and trained for 13 days by the research team. Their training included team building, confidence building, presentation skills, and workshop content. All facilitators practiced their workshop with AWWs twice. Several facilitators were trained specifically for the shorter workshop, while those who were best in leading a group and implementing participative methods were trained specifically for the longer workshops.

Delivery of ingredients

AWWs of AWCs in the intensive treatment arm received regular deliveries of lemon and sugar, the ingredients required to prepare the iron-enriched water. The first bags of ingredients were handed out directly at the workshops. The following deliveries were brought to the AWCs individually every three weeks.

Reminder Phone Call

Five weeks after the workshops, the workshop facilitators called all AWWs of AWCs in the intensive treatment arm. The facilitators asked the AWWs whether they had received their delivery of ingredients, whether they had faced any difficulties, and offered to answer any open questions.

3.2.3 Conceptual framework

The final goal of introducing the Lucky Iron Leaf to AWCs was the reduction of anaemia among children attending the AWCs. However, several assumptions need to hold for this impact to be realised. Firstly, the study builds on the assumption that AWWs are aware of iron deficiency anaemia and know about the importance of iron supplementation. The workshops introducing the Lucky Iron Leaf need to be sufficient to teach AWWs the exact process of preparing iron-fortified drinking water with the Lucky Iron Leaf and for them to accept this as a valid and effective method of iron supplementation. Furthermore, we assume that AWWs who understand the benefits of the method and the process will prepare iron-enriched drinking water for the children attending their AWC regularly, that children attend the AWC several days a week, accept and drink the water.

Several barriers might impede the realisation of health benefits. Firstly, procuring the ingredients for the preparation of iron-enriched drinking water, lemons and sugar, requires time and resources which are limited. AWWs have many tasks to complete daily and might forget the preparation of iron-enriched drinking water, have no time for it or prioritise other tasks. Additionally, problems not discussed in the one-time training might arise when preparing the iron-enriched water at the AWC, meaning that it is not prepared correctly or that AWWs are discouraged in preparing it regularly. A further key barrier might be the general dysfunctionality of the AWCs, resulting in irregularities in provision of key AWC services, and low and irregular attendance of children at the AWC (Fraker et al., 2013). Lastly, children might refuse to consume the water.

The elements of the intensive intervention aimed to address several of these barriers. The delivery of lemons and sugar removed potential budget constraints and logistical problems of procurement. It further acted as a reminder and potentially had a monitoring effect on AWWs, encouraging them to prepare the water. The additional phone calls functioned as a reminder and presented an opportunity to discuss potential questions and enhance the AWWs' understanding of the method. Most ICDS trainings target the AWWs. This intervention added a workshop for AWHs, who are usually in charge of preparing the food and therefore also likely tasked with the preparation of the iron-enriched drinking water by the AWW. This removed the additional work burden for the AWW, removed possible misunderstandings created by the

AWW explaining the task to the AWH, and added a second person at the AWC who could remember the process. Addressing general dysfunctionality of AWCs and the low number of attending children requires systemic changes beyond the reach of this study. However, the additional elements in the more intensive workshop all aimed at increasing the AWWs' motivation by reminding her of her crucial role and her influence on the well-being of the beneficiaries, by making her experience self-efficacy and by acknowledging her work.

3.3 Study design and data

We aimed to assess the effect of the two different intervention regimes in a sample of functional AWCs using an RCT with two treatment groups and a control group. Functional AWCs were selected based on observations during an unannounced visit conducted prior to the intervention roll-out. A second unannounced visit was conducted among the AWCs defined as functional directly after intervention roll-out to capture their state at a second point in time for a more accurate measure of functionality. We planned to evaluate the interventions with measures of haemoglobin levels, as indicator of anaemia, among children attending the AWCs, as well as interviews with the AWWs and AWHs six months after intervention roll-out. Due to the closure of AWCs following the outbreak of COVID-19, it was not possible to take anthropometric measures and conduct face-to-face interviews. We therefore conducted phone interviews with AWWs fourteen months after the implementation of the intervention. The timeline of the study can be seen in Figure A 3.3.

3.3.1 Sample selection and treatment assignment

Our sampling frame were all AWCs in ten blocks of Madhepura district in Bihar. We obtained lists of all AWCs registered in a block from the respective block-level offices. The first observational survey targeted all listed AWCs. The sampling frame was updated during the survey using observations of new, permanently closed, and currently vacant AWCs. A total of 1719 AWCs were visited, capturing almost the entirety of AWCs in these blocks. Out of these, only 845 AWCs met the minimum requirement of functionality for inclusion in the RCT. We defined that a minimum requirement of functionality would be fulfilled if 1) 10 or more children were present, or 2) 5 or more children and an adult were present. The adult could be the AWW, AWH, or another adult visibly involved with the children. We removed 25 AWCs from the sample that were visited during the pilot phase of the project or which were included as benchmark AWCs that were particularly closely monitored by the research team, leading to a total of 820 AWCs included in the randomization.

For randomization purposes, AWCs located at a distance of less than 100 meters to each other were treated as one unit. Treating AWCs close to each other as one unit was meant to reduce the possibility of envy and conflict between AWWs who would see their neighbour receiving treatment that they did not receive. If only one of the paired AWCs met the sample inclusion criteria, the neighbour was treated but excluded from the analysis.

AWCs in the study sample were randomly assigned into one of three groups, the control group or either of two treatment groups. We found considerable differences in AWC functionality across blocks. In addition, Mini AWCs, that provide on-site services to a maximum of 20 children, seemed to function slightly different because no AWH is hired. We therefore stratified the randomization according to block and whether the AWC was a Mini AWC or not.

Due to the nature of the intervention, AWWs were not blinded to treatment assignment. AWWs in the control group could also easily find out about the intervention by talking to their colleagues or supervisor.

3.3.2 Data

We used several sources of data for the evaluation of the intervention.

Unannounced visits

Enumerators made two unannounced visits to AWCs and recorded the presence of AWW, AWH, children, and any other adult, and their respective current activity. AWW and AWH were usually recognizable due to their uniforms. The first visit took place between November 20 to December 16, 2019 and captured all AWCs in ten blocks of Madhepura district. AWCs were observed Mondays to Saturdays between 10:15 am and 1:45 pm, as operating times were 10 am to 2 pm and this allowed a margin of 15 minutes for coming late or closing early. In addition to ongoing activities, the survey recorded physical characteristics of the AWC building, including details about its roof, walls, and floor. The second observational survey captured the subset of AWCs that were defined as functional and were included in the RCT. It took place between January 9 and February 4, 2020, following the implementation of workshops. At the time of this survey, opening hours of the AWCs were shortened due to winter coldness. Observations of AWCs were therefore made Mondays to Saturdays, between 12:00 pm and 2:00 pm. On public holidays or strike days, no observations were conducted. In addition to ongoing activities, this survey recorded indications of food and the Lucky Iron Leaf. Data collectors were blinded to treatment assignment.

Implementation information

We recorded information on the implementation of the intervention. In each training, workshop facilitators noted the actual participation of AWWs and AWHs. Data records show which AWW received the reminder phone call. Each delivery of ingredients was recorded. We can therefore capture which AWC received which component of the intervention.

Phone survey with AWWs and their supervisors

Fourteen months after the implementation of the intervention, March 22 to May 31, 2021, we conducted a phone survey with all AWWs of the ten blocks in Madhepura district as well as their supervisors. Phone numbers were collected from ICDS block-level offices and updated with the information collected during the workshops and further with the help of supervisors. In this survey, AWWs were asked about iron supplementation and the Lucky Iron Leaf as well as personal demographic characteristics, their work as AWW, and misreporting of the number of children attending AWCs. Supervisors received very similar questions.

3.3.3 Ethical considerations

The study design was reviewed by the ethics committee of the University of Göttingen, Germany, with no objections. We received approval for the study design (intervention and data collection including unannounced visits) from the ICDS Directorate in Patna, Bihar, responsible for all AWCs in Bihar, as well as the District Programme Officer in Madhepura district. We took informed consent from each AWW and supervisor personally for the phone survey. Enumerators of observational visits and the phone survey were blinded to treatment assignment. The study was registered at the AEA RCT Registry on February 11, 2020 under AEARCTR-0005449 (https://www.socialscienceregistry.org/trials/5449).

3.4 Methods

3.4.1 Estimation strategy

To evaluate the effect of the intervention, we estimate the following linear probability model:

$$Y_i = \alpha + \beta_1 T1_i + \beta_2 T2_i + \boldsymbol{x}'_{i.}\boldsymbol{\gamma} + \varepsilon_i$$

whereas Y_i is the outcome for AWC *i*. $T1_i$ and $T2_i$ indicate treatment assignment, the binary variables equal 1 if AWC *i* was assigned to the light treatment group and the intensive treatment group, respectively. The coefficients $\beta_1$and β_2 show the effects of assignment to the treatment groups, irrespective of actual participation in the workshops, receipt of Lucky Iron Leaves and other components of the intervention, i.e. the intention-to-treat effect. To increase precision, we include $\boldsymbol{x}_i$, a vector of control variables and administrative block and phone survey enumerator fixed effects. ε_i denotes the error term. As robustness checks, we estimate a model without control variables, as well as a model including only the indicator of being a Mini AWC and administrative block and phone survey enumerator fixed effects. For all models, we present coefficients of the treatment variables, β_1 and β_2, and the p-value of a test whether the two coefficients differ from another. This allows us to assess whether the intensive treatment led to greater effects than the light treatment. All analysis was conducted in Stata 16.

Outcome indicators

As we could not measure the final outcome of interest, haemoglobin levels of children attending the AWCs, we use three outcomes that capture some of the intermediate steps required for a change in children's haemoglobin levels. All three outcomes are at the AWC level. The first and second outcome are based on the phone survey with AWWs fourteen months after the workshops and twelve months after the closure of the AWCs due to the COVID-19 pandemic.

The first outcome captures whether the AWW recalled the Lucky Iron Leaves as a method of iron supplementation. It is based on the question "Which methods of iron supplementation do you know?" and is a binary variable coded as 1 if the AWW mentions the Lucky Iron Leaf without being prompted. In the following, it is referred to as *mentions LIL*.

The second outcome captures whether the AWW understood and remembered the content of the training. It is based on the question "Do you know how the Lucky Iron Leaf is used? If yes, please explain". The binary variable is coded as 1 if the AWW mentioned the key steps of using the Lucky Iron Leaf in her explanation. These key steps include boiling the Lucky Iron Leaves in water, adding lemon, and adjusting quantities and boiling time to the number of children present. In the following, it is referred to as *knows steps*. This outcome is likely influenced by the AWW's understanding of the training and her capacity to remember details of the training, but also her actual usage of the Lucky Iron Leaves. We hypothesised that AWWs who prepared iron-enriched drinking water more often are more likely to remember the process.

The third outcome is a proxy for actual usage of the Lucky Iron Leaves. It is based on the observation of the Lucky Iron Leaves in the AWCs in the second observational visit, which was conducted before the closure of the AWCs before the COVID-19 pandemic. The outcome is a binary variable coded as 1 if Lucky Iron Leaves were seen at the AWC or if data collectors observed water being stored, water being prepared, or children drinking water. This outcome is referred to as *indication of LIL*. It is a proxy for usage as several factors hinder the observation of actual usage. The observations only capture one moment in time and the iron-enriched water could have been prepared or handed out at a different time during the AWC opening hours. Moreover, some AWCs did not have a kitchen directly within or next to the building and food was prepared elsewhere, for example in the kitchen of the AWW or AWH. In this case, it was not likely to observe the Lucky Iron Leaves at the AWC. Lastly, if Lucky Iron

Leaves were stored safely while not in use, data collectors could not easily observe it at the AWC.

While we also asked AWWs directly if they had heard of the Lucky Iron Leaf, how many Lucky Iron Leaves they still had, and whether and how regularly they had used the Lucky Iron Leaves, we specifically chose variables presumably least affected by intentional misreporting for the analysis.

Covariates

As control variables we include AWC and AWW specific characteristics that might mediate the effect of the intervention, an indicator for being a Mini AWC and binary variables for the administrative blocks to account for stratification and phone survey enumerator fixed effects, as we find that enumerators recorded answers slightly differently.

We control for AWW's age and education to proxy her capacity for learning and remembering, her contact with other AWWs and her supervisor to capture her integration into the ICDS system, whether her own children attend or attended an AWC and whether she is of the same caste or same religion as the majority in the village where her AWC is located to capture her relationship with the villagers, and her perception of influence of AWWs on the functioning of AWCs to proxy her motivation to push for change. All of these variables are taken from the phone survey. In one specification of the regression with *indication of LIL* as outcome, we add further controls that might be related to the probability of using the Lucky Iron Leaves. These include the duration of transport to the AWC as proxy for ease of implementation, AWW's ownership of land and scooter as proxies for her wealth and mobility. To capture frequency of cooking at the AWC, we control for whether the AWW mentions food as one of her tasks, whether any indication of food was observed at the first observational visit, and whether 20 or more children were present at the first observational visit. Lastly, as proxies for the security of stored goods and how well the AWC can be used in inclement weather, we control for whether the AWC building had any concrete walls and whether it was possible to lock the AWC.

3.4.2 Sample and balance

In the first unannounced visit, 1719 AWCs were observed. Of those, only 845 met the minimum requirements of functionality described earlier. 25 were subsequently excluded due to their inclusion in more intensive piloting activities. The final sample of the RCT included 820 AWCs. The second unannounced visit included these 820 AWCs except one which could not be relocated. 1592 AWWs were reached during the phone survey, of which 1587 (99.7 percent) gave consent to participate in the survey.

For the analysis, we restrict the sample to those AWCs that were surveyed the second unannounced visit, whose AWW participated in the phone survey and was working as AWW already before 2020. AWWs recruited more recently, in 2020 or later, could not have participated in the intervention. This restriction leads to a sample size of 750 AWCs for the analysis. Sample sizes vary between regressions due to missing information in single outcome and control variables.

Sample characteristics

Characteristics of the AWCs and AWWs in the sample are presented in Table 3.1 and Table 3.2. Overall, functionality of the AWCs was rather low. While all AWCs were by definition open during the first unannounced visit, only 82.0 percent of these were again open at the second unannounced visit. An AWW was present in 50.5 percent of AWCs and in 35.6 percent 11 or more children were present. Learning was observed in only 12.5 percent of AWCs and any indication of food was found in 19.8 percent. More than half of AWCs had concrete walls.

Table 3.1: Sample characteristics of AWCs

	N	Total	Control	Light Treatment	Intensive Treatment
Mini AWC	750	0.087 (0.066 - 0.107)	0.080 (0.046 - 0.113)	0.083 (0.048 - 0.119)	0.097 (0.060 - 0.133)
AWC open	750	0.817 (0.790 - 0.845)	0.825 (0.777 - 0.872)	0.812 (0.763 - 0.862)	0.815 (0.767 - 0.862)
AWW present	750	0.504 (0.468 - 0.540)	0.534 (0.472 - 0.596)	0.479 (0.416 - 0.543)	0.498 (0.437 - 0.559)
Number of children					
None	750	0.243 (0.212 - 0.273)	0.239 (0.186 - 0.292)	0.237 (0.183 - 0.292)	0.251 (0.198 - 0.304)
1 to 10	750	0.351 (0.316 - 0.385)	0.351 (0.291 - 0.410)	0.338 (0.277 - 0.398)	0.363 (0.304 - 0.422)
11 to 20	750	0.257 (0.226 - 0.289)	0.259 (0.204 - 0.314)	0.237 (0.183 - 0.292)	0.274 (0.219 - 0.329)
More than 20	750	0.149 (0.124 - 0.175)	0.151 (0.107 - 0.196)	0.188 (0.138 - 0.237)	0.112 (0.073 - 0.151)
Activity observed					
(Some) Learning	750	0.127 (0.103 - 0.151)	0.124 (0.083 - 0.164)	0.146 (0.101 - 0.191)	0.112 (0.073 - 0.151)
Indication of food	750	0.199 (0.170 - 0.227)	0.175 (0.128 - 0.223)	0.171 (0.123 - 0.219)	0.247 (0.194 - 0.300)
AWC building					
Concrete walls	750	0.561 (0.526 - 0.597)	0.574 (0.512 - 0.635)	0.575 (0.512 - 0.638)	0.537 (0.476 - 0.598)
Concrete roof	750	0.360 (0.326 - 0.394)	0.359 (0.299 - 0.418)	0.367 (0.305 - 0.428)	0.355 (0.297 - 0.414)
Iron roof	750	0.441 (0.406 - 0.477)	0.438 (0.376 - 0.500)	0.442 (0.378 - 0.505)	0.444 (0.383 - 0.505)
Lockable	750	0.477 (0.442 - 0.513)	0.474 (0.412 - 0.536)	0.442 (0.378 - 0.505)	0.514 (0.452 - 0.575)

Note: Shares of AWCs with the respective characteristics, in the full analysis sample (N = 750), control group, light treatment group, and intensive treatment group, with 95 percent confidence intervals.

Characteristics of AWWs were taken from the phone survey. About two-thirds of AWWs were between 25 and 44 years old and more than half had completed higher secondary education. About 70 percent and 80 percent, respectively, stated that she or her family owned a scooter and land. Around half of all AWWs believed that AWWs had the highest influence on the functionality of AWCs.

There are no systematic differences in characteristics between control and treatment groups, indicating that randomisation was successful in creating a balanced sample (see Table 3.1 and Table 3.2).

Table 3.2: Sample characteristics of AWWs

	N	Total	Control	Light Treatment	Intensive Treatment
Age in years	742				
<25		0.047	0.065	0.046	0.031
		(0.032 - 0.062)	(0.034 - 0.095)	(0.019 - 0.073)	(0.010 - 0.053)
25-34		0.334	0.290	0.360	0.353
		(0.300 - 0.368)	(0.233 - 0.347)	(0.299 - 0.421)	(0.294 - 0.412)
35-44		0.340	0.363	0.280	0.373
		(0.305 - 0.374)	(0.303 - 0.423)	(0.223 - 0.338)	(0.313 - 0.432)
45+		0.279	0.282	0.314	0.243
		(0.247 - 0.311)	(0.226 - 0.339)	(0.255 - 0.373)	(0.190 - 0.296)
Education	750				
Primary/secondary		0.171	0.175	0.175	0.162
		(0.144 - 0.198)	(0.128 - 0.223)	(0.127 - 0.223)	(0.117 - 0.207)
Higher secondary		0.544	0.546	0.529	0.556
		(0.508 - 0.580)	(0.484 - 0.608)	(0.466 - 0.593)	(0.495 - 0.617)
Graduation		0.285	0.279	0.296	0.282
		(0.253 - 0.318)	(0.223 - 0.335)	(0.238 - 0.354)	(0.227 - 0.337)
Owns scooter	749	0.692	0.701	0.720	0.656
		(0.658 - 0.725)	(0.644 - 0.758)	(0.662 - 0.777)	(0.598 - 0.715)
Owns land	726	0.802	0.799	0.770	0.833
		(0.773 - 0.831)	(0.749 - 0.850)	(0.715 - 0.824)	(0.787 - 0.880)
Own children at AWC	741	0.526	0.561	0.492	0.525
		(0.490 - 0.562)	(0.499 - 0.623)	(0.427 - 0.556)	(0.464 - 0.586)
Of same caste/religion	744	0.577	0.590	0.556	0.582
		(0.541 - 0.612)	(0.529 - 0.652)	(0.493 - 0.620)	(0.521 - 0.643)
Contact w AWWs every week	725	0.320	0.309	0.349	0.304
		(0.286 - 0.354)	(0.251 - 0.367)	(0.287 - 0.411)	(0.246 - 0.361)
Contact w supervisor every week	722	0.104	0.098	0.123	0.092
		(0.082 - 0.126)	(0.060 - 0.135)	(0.080 - 0.166)	(0.056 - 0.129)
Influence of AWW highest	735	0.486	0.500	0.476	0.480
		(0.449 - 0.522)	(0.437 - 0.563)	(0.412 - 0.541)	(0.418 - 0.542)
Mentions food as task	749	0.322	0.315	0.312	0.337
		(0.288 - 0.355)	(0.257 - 0.373)	(0.253 - 0.372)	(0.279 - 0.395)
Transport duration	750	13.465	14.139	12.721	13.502
		(12.449 - 14.482)	(11.878 - 16.401)	(11.212 - 14.230)	(12.098 - 14.906)

Note: Shares of AWWs with the respective characteristics, in the full analysis sample, control group, light treatment group, and intensive treatment group, with 95 percent confidence intervals. Number of observations vary by characteristic due to missing values.

3.4.3 Participation in and implementation of the intervention

Participation in the workshops was high among AWWs. 272 AWWs were invited to a short workshop and 275 AWWs to a longer workshop. Among those invited to a short workshop, 239 (87.9 percent) participated and 33 (12.1 percent) did not. Among those invited to a longer workshop, 236 (85.8 percent) participated in longer workshops as intended, 24 (8.6 percent) did not participate at all, and 15 (5.5 percent) participated in a workshop that had to be turned into a short workshop due to a disruption. In addition, nine AWWs, which are not part of the sample, were also invited and received the intervention as their AWC was in immediate proximity to a treated AWC. Furthermore, five non-invited AWWs, also not part of the sample, attended due to incorrect invitation by their supervisor or the workshop taking place in their AWC.

We compared whether invited but non-participating AWWs differed systematically from invited and participating AWWs (see Table A 3.1 in the appendix). Non-participating AWWs were less likely to be present at their AWC at the second unannounced visit by about 14 percentage points (p-value 0.0576). The probability of observing any indication of food was also lower by 19 percentage points (p-value 0.0014).

AWHs of all AWCs in the intensive treatment group were invited to a workshop, but their participation was lower compared to that of AWWs. Of those invited, 146 (53.1 percent) participated in the intended workshop, 120 (43.6 percent) did not participate in any workshop, and nine (3.3 percent) participated in the short workshop with AWWs. Among those not invited to any workshop, seven participated in a workshop intended for AWHs (one was assigned to the control group, four were assigned to light treatment, two were not included in the randomisation) and five participated in the workshop for AWWs (four were assigned to light treatment, one was not included in the randomisation). The latter happened, when AWHs came along with their AWW or as a replacement of the AWW if she was not able to participate in the workshop herself.

Lucky Iron Leaves were given to AWWs participating in the workshops if their AWC was assigned to either of the two treatment groups. In case an AWW did not attend but the AWH of the same AWC attended, the AWH received the Lucky Iron Leaves for the AWC. 240 AWCs in the light treatment group and 254 AWCs in the intensive treatment group received Lucky Iron Leaves.

The ingredients were delivered to all AWWs of the intensive treatment group that received the Lucky Iron Leaves at least 3 times before the closure of the AWCs. The reminder phone calls were received by 240 AWWs (94.5 percent of the AWCs that received the Lucky Iron Leaves). The remaining AWWs were either not reached or no valid phone number was available. In the reminder phone calls, all AWWs claimed that they used the Lucky Iron Leaves regularly. All except one AWW reported that they had received their delivery of ingredients. A large majority of 74 percent reported no difficulties encountered in the use of Lucky Iron Leaves.

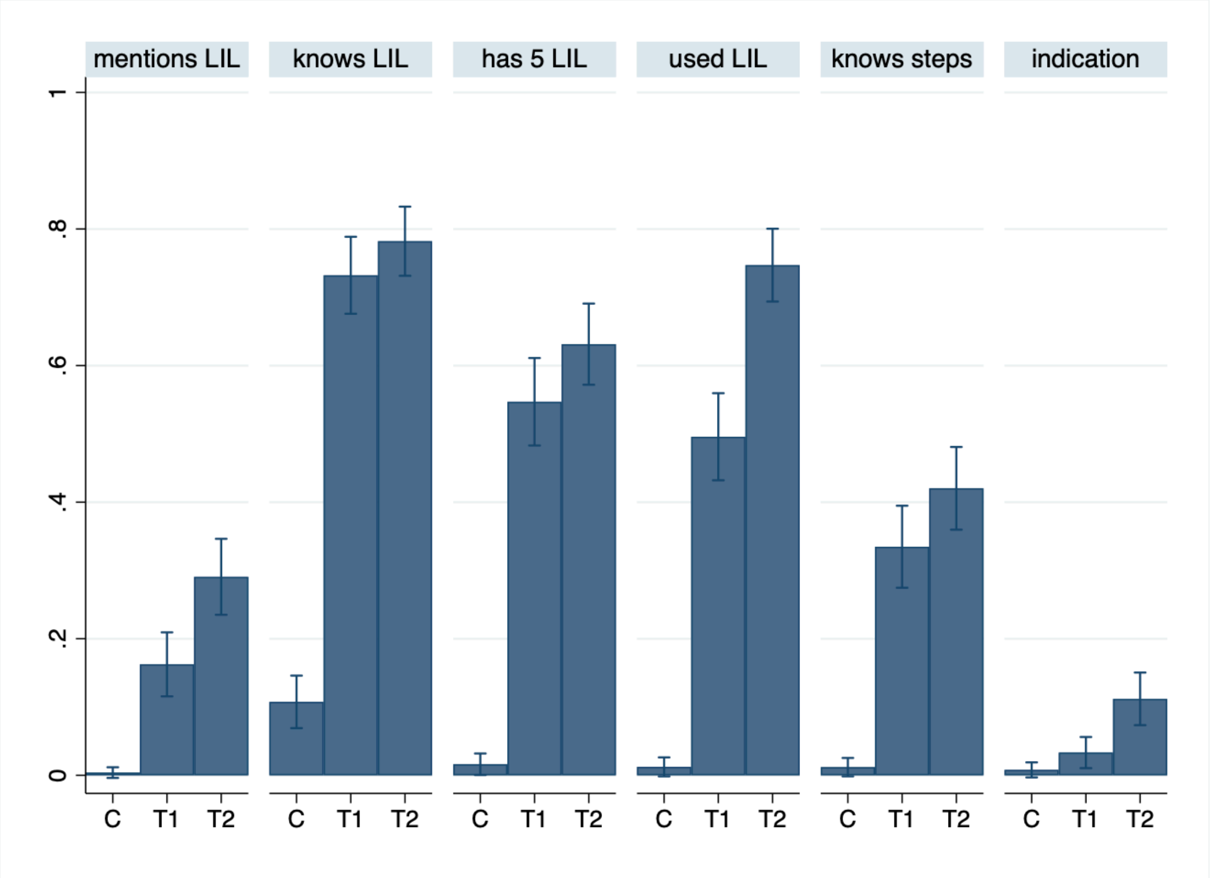
3.5 Results

Figure 3.1 descriptively presents results of the interventions by comparing shares of AWWs across treatment and control groups. The indicators shown include indicators that could potentially be affected by intentional misreporting. In the light and intensive treatment groups, 16.3 percent and 29.1 percent of AWWs mentioned the Lucky Iron Leaf when asked about iron supplementation methods during the phone interview, as compared to 0.4 percent in the control groups. When asked specifically whether they know the Lucky Iron Leaf, 73.2 percent and 78.2 percent in the light and intensive treatment groups and 10.8 percent in the control group answered that they did. It is very possible that AWWs in the control group talked to colleagues who benefited from the training and had therefore heard about it. However, the information does not seem to have spread a lot. 54.8 percent and 63.2 percent in the treatment groups reported that they still have all five Lucky Iron Leaves that were handed out to them while 49.6 percent and 74.7 percent reported to have regularly used them. The Lucky Iron Leaves and indications of their use were seen in only 3.3 percent of AWCs in the light treatment group and 11.2 percent in the intensive treatment group at the second observational visit.

Table 3.3 presents the results of the regression analysis for the three outcomes *mentions LIL*, *knows steps,* and *indication of LIL*. The share of AWWs mentioning the Lucky Iron Leaf without being prompted (*mentions LIL*) is 17.0 percentage points and 26.8 percentage points higher in the light and intensive treatment groups compared to the control group. The difference between the two treatment groups is statistically significant. The share of AWWs remembering the detailed steps of how to use the Lucky Iron Leaf (*knows steps*) are 28.1 and 34.1 percentage points higher in the treatment groups compared to the control group. However, the difference between the two treatment groups is not statistically significant for this outcome. The share of AWCs with an indication of the Lucky Iron Leaf being used (*indication of LIL*) is higher by 2.7 and 10.5 percentage points. The difference between the two treatment groups is statistically significant.

The results are very similar for models without control variables, with control variables only for administrative block, being a Mini AWC, and phone survey enumerator, and in the case of *indication of LIL* with additional control variables capturing the likelihood of food being served (see Table A 3.2).

Figure 3.1: Mean outcomes across treatment arms



Note: Shares of AWWs who (1) mention the Lucky Iron Leaf (LIL) without being prompted, (2) state to know the LIL (without or with being prompted), (3) report to have 5 LIL, (4) report to ever have used the LIL, (5) know the steps of using the LIL, and (6) share of AWCs where an indication of LIL use was observed, with 95 percent confidence intervals, analysis sample, N = 750.

Table 3.3: Estimated treatment effects

	(1) Mentions LIL	(2) Knows steps	(3) Indication
Light treatment	0.170***	0.281***	0.0269
	(0.0000)	(0.0000)	(0.0624)
Intensive treatment	0.268***	0.341***	0.105***
	(0.0000)	(0.0000)	(0.0000)
p-value	.0017	.1496	.001

Note: Estimation results of a linear probability model with robust standard errors, N = 673 (1) and (3) and 672 (2); controls for AWW's age and education (capacity for learning), contact with AWWs and supervisor (integration into system), children attending AWC, same caste/religion (relation to village), perceived influence of AWWs (willingness to learn, motivation), being Mini AWC, block, phone survey enumerator. p-values in parentheses. p-value of test of difference between coefficients. * p < 0.05, ** p < 0.01, *** p < 0.001

We observe deviations in the response pattern for one phone survey enumerator and therefore re-run the analysis for the sub-sample excluding her interviews (see Table A 3.3). The results are robust to excluding her interviews from the sample, and the same is true for all other enumerators.

3.6 Discussion

We find that introducing the Lucky Iron Leaves to AWWs via workshops was relatively successful in a sense that the method was understood and accepted. However, the observed indication of usage was low. While we do not have data on health outcomes, given the low observed usage and short time for potential usage before the closure, we do not believe that health benefits for children attending the AWCs were realised.

Our findings show that a substantial share of AWWs remembered the Lucky Iron Leaf and remembered the preparation process in sufficient detail to enable its correct use. Nevertheless, there is a considerable share that did not mention the Lucky Iron Leaf without being prompted and that did not remember all key steps. This is less surprising when considering that the survey took place 14 months after the workshops and AWWs could have used the Lucky Iron Leaves in their AWCs only for two months before their closure. Additionally, AWWs were involved as key actors in the public health response to the COVID-19 pandemic and had to perform multiple new tasks even during the closure of AWCs. The lower share of AWWs mentioning the Lucky Iron Leaf without being prompted compared to the share remembering the process in detail suggests that failure to recall might be a relevant factor.

We observed indications of usage of the Lucky Iron Leaf or actual usage only in a small fraction of AWCs. This outcome is only a proxy of actual usage as observational visits lasted for just about 10 min. Iron-enriched drinking water could have been prepared at a different time or location. Nevertheless, even self-reported usage in the light treatment group, a measure presumably prone to misreporting, is much lower than the share claiming to know the Lucky Iron Leaf when prompted, further indicating low actual use.

Even if iron-enriched water is prepared regularly, regular attendance of children and their their parents' acceptance of the water is required for health benefits to materialise. We observed that attendance overall was rather low. The anecdotal evidence regarding the acceptance of the iron-enriched water is mixed. Several AWWs mentioned in the reminder phone calls that children did not like the taste of the iron-enriched water. Enumerators of the second unannounced visit also reported that AWWs told them about children disliking the taste. Other AWWs stated that children were enthusiastic about the water and some children even said the water tasted like lemonade. Few reports during the reminder phone calls indicated that AWWs had to justify the use of the Lucky Iron Leaf towards the children's parents. Some parents requested the AWW not to give iron-enriched water to their children, claiming that it made them sick.

For all outcomes we observe lower shares in the light compared to the intensive treatment group with substantial and statistically significant differences for mentioning the Lucky Iron Leaf without being prompted as well as observed and self-reported usage. A higher share of AWWs mentioning the Lucky Iron Leaf without being prompted may be a reflection of higher usage but might also be an effect of more intense exposure due to the longer workshop and several reminders. We conclude that the light treatment with the shorter workshop is sufficient to establish an understanding of the method, but that successful implementation is more likely with the intensive treatment.

AWWs seem to be an appropriate target group for this intervention. AWWs have a higher educational background and a better understanding of iron deficiency anaemia compared with the general population. In our sample of AWWs, 83 percent completed upper secondary school, compared to 23 percent of the general female adult population in Madhepura district according to the NFHS-5 reports (International Institute for Population Sciences (IIPS), 2021). Likewise, the workshops showed that awareness of anaemia is nearly universal among participating AWWs. 99.1 percent (85.6 percent) of AWWs knew at least one (two) methods of iron supplementation in the survey. In contrast, only 4 percent of households in the study of Ebert et al. had heard of anaemia prior to their intervention (Ebert et al., 2020). In their study, even after the intervention, only 7 percent of the treatment group reported having heard of anaemia, which was a key element in the explanations introducing the Lucky Iron Leaf, indicating very little learning. A lack of information about underlying health issues is a common obstacle to take-up of preventive health care products (Dupas, 2011). We assume that the AWWs' characteristics facilitated their understanding of the workshop contents and the acceptance of the Lucky Iron Leaf as a valid method for iron supplementation.

Yet, the observed low functionality of many AWCs in Madhepura seems to impede regular usage of the Lucky Iron Leaf. Through the unannounced visits we observed that in a majority of AWCs, services such as pre-school education and food were not provided as intended. Looking at more functional AWCs, proxied by the number of attending children during the first unannounced visit, the likelihood of an indication of usage being observed increases by a factor of four and six for the light and intensive treatment, respectively. Better existing functionality therefore seems to be linked to a higher probability of using the Lucky Iron Leaf. Anecdotal evidence further suggests that regular use is impeded by an actual or perceived lack of resources. In the short workshops, many AWWs expressed a concern that not having lemons and cooking fuel will prevent them from using the Lucky Iron Leaf. During the second unannounced visit, enumerators were approached by AWWs with questions about lemons and sugar. During the reminder phone calls, AWWs also reported a lack of lemons and sugar or resources to buy these ingredients as well as a lack of cooking fuel. AWCs in the intensive treatment arm received lemons and sugar regularly. The anecdotal evidence suggests that these deliveries might be a major factor explaining the differences in observed and self-reported usage between the two treatment groups.

Costs of the intervention

A major advantage of the Lucky Iron Leaf as a method for iron supplementation in AWCs is its low cost and logistical requirements. The five Lucky Iron Leaves required to cater to 40 children attending one AWC can be used for up to five years. At the time of writing, a single unit was available for 1100 INR in the Indian market. However, the company producing Lucky Iron Leaves has several schemes supporting public health efforts in vulnerable population groups, potentially lowering the price for the ICDS. Production costs of the cast iron ingots were 3.30 USD a piece at the time of this study. In addition, there is a one-time cost for training AWWs (and AWHs). In our study, the costs for the workshops targeting 556 AWWs and 279 AWHs amounted to about 855,900 INR including the training of facilitators, salary, transportation, materials, and refreshments. Continuous expenses arise for lemons, sugar, and cooking fuel. To prepare water for 40 children, three small lemons and four tablespoons of sugar are required. During the study period, this amounted to 7 to 13 INR per AWC per day for lemons and sugar. Depending on the strategy for implementation, additional costs for delivery of ingredients as well as reminder phone calls can arise. Assuming a continuous use of the Lucky Iron Leaves for five years on 25 days per month, the costs per targeted AWC per year in our study amounted to about 3500 INR, excluding delivery and cooking fuel. If 40 children regularly consume the iron-enriched water, this corresponds to 88 INR per child per year.

While the costs of the intervention would be low per child for the ICDS, limited resources in the current state of the AWC system seem to be a major barrier. Running costs of using the Lucky Iron Leaves could be reduced in several ways. A major factor driving the costs are lemons. As lemon prices vary substantially over the year, costs could be reduced by providing iron supplementation with the Lucky Iron Leaf only during the season with low lemon prices. Moreover, the ICDS project of creating kitchen gardens for the AWCs could include a lemon tree. The sugar used in the iron-enriched water is for taste only and could potentially also be substituted with herbs from a kitchen garden. Ultimately, adjusting the lunch menu to include fruit acid from tomato, tamarind or lemon on several days a week would allow to incorporate the Lucky Iron Leaves in the preparation of food. This would avoid the expenses for fuel to boil the water and eliminate the additional effort of preparing iron-enriched drinking water.

Replicability of the intervention

The implementation of the workshops relied on the structure of the ICDS and the light treatment was designed similar to how the intervention would likely be implemented in the institutional framework of the ICDS. However, our workshops differed in two key aspects from other trainings for AWWs conducted by the ICDS. Firstly, the workshop facilitators in our intervention were women, mainly from Madhepura City and young, most under 25 years old, while the ICDS trainings we observed were conducted by men. Second, our workshops were conducted in small groups, centred on participants, and were highly interactive with practice elements for each participant. The workshops had a friendly rather than teacherly atmosphere. Facilitators were perceived as approachable and usually reported chatting with AWWs. ICDS trainings that we observed seemed to be more formal, for large groups of AWWs, for example an entire block, and much less interactive. It is unclear whether these differences influence the impact of the intervention and its replicability. While male and more senior trainers might be perceived as a stronger authority and might be taken more seriously in this highly patriarchal context, they could be less suitable for instilling a sense of empowerment in the AWWs compared to female trainers. Interactive methods have been shown to lead to greater learning, potentially increasing the share of AWWs remembering exactly how to use the Lucky Iron Leaves. On the other hand, these methods might have also raised scepticism as AWWs would not be used to them from other trainings. Reports from workshop facilitators indicated that in some groups, AWWs and AWHs were actually enthusiastic about the activities and games and participated actively, other groups were more reluctant and wanted to rush through the activities.

Further aspects could influence the replicability of the intervention when implemented through the ICDS. As a training conducted by ICDS, the use of Lucky Iron Leaves might be perceived as an institutional order, potentially leading to greater compliance and use of the method compared to a workshop conducted by researchers. However, it is also known that many main tasks of AWCs, such as providing daily meals, are not implemented as intended (Fraker et al., 2013). Due to imperfect monitoring and leakage of funds, an additional task could similarly be lacking in implementation.

Limitations

This study suffers from several limitations. The biggest limitation was caused by the COVID-19 pandemic and resulting long-term closure of AWCs. It was not possible to measure haemoglobin levels of children attending AWCs as the health outcome of interest and the study was limited to the analysis of intermediate stages in the effect chain. However, we argue that the outcomes of remembering how to use the new method and observed indications of usage capture intermediate steps relevant for an effect on the health outcome.

Secondly, it is not possible to differentiate which component of the intensive treatment led to its greater effectiveness compared to the light treatment. The first observational data collection revealed overall low functionality, leading to a smaller sample of at least partially functional

AWCs than originally expected. As a result, the number of treatment arms was smaller than initially planned to retain sufficient statistical power and the different components of the intensive treatment were combined into one. All components of the intensive treatment that differ from the light treatment, including empowerment aspects in the workshops, training for AWHs, reminder phone calls, as well as the delivery of ingredients, have to be interpreted jointly.

Further limitations also caused by the COVID-19 pandemic relate to the accuracy of our outcomes. The phone survey took place fourteen months after the workshops and twelve months after the closure of AWCs, impeding recollection. Potentially this was amplified by the long closure period of AWCs during which no food or water were prepared and thus the Lucky Iron Leaves could not be used. However, as AWWs were involved in the response measures to the pandemic, an earlier date for the survey was not feasible. Further planned unannounced visits would have improved the measures but could not be conducted.

Another limitation pertains to possible distortions in answer patterns during the phone survey. Several workshop facilitators also acted as enumerators during the phone survey. Due to COVID-19 restrictions, training and monitoring of the phone survey had to be conducted remotely. As we needed a team of trusted, experienced enumerators who were well informed about the intervention to ensure correct posing of questions and coding of answers, we opted for this setup. As a result, some enumerators could potentially have recognised individual AWWs, therefore knowing their treatment status. Similarly, AWWs could potentially have recognised the enumerator over the phone and adjust answer patterns. For example, AWWs could be reminded about the Lucky Iron Leaf by being called by the person who also conducted their workshop and could then feel obliged to answer positively about the Lucky Iron Leaf. To address this, the research team assigned phone calls randomly, which led to a minimal overlap between enumeration and workshop facilitation. Nevertheless, there were single incidences where enumerators reported that they recognised the AWW and exchanged with her further about her experience with the Lucky Iron Leaf.

3.7 Conclusion

We implemented an intervention introducing a low-cost method of iron supplementation, the Lucky Iron Leaf, to AWCs in Madhepura district in Bihar, India. We evaluated its effect on intermediate outcomes, knowledge about the method and observed indications of usage, because the intended health outcomes among children could not be measured due to the COVID-19 pandemic. Our findings indicate that AWWs, due to their high education and awareness of health problems in general and regarding anemia specifically, are a suitable target group to learn about this method and to provide iron supplementation to beneficiaries of the AWC services even in remote areas. However, observed usage of the Lucky Iron Leaves was very low. We identify the overall low functionality of AWCs and lack of resources as major barriers to the implementation of the iron supplementation method. Institutional support and structural improvements in the AWC system would be necessary to successfully establish iron supplementation in the daily routine of AWCs.

3.8 Appendix

3.8.1 Tables

Table A 3.1: Differences between compliers and non-compliers

	Assigned	Participated	Not participated	p-value
N	499	447	52	
Mini AWC	0.090	0.089	0.096	0.8741
AWC open	0.814	0.819	0.769	0.3861
AWW present	0.489	0.503	0.365	0.0598
Number of children				
None	0.244	0.239	0.288	0.4367
1 to 10	0.351	0.345	0.404	0.3972
11 to 20	0.257	0.262	0.212	0.4337
More than 20	0.148	0.154	0.096	0.2645
Activity observed				
(Some) Learning	0.128	0.128	0.135	0.8851
Indication of food	0.210	0.230	0.038	0.0013
AWC building				
Concrete walls	0.555	0.550	0.596	0.5301
Concrete roof	0.361	0.356	0.404	0.4948
Iron roof	0.443	0.436	0.500	0.3820
Lockable	0.479	0.474	0.519	0.5400
AWW age in years				
<25	0.038	0.038	0.038	1.0000
25 to 34	0.356	0.364	0.288	0.2813
35 to 44	0.328	0.333	0.288	0.5225
45+	0.277	0.265	0.385	0.0679
AWW education				
Primary/secondary	0.168	0.170	0.154	0.7685
Higher secondary	0.543	0.544	0.538	0.9438
Graduation	0.289	0.286	0.308	0.7485
Owns scooter	0.687	0.688	0.673	0.8227
Owns land	0.803	0.805	0.784	0.7248
Own children at AWC	0.509	0.515	0.462	0.4694
Of same caste/religion	0.570	0.569	0.577	0.9116
Contact w AWWs every week	0.326	0.326	0.320	0.9280
Contact w supervisor every week	0.107	0.105	0.122	0.7110
Influence of AWW highest	0.478	0.466	0.588	0.0975
Mentions food as task	0.325	0.320	0.373	0.4481
Transport duration	13.126	12.776	16.135	0.0492

Note: AWWs and AWCs with the respective characteristics, in the sample of AWWs assigned to any treatment group, the sample who were assigned and participated in any workshop, the sample who were assigned but did not participate in any workshop; p-value of test of difference between share of those who participated and those who did not participate.

Table A 3.2: Estimated treatment effects - robustness

	(1) Mentions LIL	(2) Mentions LIL	(3) Mentions LIL	(4) Knows steps	(5) Knows steps	(6) Knows steps	(7) Indication	(8) Indication	(9) Indication	(10) Indication
Light treatment	0.159***	0.160***	0.170***	0.264***	0.263***	0.281***	0.0254*	0.0245	0.0269	0.0262
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0496)	(0.0593)	(0.0624)	(0.0800)
Intensive treatment	0.287***	0.274***	0.268***	0.354***	0.348***	0.341***	0.104***	0.103***	0.105***	0.0944***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Fixed effects	no	yes	yes	no	yes	yes	no	yes	yes	yes
Controls	no	no	yes	no	no	yes	no	no	yes	yes+
p-value	.0006	.0001	.0017	.0312	.0333	.1496	.0006	.0005	.0001	.0033
Observations	749	749	673	747	747	672	750	750	673	654

Note: : Estimation results of a linear probability model with robust standard errors; models (1), (4), (7) include no control variables; models (2), (5), (8) include fixed effects for blocks and phone survey enumerators and being mini AWC; models (3), (6), (9), additionally control for AWW's age and education (capacity for learning), contact with AWWs and supervisor (integration into system), children attending AWC, same caste/religion (relation to village), perceived influence of AWWs (willingness to learn, motivation); model (10) additionally controls for whether any indication of food was observed, whether 20 or more children were present, whether the AWC building had any concrete walls, whether the AWC building was lockable, and observational survey day. p-values in parentheses. p-value of test of difference between coefficients. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A 3.3: Estimated treatment effects - robustness: excluding enumerator

	(1) mentions LIL	(2) knows steps	(3) indication
Light treatment	0.0818***	0.226***	0.0249
	(0.0003)	(0.0000)	(0.0680)
Intensive treatment	0.169***	0.268***	0.114***
	(0.0000)	(0.0000)	(0.0000)
p-value	0.0061	0.3304	0.0004
Observations	585	584	585

Note: Estimation results of a linear probability model with robust standard errors, sample excludes observations made by one enumerator due to quality concerns; controls for AWW's age and education (capacity for learning), contact with AWWs and supervisor (integration into system), children attending AWC, same caste/religion (relation to village), perceived influence of AWWs (willingness to learn, motivation), being Mini AWC, block, phone survey enumerator. p-values in parentheses. p-value of test of difference between coefficients. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.8.2 Figures

Figure A 3.1: Lucky Iron Leaf

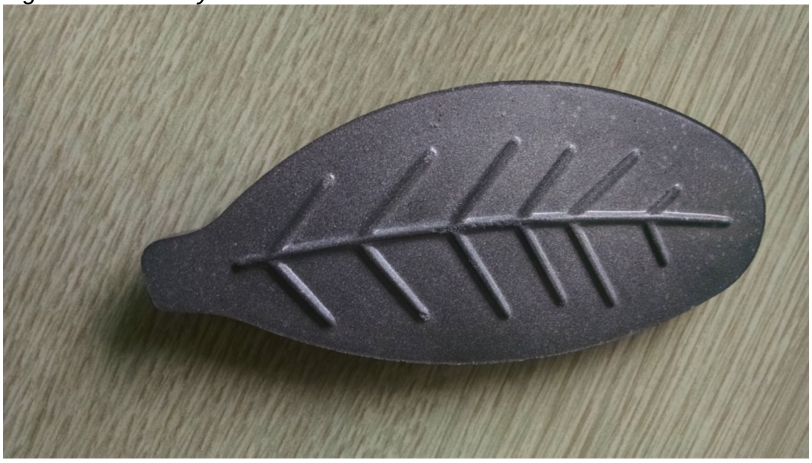
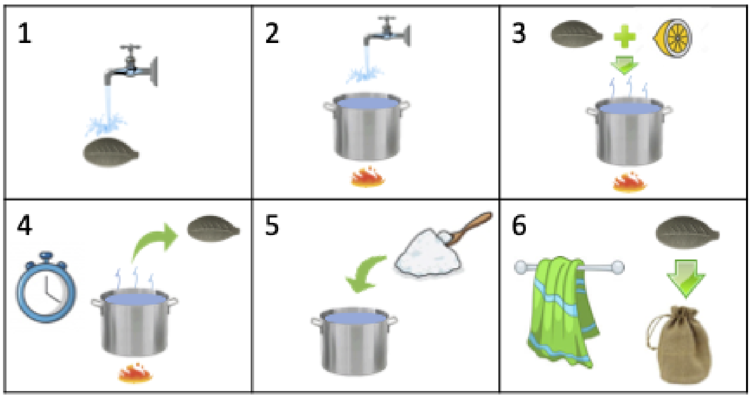
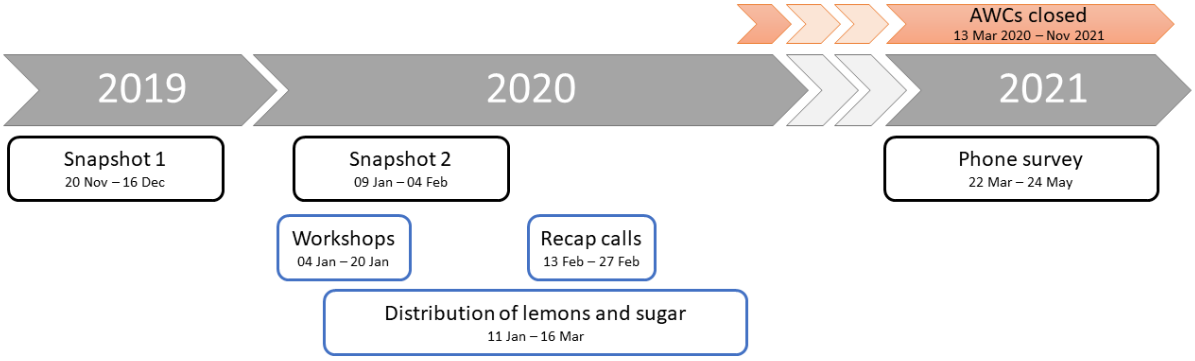


Figure A 3.2: Manual



👫	🚰	🍋	⏱	🧂
0 - 20	2 l	2	10 min	2
21 - 30	3 l	2.5	16 min	3
31 - 40	4 l	3	22 min	4

Figure A 3.3: Project timeline



Chapter 4

Sensitivity bias in questions on sexual behaviour, sexual health, drug consumption, and intimate partner violence in an adult population in urban Tanzania

with: Sebastian Vollmer and Till Bärnighausen

Abstract

Risky health behaviour is stigmatized in many societies, potentially causing misreporting in surveys capturing health behaviour. Such sensitivity bias affects conclusions derived from these surveys, potentially impacting policy making. It is therefore important to identify the presence of sensitivity bias in survey questions to allow for corrections in estimations of prevalence. In this study, we measure sensitivity bias in ten sensitive questions on sexual behaviour, sexual health, drug consumption, and intimate partner violence in an adult population in Dar es Salaam, Tanzania. The sensitivity bias is estimated by comparing responses to list experiments and corresponding direct questions. We measure the presence and magnitude of sensitivity bias in the total sample and analyse heterogeneities by demographic characteristics. We find sensitivity bias in some questions, but not others, and heterogeneities by demographic characteristics, especially between men and women.

4.1 Introduction

Risky health behaviour, including drug use, alcohol consumption, certain sexual activities, is stigmatized in many societies (Airhihenbuwa et al., 2014). While social stigma may not directly prevent such behaviour, it may prevent individuals from admitting that they do engage in it. This leads to concerns of misreporting in surveys capturing health behaviour. Respondents may decide to withhold the truthful answer but instead give a response that is perceived as socially preferred in order to avoid negative consequences such as embarrassment or even physical threats. This misreporting has been defined as social desirability bias or sensitivity bias (Blair et al., 2020). If sensitivity bias severely affects surveys on self-reported health behaviour and health, conclusions derived from these surveys are affected. The prevalence of risky health behaviour may be underestimated. Moreover, when it is not possible to correctly identify who engages in risky health behaviour, the identification of determinants of such behaviour is impeded. Biases in prevalence estimates and incorrectly identified determinants negatively influence policy making.

List experiments, also called item count technique or unmatched count technique, are an increasingly popular method to reduce sensitivity bias in survey questions (Lupu & Michelitch, 2018). In a list experiment, survey respondents are split into a treatment and a control group. The control group receives a list of *J* statements, the control statements, while the treatment group receives the same list of *J* statements and an additional *J+1th* statement about a sensitive topic. Respondents are asked to report how many of the statements are true for them. As this method makes it impossible to infer the individual response to the sensitive item, an increased sense of privacy is created. Respondents are therefore more likely to answer the sensitive item, e.g., whether or not they engage in a risky behaviour, truthfully. The difference in the mean response between control and treatment group gives an estimate of the prevalence of the behaviour in the surveyed population. The sensitivity bias can be estimated by comparing the prevalence according to the list experiment with the prevalence according to the responses to a direct question regarding the same sensitive item.

List experiments have been used to estimate prevalence of behaviour or opinions and to reveal sensitivity bias in a wide range of topics, including racism (Gilens et al., 1998; Kuklinski et al., 1997), vote preferences (Gonzalez-Ocantos et al., 2012; Holbrook & Krosnick, 2010), illegal migration (McKenzie & Siegel, 2013), and opinions on same sex marriage (Lax et al., 2016). Health topics studied with list experiments include condom use (Jamison et al., 2013; LaBrie & Earleywine, 2000; Lépine, Treibich, Ndour, et al., 2020), intimate partner violence (Agüero & Frisancho, 2017; Cullen, 2020; Gilligan et al., 2021; Peterman et al., 2018), and abortion

(Moseson et al., 2017). Evidence of sensitivity bias in health topics is very mixed and strongly depends on the setting and specific topic (Blair et al., 2020). This study contributes to this evidence by estimating sensitivity bias on a set of ten outcomes related to health in the context of urban Tanzania. Including multiple outcomes allows us to compare magnitudes of sensitivity bias for different topics. Moreover, we analyse how sensitivity bias varies with demographic characteristics of respondents.

We use data from a face-to-face survey to measure sensitivity bias in a range of health behaviours in an adult population in Dar es Salaam, Tanzania. The outcomes include alcohol consumption, drug use, experience of intimate partner violence, sexual behaviour, and sexual health. According to the social reference theory developed by Blair and colleagues, sensitivity bias is a concern if the respondent (1) has a social referent in mind when considering his or her response, (2) perceives that the social referent can infer the response to the sensitive question, (3) has a perception about what response the social referent prefers, and (4) perceives that not responding as preferred by the social referent would entail costs (Blair et al., 2020). Given the survey design and topics, the four proposed conditions for sensitivity bias to be present seem to be fulfilled. The social referent could be the respondent him/herself or the data collector. The response to the sensitive question can be inferred directly due to the format of the face-to-face interview when asked directly. As the data collector comes from the same cultural context, the respondent is likely to perceive that he or she has a similar understanding of the socially preferred response. Costs of giving a socially undesirable response are likely to be social, such as embarrassment, but could include physical for some questions. Same-sex relations, one of the topics covered, are a criminal offence and punishable by life imprisonment in Tanzania. Sensitivity bias is therefore highly likely to be a concern in this setting for the questions analysed here.

4.2 Methods

4.2.1 Study Population and Setting

The data comes from a study that was nested within the Dar es Salaam Urban Cohort Study (DUCS) in Dar es Salaam, Tanzania. DUCS is a Health and Demographic Surveillance System (HDSS) that aims to collect information on the health status of an urban population to explain health disparities due to socioeconomic, urban-living, and environmental influences on health and health behaviour (Leyna et al., 2017). The HDSS surveys all residents, totalling 21,000 households, in seven administrative streets in Ilala region in Dar es Salaam biannually (Leyna et al., 2017). For a nested study, 4,000 men and women aged 40 years and above were randomly selected from those enrolled in the HDSS. The selected individuals were visited for home-based face-to-face interviews from June 2017 to June 2018. Participation in the study was voluntary and written informed consent was sought from all participants before conducting the survey.

4.2.2 Study Design

Data collectors used a computer-assisted personal interviewing (CAPI) system to conduct interviews in Swahili. Respondents were asked questions on sociodemographic and clinical factors, as well as physical and mental functions. The questionnaire also included a set of nine list experiments capturing ten sensitive items on alcohol consumption, intravenous (IV) drug use, experience of intimate partner violence, sexual behaviour and health. While respondents received an explanation of how to answer these list questions, their purpose of revealing sensitive answers was not explained. Later in the questionnaire, the respondents received the same sensitive items as direct questions. The tablets used for the interviews randomized each respondent to the control or treatment group of the list experiments. Randomization was done separately for each list experiment. Respondents in the control group received a list of four control items while respondents in the treatment group received a list of five items including the same four control items plus the respective sensitive item. Respondents were then asked to report the total number of items that were true for them, without specifying which item they considered as true. Data collectors took time to explain the concept of providing the number of "yes" responses instead of answering the individual items on the list, but respondents were not provided with marbles or other objects to help counting. For one outcome, alcohol consumption, respondents were randomized into either the control group or one of two treatment groups. All three groups received the same four control items while the treatment groups received one of two additional items capturing the sensitive question. The phrasing of control and sensitive items is shown in Table A 4.1. The selection of control items can influence respondents' trust in the method. To avoid that the sensitive item stands out from the other items in the list and raises suspicions about the purpose of the survey, it is recommended to select control items on topics that are related to the sensitive item (Aronow et al., 2015; Kuklinski et al., 1997). In this study, each list was chosen to have a range of items that appeared to be in line with the actual survey content, including basic questions about daily life, family, social and health behaviours, physical ability, health history, and sensitive information such as sexual history and drug use. Following the recommendation by Glynn (Glynn, 2013), control items were chosen to be negatively correlated with each other. Direct questions concerning the sensitive items appeared later in the survey and were posed to all participants. Assuming that treatment and control groups similarly agree or disagree to the control items, the difference in the mean score, the mean response to the list question, between the treatment group and the control group represents the proportion of the participants who consider the sensitive item to be true for them. This difference therefore provides an estimate of the self-reported prevalence of the behaviour captured by the sensitive item (Blair & Imai, 2012; Imai, 2011), in the following referred to as list estimate. Comparing the list estimate with the prevalence reported in the direct question, the direct estimate, is defined as the revealed sensitivity bias for this question (Blair & Imai, 2012).

4.2.3 Outcomes

We estimate sensitivity bias in ten outcomes capturing (A) alcohol and drug consumption, (B) intimate partner violence, and (C) sexual behaviour and (D) sexual health. Alcohol and drug consumption were measured by whether or not the respondent (1) drank alcohol in the past 30 days, (2) drank six or more alcoholic drinks in the past 30 days, and (3) ever used IV drugs. Intimate partner violence was measured by whether or not the respondent ever experienced (1) physical or (2) sexual violence committed by their partner. Sexual behaviour was measured by whether or not the respondent (1) ever had sex with a person of the same gender, (2) ever paid someone in exchange for sex (males) or has ever been paid for sex (females), and (3) had sex with two or more different people in the past 12 months. Sexual health was measured by whether or not the respondent (4) had a sexually transmitted infection (STI) in the past 12 months and (5) ever tested positive HIV.

4.2.4 Statistical Analysis

For the analysis of the list experiments, observations with missing values in either the list questions or the respective direct questions from the sample need to be dropped. We create a data set of non-missing observations for each outcome separately, leading to a different sample size for each outcome (see Table A 4.2).

First, we evaluate the design of the list experiments. We address the three standard assumptions of list experiments, treatment ignorability, no design effects, no liars (Imai, 2011), related design effects on the direct question, as well as floor and ceiling effects (Blair & Imai, 2012). We also conduct the joint test of the three standard assumptions and an assumption of monotonicity suggested by Aronow and colleagues (Aronow et al., 2015). Additionally, we test whether non-response to list questions or direct questions may influence the analysis of sensitivity bias. We define non-response as the response of either “don’t know” or “refuse to answer”. We test whether rates of non-response differ between direct and indirect question and between control and treatment groups using t-tests.

We then estimate the revealed sensitivity bias for each outcome using the approach proposed by Blair and Imai (Blair & Imai, 2012). For each outcome, we proceed as follows. We use the *ictreg* function of the *List* package to run a linear regression on the list question without covariates (Blair & Imai, 2010). A binary logistical regression is used for the direct question. Without covariates, the former basically results in the standard difference-in-means estimator and the latter gives the proportion answering "Yes" to the direct question. We then generate predicted probabilities using the *predict* function of the *List* package in R. The predicted probability of the outcome being true, i.e. the predicted prevalence of the outcome, using the direct question is referred to as the direct estimate. This estimate is hypothesised to be contaminated by sensitivity bias. The predicted prevalence using the list question is referred to as the list estimate and is hypothesised not to be affected by sensitivity bias if the experiment was successful. We then calculate the difference between the two estimates. This difference measures the revealed sensitivity bias.

For a heterogeneity analysis, this approach is repeated while adding a set of covariates covering basic demographic characteristics. The covariates include gender (male, female), marital status (married, not married), religious affiliation (muslim, christian), age (40 to 49 years, 50 to 59 years, 60 years and above), education (none, primary, secondary or higher), and wealth quartiles. For each outcome, we run regressions controlling for each covariate separately. The sensitivity bias is then estimated for each subgroup of the sample. The sample sizes in the heterogeneity analysis are smaller compared to the previous analysis as observations with missing information in any of the covariates are dropped from the samples for each outcome.

Stata version 16.0 was used for data preparation and analysis of non-response, while R 4.2.0 was used for analysis sensitivity bias and related assumptions.

4.2.5 Ethical considerations

Ethical approval for the study was received from the Institutional Review Boards of Muhimbili University of health and Allied Sciences, Tanzania (2015-04-22/AEC/Vol.IX/82) and Harvard T.H. Chan School of Public health, USA (14-4282). Written informed consent was obtained from all respondents before the interview.

4.3 Results

4.3.1 Respondent characteristics

A total of 2,270 individuals participated in the survey, of which a majority (67.8%) was female (see Table A 4.3 for sample characteristics). Almost half of the respondents were aged between 40 and 49 years (48.0%), followed by 26.9% in the age bracket of 50 to 59 years. Most of the respondents (61.4%) had primary education and more than two-thirds were currently married (70.3%). The sample was split relatively evenly among individuals identifying as Muslim (53.9%) and Christian (45.6%).

4.3.2 Testing assumptions

Treatment ignorability requires that treatment assignment is independent of potential outcomes in the list questions and direct questions. To address this, respondents were randomly assigned into the treatment group for each experiment with a probability of 50%. One exception was the list experiment regarding alcohol consumption. As this list experiment contained two sensitive items, the sample was split in thirds. One third received the list of four control items, one third received the additional sensitive item on the consumption of any alcohol in the past month (*drank alcohol*), and the last third received the additional sensitive item on the consumption of six or more alcoholic drinks (*drank 6+ alcoholic drinks*). The control group for these two outcomes is therefore the same. The number of individuals in treatment and control groups for each outcome can be seen in Table A 4.2. We tested whether the groups were balanced regarding demographic characteristics using t-tests. We could not identify systematic imbalances in demographic characteristics across treatment and control groups for any outcome.

The assumption of no design effects requires that the individual's response to the control items in the list is not affected by the presence of the sensitive item (Imai, 2011). We use the *List* package in R by Blair and Imai (Blair & Imai, 2012) to test whether the mean for support for the control items is the same on average across treatment and control groups. As the response to control items alone cannot be directly observed in the treatment group, the approach is to test whether adding the sensitive item to the list increases the mean response in the treatment group but not by more than 1. The null hypothesis of no design effects was not rejected for any outcome (see Table A 4.4, column 3).

The no liars assumption stipulates that respondents answer the sensitive item in the list experiment truthfully (Imai, 2011). Those individuals for whom the sensitive item is true are assumed to include this sensitive item in the item count. Answers to the control items do not have to be truthful as long as they are on average the same across treatment and control groups. While we cannot conclusively test this assumption, we can exclude obvious violations. Answering “yes” to the direct question indicates the sensitive response. Among respondents in the treatment group who give the sensitive response in the direct question, the response to the list question should not be zero. While the no liars assumption does not directly require answers to the direct question to be truthful, we assume that those who answer “yes” to the direct question answer truthfully as there would not be any social pressure to give the socially undesirable answer. We therefore drop all observations that violate the no liars assumption in this manner from our analysis of sensitivity bias. Across the ten outcomes, between 0 and 15 individuals in the treatment groups reported “yes” to the direct question but answered 0 to the list question (see Table A 4.4, column 2). These observations were dropped from the analysis of sensitivity bias.

Aronow and colleagues propose a joint test of the three standard assumptions and an additional assumption of monotonicity (Aronow et al., 2015). Monotonicity requires that respondents do not claim to engage in the risky behaviour if they actually do not do so, meaning there are no false confessions. Similar to the approach for identifying violations of the no liars assumption, this test is based on the idea that respondents answering “yes” to the direct question should also answer “yes” to the sensitive item. Assuming that control and treatment groups answer control items on average the same, the difference between the two groups should be exactly 1 for those admitting the behaviour when asked directly. This difference comes from the treatment group including the sensitive item in their answer to the list question which is not included in the list for the control group. We test this using a simple t-test for the subgroup of respondents affirming in the direct question. The null hypothesis of the difference being equal to 1 is rejected for all outcomes (see Table A 4.4, column 5). This raises concerns about the validity of the list experiment as it indicates that not all assumptions are fulfilled. However, monotonicity is not directly required for the estimator used here to be unbiased.

Next, we test a second type of design effect. The response to the direct question should not be affected by assignment to the treatment group in the list experiment. Respondents should not be influenced in their response to the direct question by having seen the sensitive item in the list question. While this effect is not part of the standard assumptions, a violation could raise concerns regarding the design of the list experiment. We use chi-squared tests to assess whether the mean response to the direct question differs between treatment and control groups. The null hypothesis of no difference was not rejected for any outcome (see Table A 4.4, column 4).

Floor and ceiling effects are a further concern that could invalidate the experiment as they indirectly reveal the respondents' response to the sensitive item, removing the cover of anonymity provided by the method (Blair & Imai, 2012). Ceiling effects occur if a respondent wants to answer "yes" to all items in the list, including the sensitive item. This would reveal the response to the sensitive item. In order to hide the response, the respondent might answer with a lower item count (4 instead of 5), which would violate the no liars assumption. Floor effects occur if the respondent wants to answer "yes" only to the sensitive item and feels that control items all trigger a response "no", therefore revealing the response to the sensitive item. To avoid this, he or she might be pushed to answer with an item count of zero, again violating the no liars assumption. Both effects would lead to an underestimate of the true prevalence. Following the recommendation by Glynn, control items were selected to be negatively correlated, reducing the risk of floor and ceiling effects (Glynn, 2013). We also check the answer distribution to the list questions and whether the mean response is close to 2 among the control group to assess the probability of floor and ceiling effects (see Table A 4.5). The share of respondents answering with an item count of 4 is very low, above 1% only for the two outcomes on alcohol consumption. We therefore do not believe that ceiling effects are a concern. The share of respondents answering with an item count of 0, however, is higher for several outcomes, between 1.4% and 8.4%. One notable exception is found for the outcome *sex with same gender*, where the share is 20.6%. For this outcome, a floor effect could be a concern. This is also the outcome with the lowest mean, at 1.09. Means for all other outcomes are larger, between 1.41 and 1.83.

4.3.3 Non-response patterns

For the analysis of non-response patterns, we use the entire sample of 2,270 individuals. Between 21 and 35 observations are missing in the list questions (see Table A 4.6). In most cases, these observations contain missing values for all list questions, meaning that these respondents did not answer the entire survey module. For the direct questions, between 16 and 45 values are missing with the exception of the outcome *HIV*. This question has 107 missing values. For most outcomes, the rate of non-response is not statistically different between the list questions and the direct questions. The outcome *HIV* is the exception with significantly more missing values in the direct question compared to the list question (p-value < 0.0001). The outcome *STI* also has a higher rate of non-response in the direct question, with a p-value just above the 5-percent threshold (p-value = 0.0555).

A larger concern is whether there is any difference in non-response across treatment groups, as this may bias the estimation of sensitivity bias. In the list questions, we cannot find any difference in non-response across treatment and control groups (Table A 4.7). For one of the direct questions, *Drank 6+ alcoholic drinks,* the rate of non-response seems to be lower in the treatment group (p-value = 0.0143) (Table A 4.8). For the outcome *HIV*, non-response seems higher in the treatment group, but the test just misses statistical significance (p-value = 0.0587).

Overall, non-response does not seem to pose a threat to the estimation of sensitivity bias in this context.

4.3.4 Sensitivity bias

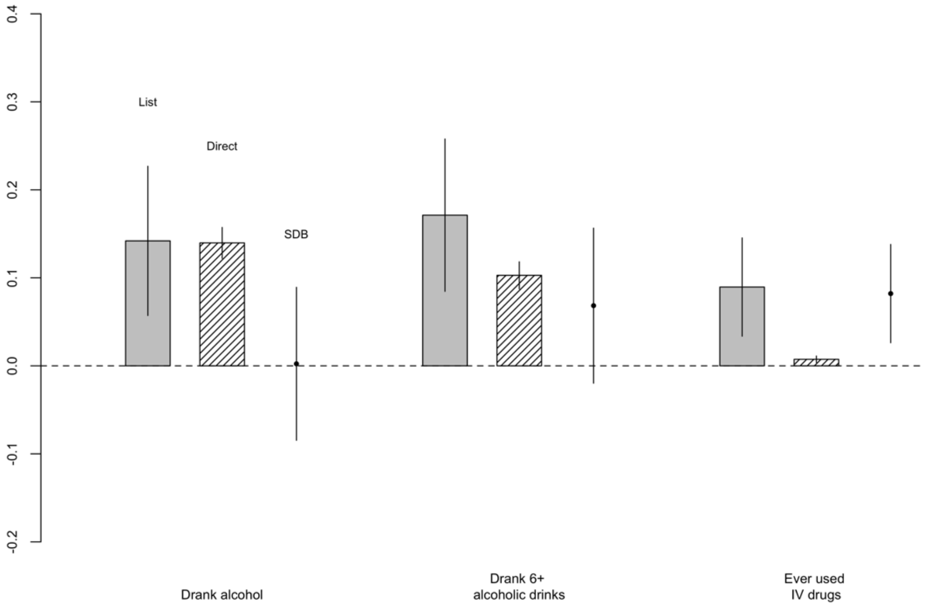
Sensitivity bias is measured by the difference between the list estimate and the direct estimate. We present the revealed sensitivity bias for all outcomes in Table 4.1 and Figure 4.1 to Figure 4.4. While the direct estimates have rather narrow confidence intervals for all outcomes, the list estimates are less precise and as a result, the estimates of the sensitivity bias also have rather wide confidence intervals. This is due to the fact that only half of respondents received the sensitive item in the list questions and control items add additional noise.

Table 4.1: Total sensitivity bias

	N	**Prevalence in direct question**	**Direct estimate**	**List estimate**	**Sensitivity bias**
Drank alcohol	1491	0.1395	0.1397	0.1420	0.0023
			(0.0090)	(0.0433)	(0.0443)
Drank 6+ alcoholic drinks	1481	0.1026	0.1028	0.1712	0.0684
			(0.0079)	(0.0441)	(0.0449)
Ever used drugs	2217	0.0072	0.0075	0.0895	0.0821
			(0.0019)	(0.0285)	(0.0285)
Physical IPV	2205	0.2032	0.2032	0.2826	0.0793
			(0.0087)	(0.0319)	(0.0328)
Sexual IPV	2210	0.0321	0.0324	0.1718	0.1394
			(0.0038)	(0.0294)	(0.0296)
Sex with same gender	2214	0.0072	0.0075	0.0584	0.0510
			(0.0019)	(0.0320)	(0.0321)
Transactional sex	2201	0.1908	0.1911	0.1084	-0.0828
			(0.0084)	(0.0327)	(0.0336)
Sex with multiple partners	2214	0.0393	0.0395	0.1222	0.0827
			(0.0042)	(0.0274)	(0.0277)
STI	2213	0.0181	0.0183	0.1766	0.1584
			(0.0029)	(0.0285)	(0.0286)
HIV	2146	0.0527	0.0529	0.1336	0.0807
			(0.0048)	(0.0322)	(0.0325)

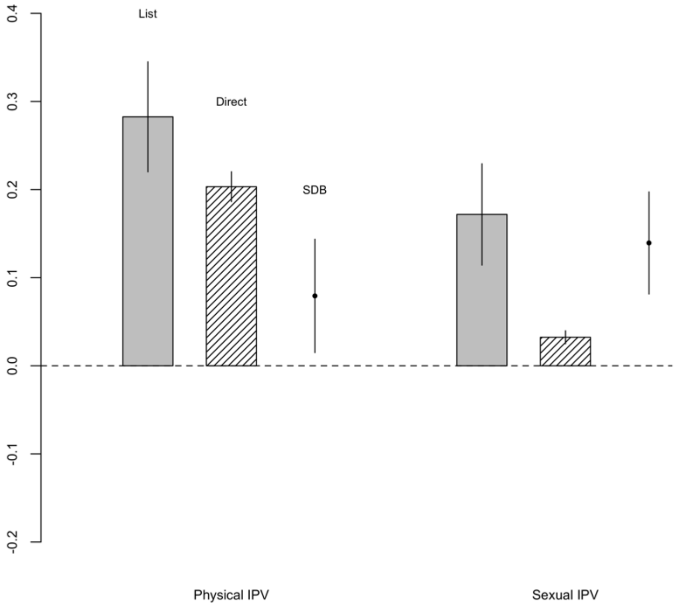
Note: Standard errors in parentheses. STI = sexually transmitted infections; HIV = human immunodeficiency virus; IPV = intimate partner violence.

Figure 4.1: Sensitivity bias in alcohol and drug consumption



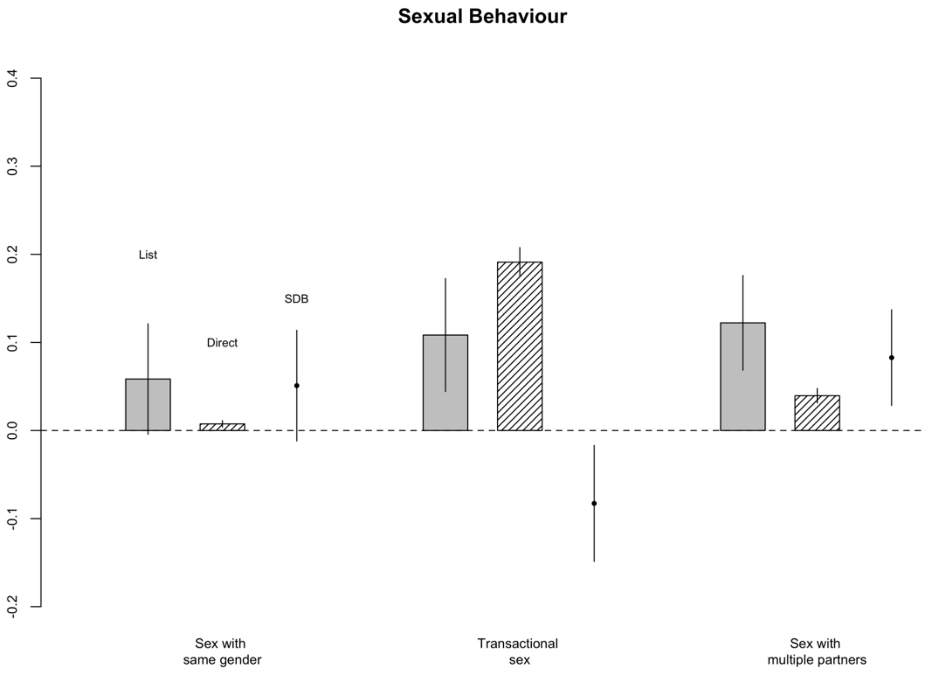
Note: Graph shows mean estimates with 95%-confidence intervals. List = List estimate; Direct = Direct estimate; SDB = sensitivity bias.

Figure 4.2: Sensitivity bias in intimate partner violence



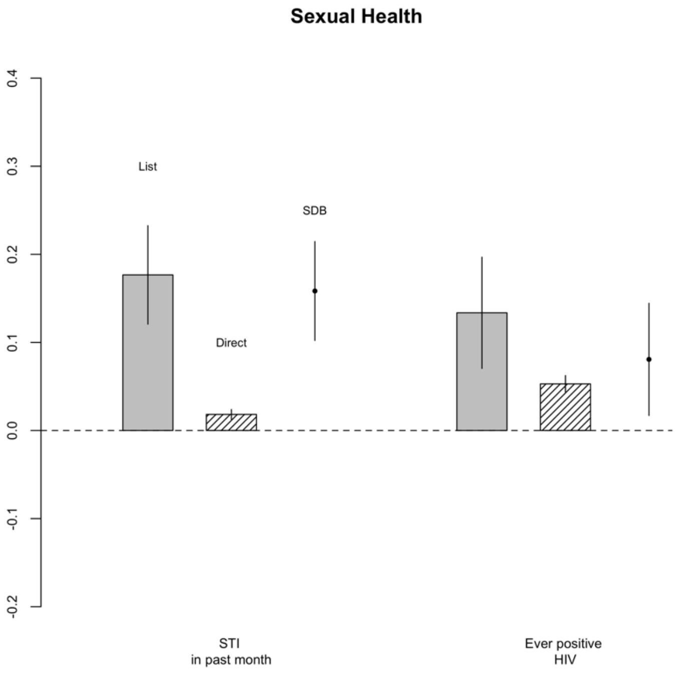
Note: Graph shows mean estimates with 95%-confidence intervals. List = List estimate; Direct = Direct estimate; SDB = sensitivity bias.

Figure 4.3: Sensitivity bias in sexual behaviour



Note: Graph shows mean estimates with 95%-confidence intervals. List = List estimate; Direct = Direct estimate; SDB = sensitivity bias.

Figure 4.4: Sensitivity bias in sexual health

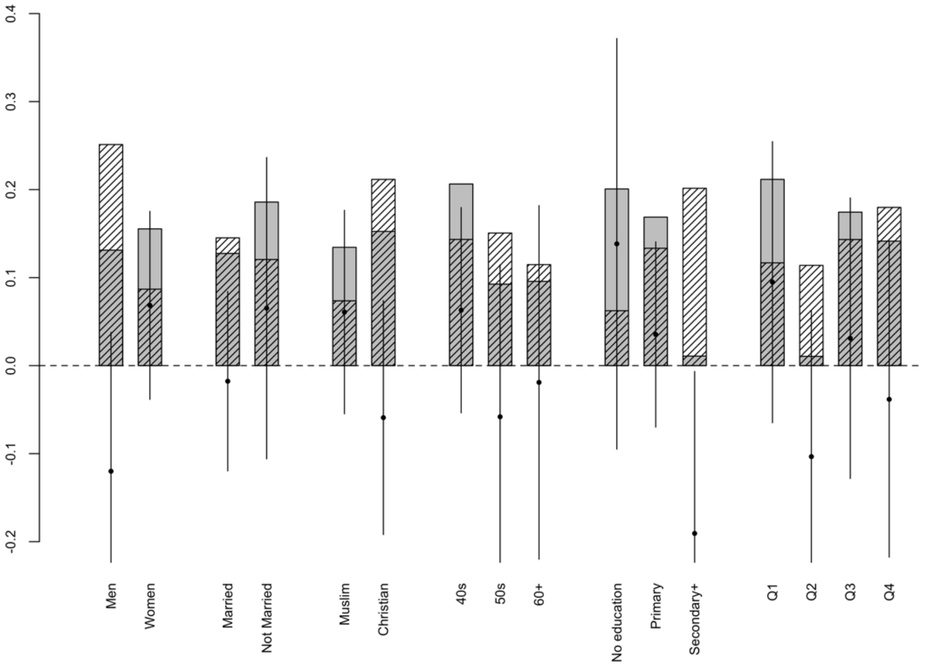


Note: Graph shows mean estimates with 95%-confidence intervals. List = List estimate; Direct = Direct estimate; SDB = sensitivity bias.

The revealed sensitivity bias varies greatly between outcomes. For the two outcomes on alcohol consumption (*drank alcohol* and *drank 6+ alcoholic drinks*), sensitivity bias is estimated to be 0.0 percentage points (95%-CI: -0.085-0.098) and 6.8 percentage points (95%-CI: -0.020-0.157), respectively. Neither bias is statistically different from 0. This indicates that alcohol consumption may not be a highly sensitive topic in this context of urban Tanzania. Drug use, a more sensitive topic, is underreported in the direct question by 8.2 percentage points (95%-CI: 0.026-0.138). The experience of physical and sexual intimate partner violence is underreported by 7.9 percentage points (95%-CI: 0.015-0.143) and 13.9 percentage points (95%-CI: 0.081-0.197), respectively. This is considerable given that direct report of physical intimate partner violence is already relatively high at 20.3 percent. All of these estimates are statistically significantly different from 0, indicating that there is a non-zero sensitivity bias. Looking at sexual behaviour, the smallest sensitivity bias is estimated for *sex with same gender,* at 5.1 percentage points (95%-CI: -0.012-0.114), which is not statistically significant. Given that the cost of providing the socially unpreferred response is a potential jail sentence if reported, the list experiment may not provide sufficient privacy or trust to the respondents for them to give a truthful response. *Transactional sex* seems to be overreported by 8.3 percentage points (95%-CI: -0.149- -0.017) in the direct question. The sensitivitiy bias for *sex with multiple partners* is estimated at 8.3 percentage points (95%-CI: 0.028-0.137). Sensitivity bias is also positive and statistically significant for the two outcomes of sexual health, *STI* and *HIV,* at 15.8 (95%-CI: 0.102-0.215) and 8.1 percentage points (95%-CI: 0.017-0.144), respectively.

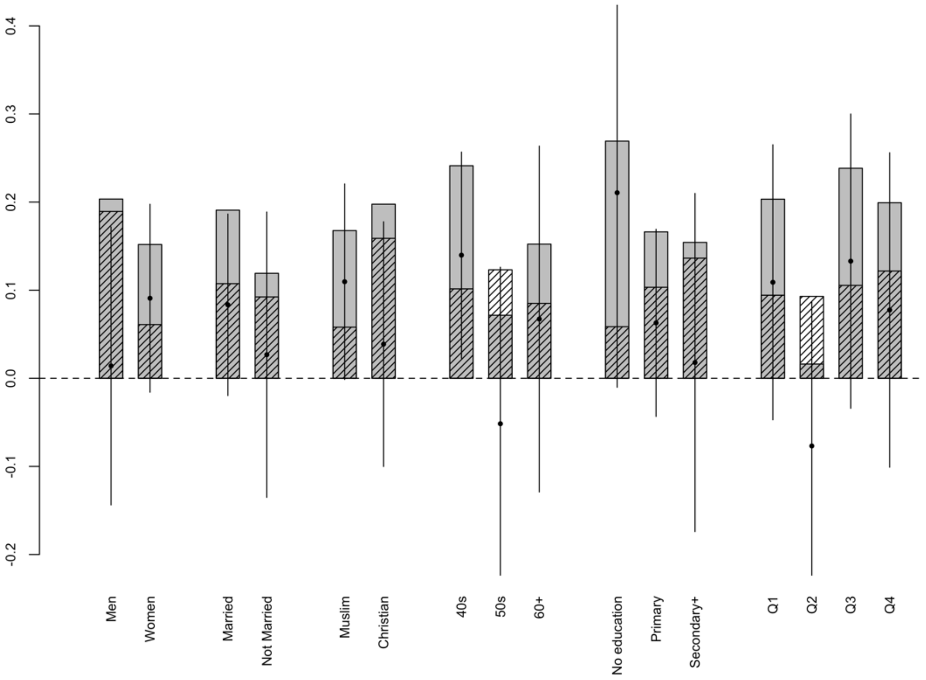
In the heterogeneity analysis, some noteworthy patterns emerge (see Figure 4.5 to Figure 4.14). For the outcome *drank alcohol* (Figure 4.5)*,* the bias varies between over- and underreporting between subgroups. While men overreport alcohol consumption, women tend to underreport, indicating that social pressure differs between men and women in this context. Similarly, Muslim respondents seem to underreport while Christian respondents overreport in this direct question. Younger and less educated respondents underreport, while those above 50 years, those with higher education and those in the lower quartiles overreport alcohol consumption. However, these biases are not statistically different from zero. Heavier alcohol consumption, *drank 6+ alcoholic drinks* (Figure 4.6), is underreported among all subgroups except those aged 50 to 59 years and those in the second lower wealth quartile. Again, these biases are not statistically different from zero. Sensitivity bias in drug consumption is positive and statistically significant for most subgroups. While sensitivity bias seems slightly lower among women, Muslims, individuals aged 40 to 49 years, and individuals in wealth quartile one and three, these differences are not large and not statistically significant.

Figure 4.5: Sensitivity bias in drug consumption - drank alcohol, across groups



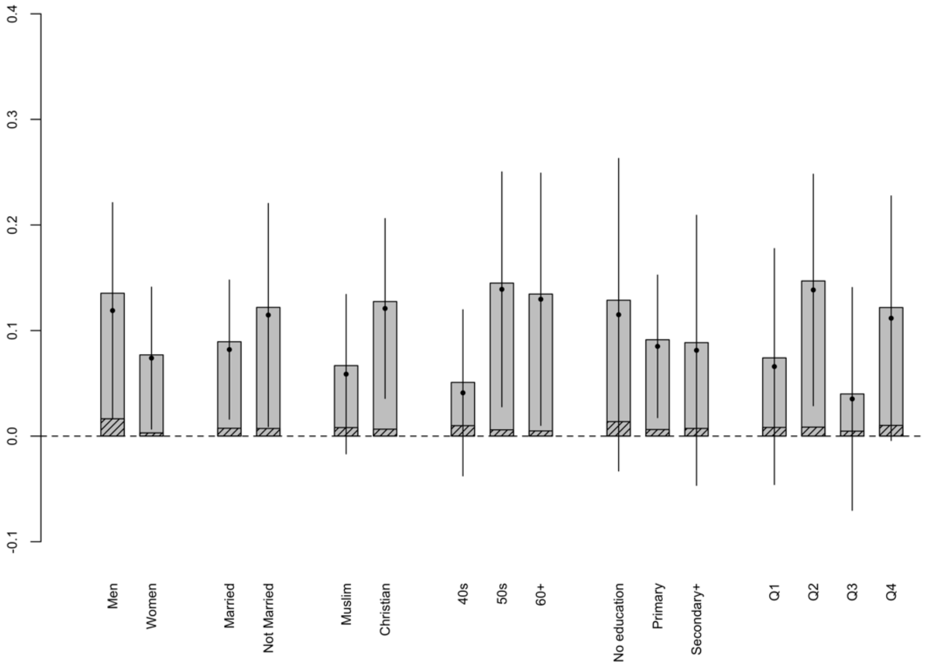
Note: Grey bars represent direct estimates, shaded bars represent list estimates, sensitivity bias is shown as point with 95%-confidence intervals.

Figure 4.6: Sensitivity bias in drug consumption - drank 6+ alcoholic drinks, across groups



Note: Grey bars represent direct estimates, shaded bars represent list estimates, sensitivity bias is shown as point with 95%-confidence intervals.

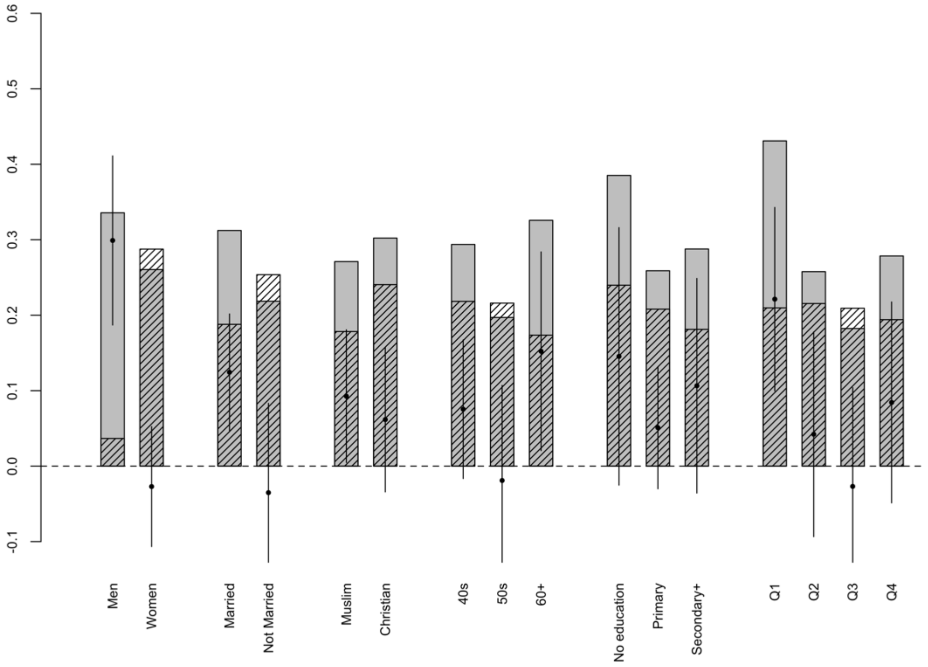
Figure 4.7: Sensitivity bias in drug consumption - ever used IV drugs, across groups



Note: Grey bars represent direct estimates, shaded bars represent list estimates, sensitivity bias is shown as point with 95%-confidence intervals.

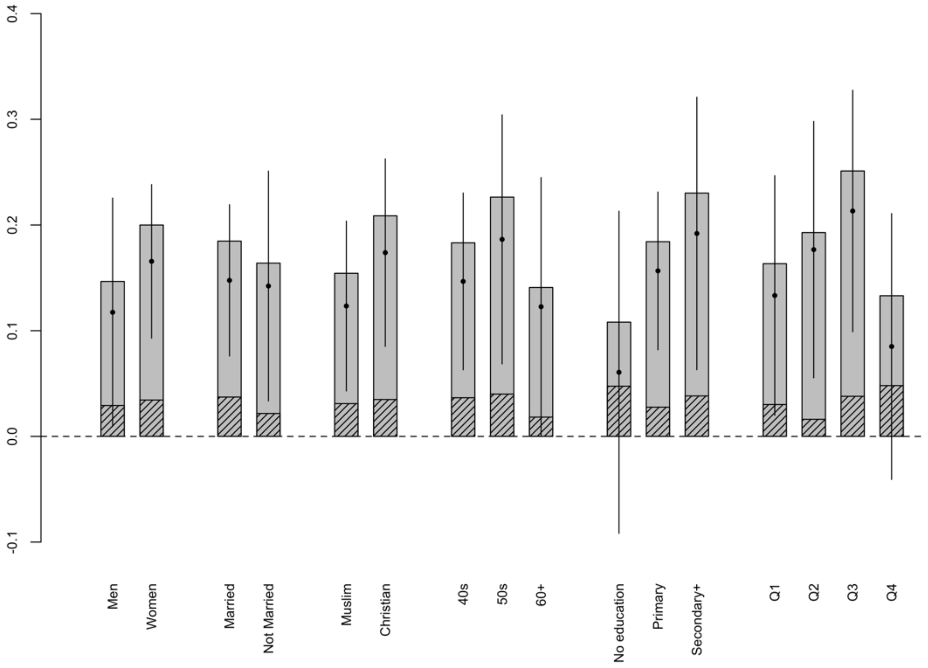
For *physical IPV* (Figure 4.8), a stark contrast emerges between men and women. Men show a very high sensitivity bias of 29.2 percentage points (95%-CI: 0.187-0.411) and women a comparatively small and insignificant overreporting. This indicates that reporting physical IPV is socially highly undesirable for men despite seemingly high prevalence, while this is less of a concern for women. Apart from men, sensitivity bias is statistically non-zero for the subgroups of married respondents, Muslims, those 60 years and older, and those in the lowest wealth quartile, but these differences between subgroups are not statistically significant. The bias in reporting of *sexual IPV* (Figure 4.9) is positive and statistically significant for almost all subgroups except for those with no education and those in the highest wealth quartile. Again, the differences between subgroups are not considerable and not statistically significant.

Figure 4.8: Sensitivity bias in intimate partner violence - physical IPV, across groups



Note: Grey bars represent direct estimates, shaded bars represent list estimates, sensitivity bias is shown as point with 95%-confidence intervals.

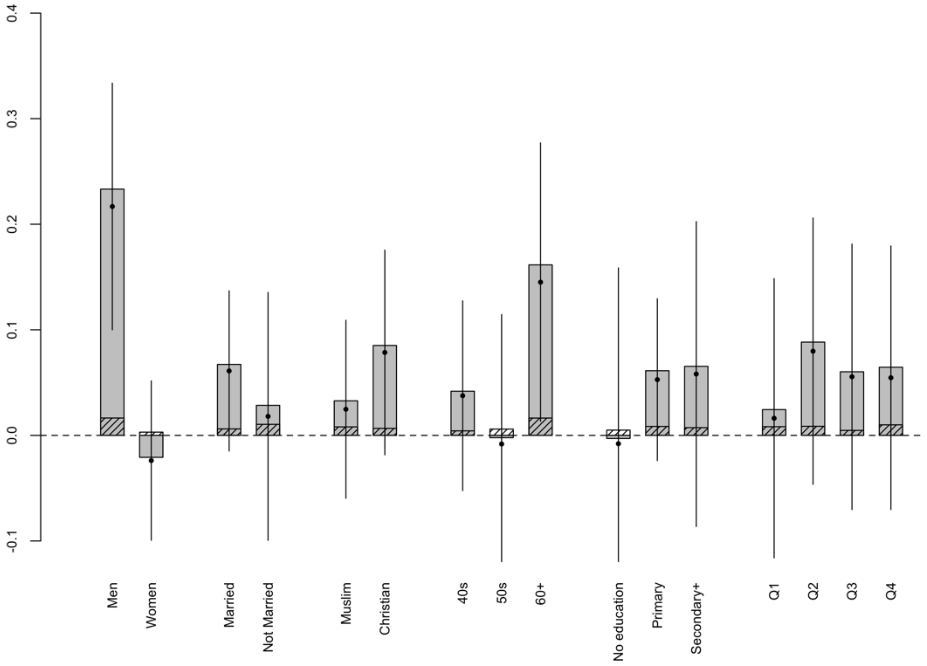
Figure 4.9: Sensitivity bias in intimate partner violence - sexual IPV, across groups



Note: Grey bars represent direct estimates, shaded bars represent list estimates, sensitivity bias is shown as point with 95%-confidence intervals.

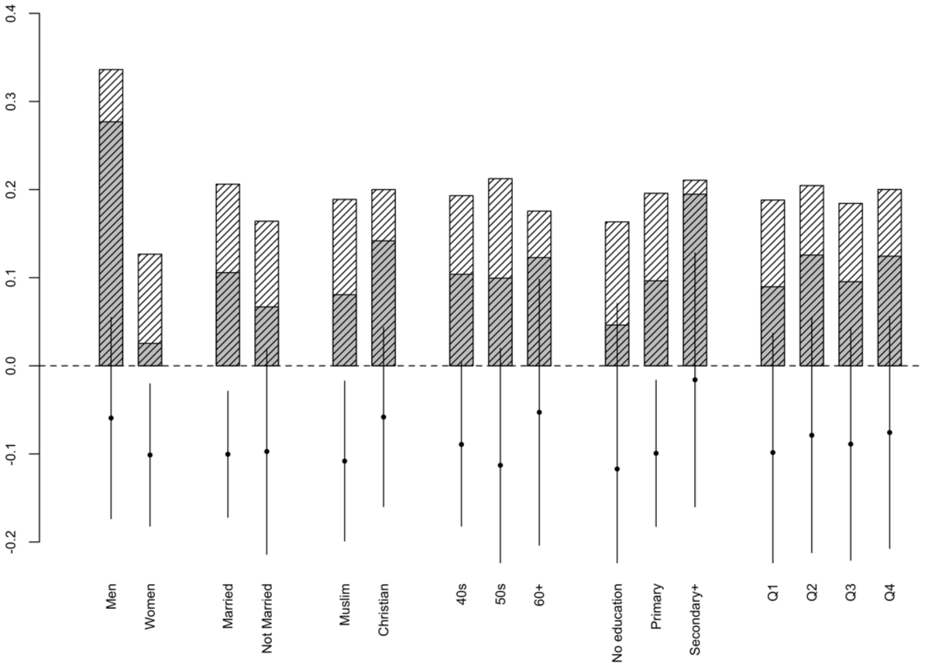
As in the estimation for the entire sample, sensitivity bias is not statistically significant for the outcome *sex with same gender* (Figure 4.10) for most subgroups. The two exceptions are men, where we find underreporting by 22.4 percentage point (95%-CI: 0.100-0.334), and those aged 60 years and older, where underreporting is 15.2 percentage points (95%-CI: 0.013-0.277). Sensitivity bias is negative, indicating overreporting, for all subgroups for *transactional sex* (Figure 4.11), although it is statistically significant only for some subgroups, including women, married respondents, Muslims, those who completed primary school. Having had multiple partners in the past 12 months (Figure 4.12), is underreported by most subgroups. In this outcome, again a contrast is seen between men, with basically no sensitivity bias, and women, who seem to underreport this behaviour by 12.7 percentage points (95%-CI: 0.051-0.188).

Figure 4.10: Sensitivity bias in sexual behaviour - sex with same gender, across groups



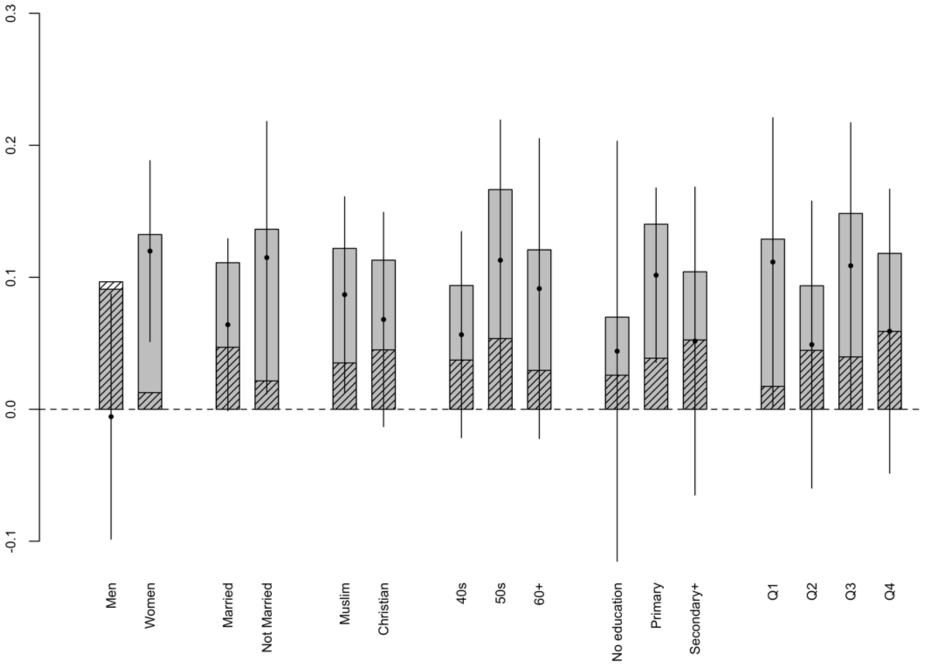
Note: Grey bars represent direct estimates, shaded bars represent list estimates, sensitivity bias is shown as point with 95%-confidence intervals.

Figure 4.11: Sensitivity bias in sexual behaviour - transactional sex, across groups



Note: Grey bars represent direct estimates, shaded bars represent list estimates, sensitivity bias is shown as point with 95%-confidence intervals.

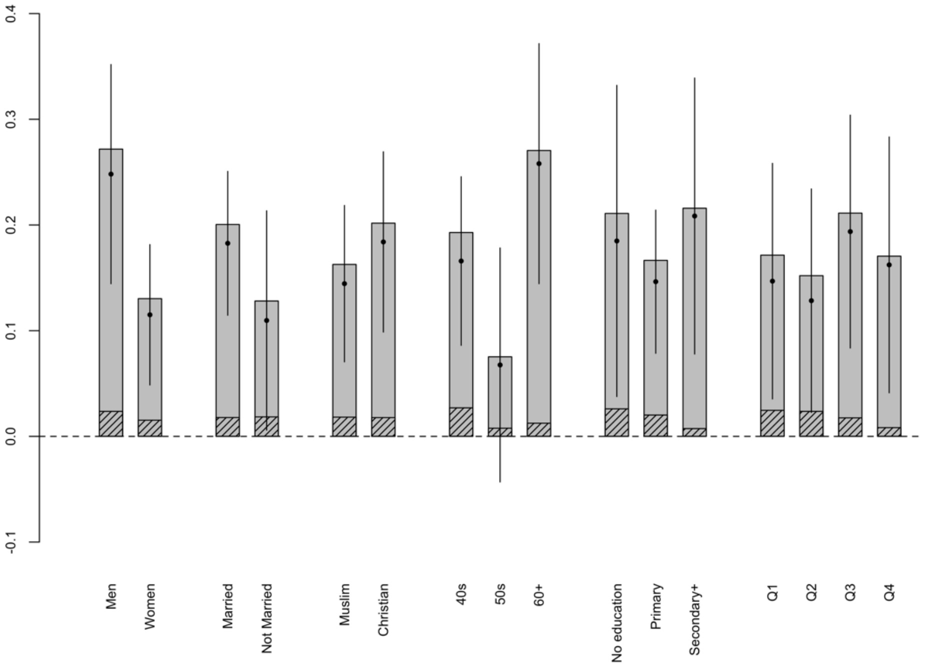
Figure 4.12: Sensitivity bias in sexual behaviour - sex with multiple partners, across groups



Note: Grey bars represent direct estimates, shaded bars represent list estimates, sensitivity bias is shown as point with 95%-confidence intervals.

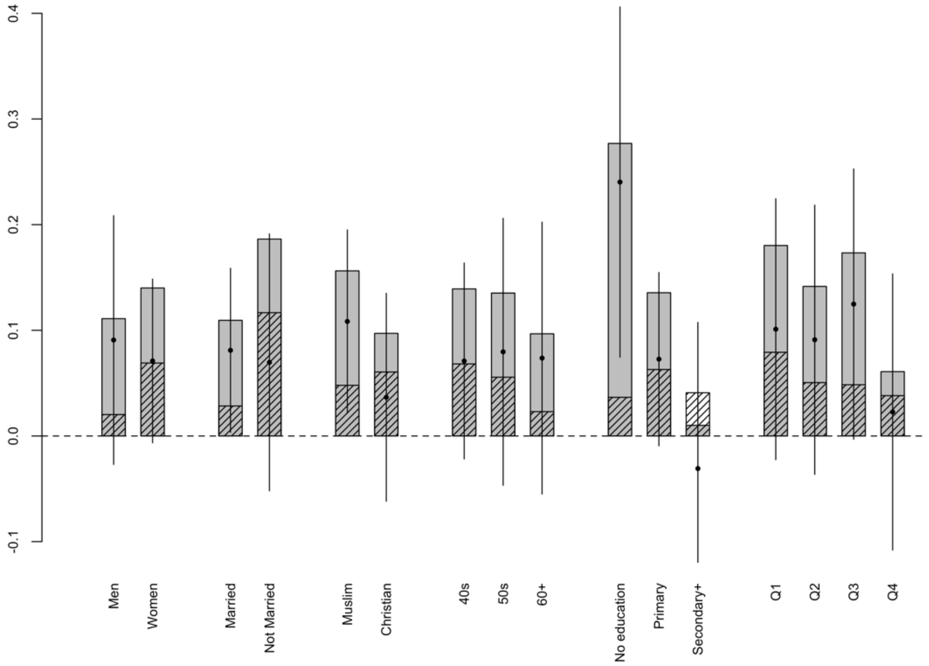
Out of the two outcomes for sexual health, having had a STI in the past 12 months seems to be more affected by sensitivity bias than ever having tested positive for HIV. Note that the phrasing of the sensitive item on STIs includes HIV as a sexually transmitted disease. The bias in *STI* (Figure 4.13) is positive and statistically significant for almost all subgroups. The bias is somewhat larger among men compared to women and among married respondents compared to unmarried respondents, but this difference is not statistically significant. The bias in *HIV* (Figure 4.14) is only statistically different from zero for respondents with no education and, smaller but still positive, for Muslim respondents.

Figure 4.13: Sensitivity bias in sexual health - STI, across groups



Note: Grey bars represent direct estimates, shaded bars represent list estimates, sensitivity bias is shown as point with 95%-confidence intervals.

Figure 4.14: Sensitivity bias in sexual health - ever tested positive for HIV, across groups



Note: Grey bars represent direct estimates, shaded bars represent list estimates, sensitivity bias is shown as point with 95%-confidence intervals.

4.4 Discussion

In this sample of adults in urban Tanzania, we find evidence of sensitivity bias in some questions on health and health behaviour but not for others. Alcohol consumption does not seem to be affected by sensitivity bias overall, but there is indication of differences across subgroups. Men do not seem to feel pressured to hide alcohol consumption, on the contrary they seem to overreport this. In the context of urban Tanzania, this is reasonable as men would often sit together in bars and share a beer with peers in the evenings. Not participating in this could indicate lack of financial liquidity or lack of social contact. Meanwhile, alcohol consumption seems to be less socially desirable for women, pushing them to underreport when asked directly. To our knowledge, this outcome has only been studied in the context of college student-athletes in the US where substantial bias was found (Druckman et al., 2015). The same study also found sensitivity bias in the use of performance-enhancing drugs, which is not directly comparable with IV drug use among a non-athlete adult population as in our study.

We find sensitivity bias in questions on physical and sexual IPV. Most other studies focus on physical IPV and women only. Among women in Rwanda and Nigeria (Cullen, 2020) as well as rural Burkina Faso (Lépine, Treibich, & D'Exelle, 2020), sensitivity bias was found. No difference between list estimate and direct question was found among poor women in rural

Ethiopia (Gilligan et al., 2021) and peri-urban Peru (Agüero & Frisancho, 2017). Interestingly, our results indicate a much larger sensitivity bias in reporting physical IPV among men, although this is not the case for sexual IPV. In this setting, physical IPV does not seem to be a very sensitive topic for women, while it does lead to considerable underreporting among men. Two other studies in rural Kenya and Peru also included men in this question as victims of IPV (Castilla & Murphy, 2021; Porter et al., 2021) and find no difference in the list estimate between men and women, but neither study includes a direct question and therefore an estimate of the sensitivity bias. Differences in sensitivity bias were found by level of education among poor adult women in Lima, Peru (Agüero & Frisancho, 2017). This study found no sensitivity bias among less educated women while they did for women with tertiary education.

Of the three outcomes on sexual behaviour, we find evidence of sensitivity bias for transactional sex and having had sex with multiple partners. Engagement in transactional sex seems to be overreported by men and women, opposite of common assumptions. While transactional sex poses a health concern due to its link with increased risk of HIV (Wamoyi et al., 2016), it does not seem to be negatively viewed in this context. Qualitative studies describe the close link between money and love in relationships and the gendered expectations that men offer material support to the female partner, also as a sign of commitment (Stoebenau et al., 2016). From a man's perspective, paying for sexual intercourse as the question states could therefore be taken as indication that he is able to fulfil his expected role in his relationships. From a woman's perspective, being paid could be understood as being able to form relationships with men who can provide for her. The connotation of the question would then be positive. The exact phrasing of the question is likely to be very important in triggering either negative or positive meaning. Few other studies analyse these outcomes of sexual behaviour. One study among men in the US find some overreporting in the direct question on sex buying (Roe-Sepowitz et al., 2019), mirroring our finding of overreporting. A review on interviewing tools for sexual behaviour in low- and middle-income countries reports that most studies compare audio computer-assisted self-interview with face-to-face interviews and sensitivity bias is only found in some settings (Phillips et al., 2010). For example, such a comparison among adolescents in rural Malawi provides some indication of sensitivity bias in reporting sexual experience (Kelly et al., 2013). However, evidence from list experiments on these outcomes is scarce.

Sexual health is studied even less with list experiments, although our results indicate that in this setting, sensitivity bias is present in questions on sexually transmitted diseases and HIV. We know of only one study measuring HIV-positive rate with a list randomisation tool in Kwa-Zulu Natal, South Africa (Haber et al., 2018). However, the list randomisation in that survey led to high refusal rates and the estimate of HIV-positive rate was not closer to the known true HIV rate compared to the direct question.

The list experiments seemed to work in reducing sensitivity bias to some degree in this study. Nevertheless, some concerns regarding their validity for some outcomes remain. A major question regarding the effectiveness of list experiments is the trade-off between an unbiased estimate and a larger variance (Blair et al., 2020). It is evident in our results that direct estimates are more narrowly estimated than the list estimates. However, for several outcomes analysed here, list estimates are considerably different from the direct estimates, indicating a large sensitivity bias. For these outcomes and in this setting, list experiments may therefore be a better option compared to the direct question. Due to the large variance, a larger sample size is required when using a list experiment compared to the direct question (Blair et al., 2020). Corstange (Corstange, 2009) recommends a sample size close to 2000, which is achieved in this study for all outcomes except for those on alcohol consumption. List experiments are weak tools for the precise estimation of rare behaviours (Ahlquist, 2018). The outcomes in this study are not very rare and therefore do not seem to be affected by this concern. The only exception could be sex with same gender, the most sensitive outcome in this context.

The study is also subject to some limitations. The questionnaire included nine list experiments which appeared one after the other. This may raise suspicions about the purpose of these questions among respondents, especially for those who are randomized to receive the sensitive item in several list experiments. However, the results do not show a pattern of reducing sensitivity bias across list experiments when ordered as they appeared in the questionnaire. There were also no reports of changed answer behaviour in the last compared to the first list questions by data collectors.

While list experiments may reduce sensitivity bias, the increased complexity of the method may introduce other forms of errors (Jerke et al., 2019; Kramon & Weghorst, 2019). Non-compliance with instructions appeared to be high in a list experiment on sexual and reproductive health in Cameroon and Côte d’Ivoire (Chuang et al., 2021). For political attitudes, it has been shown that inattentiveness among respondents can lead to biased estimates (Alvarez et al., 2019). In this study, there was no mechanism to filter out inattentive respondents with test questions.

4.5 Conclusion

Our results indicate the presence of sensitivity bias in the responses on health behaviour and health outcomes in this sample of adults in urban Tanzania. However, our findings for the outcomes alcohol consumption and transactional sex were different from common assumptions. List experiments are therefore valuable in correcting such assumptions. Moreover, sensitivity bias seems to vary not only between topics but also between population groups. Especially men and women seem to be affected differently by social pressure or expectations in terms of responding to sensitive questions. This bias influences our knowledge about health behaviour and health derived from surveys. Surveys should therefore be designed to reduce the risk of sensitivity bias. List experiments may be one method of reducing sensitivity bias. However, it is unclear whether sensitivity bias is completely removed. Moreover, the feasibility of implementing list experiments and their relative success in provoking more truthful responses strongly depends on the specific setting and topics.

4.6 Appendix

Table A 4.1: List of items

Outcome	Order of list experiments	Statements (English)	Statements (Swahili)
Drank alcohol AND Drank 6+ alcoholic drinks	List 1	1. I always eat three meals a day. 2. My house does not have running water. 3. I speak more than one language. 4. I have been diagnosed with diabetes or high blood sugar. 5 (Randomized – ARM 1). I drank alcohol in the last 30 days. 5 (Randomized – ARM 2). I drank 6 or more alcoholic drinks (3 or more Castle lite, 3 or more Ndovu, 2 or more Castle Lager, 2 or more Safari lager) in one occasion in the last 30 days.	1. Kwa kawaida nakula milo mitatu 2. Nyumba yangu ina maji ya bomba 3. Ninaongea lugha zaidi ya moja 4. Nimegundulika kuwa nina kisukari au wingi wa sukari mwilini 5 (Randomized – ARM 1). Nimekunywa pombe katika siku 30 zilizopita. 5 (Randomized – ARM 2). Nimekunywa vinywaji 6 au zaidi (ambavyo ni sawa na bia 3 au zaidi za ujazo wa chupa ndogo(castle light) au bia 2 au zaidi za ujazo wa chupa kubwa (Castle lager) ndani ya siku 30 zilizopita.
Ever used drugs	List 3	1. I belong to one of the religious denominations. 2. I exercise every day. 3. I drink soft drinks every day. 4. I have had more than 4 lifetime sex partners. 5 (Randomized). I have ever used injection drugs.	1. Mimi ni muumini wa mojawapo ya madhehebu ya dini 2. Ninafanya mazoezi kila siku 3. Huwa ninakunywa vinywaji laini(juisi , soda) kila siku 4. Nimeshakuwa na wapenzi zaidi ya wanne maishani 5 (Randomized). Nimewahi kujidunga madawa ya kulevya.
Physical intimate partner violence	List 7	1. I have more than four children. 2. I sleep more than 8 hours daily. 3. I have been diagnosed with high blood pressure in the last 12 months. 4. I was younger than 15 years old the first time I had sex. 5 (Randomized). I have ever experienced physical violence committed by my partner.	1. Nina watoto zaidi ya wanne(4) 2. Huwa nina lala zaidi ya masaa nane(8) kila siku 3. Nimegundulika kuwa na presha(ya damu) katika miezi 12 iliyopita. 4. Nilikuwa na chini ya miaka 15 nilivyoshiriki ngono(jamiiana) kwa mara ya kwanza 5 (Randomized). Mwenza wangu amewahi kunipiga.
Sexual intimate partner violence	List 8	1. I normally go to visit friends several times per week 2. I usually wake up before 7am. 3. I visited a health clinic in the past 3 months. 4. I have trouble walking across the room without assistance.	1. Huwa natembelea marafiki mara kadhaa katika wiki 2. Kwa kawaida huwa nina amka saa 1 asubuhi 3. Nilitembelea kliniki ya afya katika miezi 3 iliyopita 4. Ninapata shida kutembea(chumbani) bila ya msaada

		5 (Randomized). I have ever experienced sexual violence committed by my partner	5 (Randomized). Mwenza wangu amewahi kunibaka
Sex with same gender	List 2	1. I like reading English newspapers. 2. I do not own a cell phone. 3. I had malaria as a child. 4. I know someone who has disclosed their HIV positive status to me. 5 (Randomized – Males). I have ever had sex with a man. 5 (Randomized – Females). I have ever had sex with a woman.	1. Napenda kusoma magazeti ya kiingereza 2. Sina simu ya kiganjani 3. Niliugua malaria utotoni 4. Ninajua mtu anayeishi na virusi vya ukimwi ambaye amenishirikisha majibu yake yeye mwenyewe 5 (Randomized – Males). Nimewahi kufanya ngono(kujamiiana) na mwanaume. 5 (Randomized – Females). Nimewahi kufanya ngono(kujamiiana) na mwanamke.
Transactional sex	List 4	1. I own a vehicle. 2. I have back pain most days. 3. I exercise regularly. 4. I live with my spouse or partner. 5 (Randomized – Males). I have ever paid someone in exchange for having sexual intercourse. 5 (Randomized – Females). I have ever been paid in exchange for sexual intercourse.	1. Nina miliki gari 2. Huwa nina maumivu ya mgongo karibu kila siku(siku nyingi katika wiki) 3. Huwa nafanya mazoezi mara kwa mara 4. Ninaishi na mwenza wangu 5 (Randomized – Males). Nimewahi kununua/kumlipa mtu ili kufanya naye ngono (kujamiiana). 5 (Randomized – Females). Niwewahi kulipwa ili kufanya ngono (kujamiiana).
Sex with multiple partners	List 5	1. I watch television regularly. 2. I visit my ancestors home every holiday. 3. I have lived in Dar es Salaam for most of my life. 4. My hearing has worsened in the past few years. 5 (Randomized). I have had sexual intercourse with two or more different people in the last 12 months.	1. Huwa naangalia luninga(TV) mara kwa mara 2. Huwa natembelea nyumbani kwa asili yetu kila likizo(sikukuu) 3. Nimeishi Dar es Salaam muda mwingi wa maisha yangu 4. Masikio yangu yamepunguza sana kusikia katika miaka michache iliyopita 5 (Randomized). Nimefanya ngono(jamiiana) na watu wawili tofauti katika miezi 12 iliyopita.
STI	List 6	1. I like reading Swahili newspapers. 2. I prepare food for my family every day. 3. My vision has worsened in the past few years. 4. I have smoked cigarettes in the past 12 months.	1. Napenda kusoma magazeti ya kiswahili 2. Huwa naandaa chakula kwa ajili ya familia yangu kila siku 3. Macho yangu yamepungua sana nguvu ya kuona katika miaka michache iliyopita 4. Nimevuta sigara katika miezi 12 iliyopita

		5 (Randomized). I have had a disease which I got through sexual contact during the last 12 months.	5 (Randomized Nilipata ugonjwa uliotokana na kufanya ngono(kujamiiana) katika miezi 12 iliyopita.
HIV	List 9	1. I read the local newspaper almost every day. 2. I have at least one child. 3. I go to a religious worship most weeks. 4. I have ever had a stroke or mini stroke. 5 (Randomized). I have tested positive for HIV.	1. Huwa ninasoma magazeti ya kitanzania karibu kila siku 2. Nina angalau mtoto mmoja 3. Huwa nahudhuria ibada katika katika wiki nyingi(karibu kila wiki) 4. Nimewahi kupata kiharusi 5 (Randomized). Nimeathirika na VVU(HIV).

Table A 4.2: Sample sizes

	All	**Non-missing**	**No violations**	**Control**	**Treatment**
Drank alcohol	2270	1494	1491	749	742
Drank 6+ alcoholic drinks	2270	1482	1481	749	732
Ever used drugs	2270	2218	2217	1103	1114
Physical IPV	2270	2215	2205	1085	1120
Sexual IPV	2270	2215	2210	1124	1086
Sex with same gender	2270	2218	2214	1099	1115
Transactional sex	2270	2216	2201	1123	1078
Sex with multiple partners	2270	2214	2214	1110	1104
STI	2270	2214	2213	1123	1090
HIV	2270	2146	2146	1084	1062

Note: STI = sexually transmitted infections; HIV = human immunodeficiency virus; IPV = intimate partner violence.

Table A 4.3: Socioeconomic characteristics

	Observations (N)	Frequency (%)
Gender		
Male	730	32.16
Female	1540	67.84
Age (years)		
40-49	1089	47.97
50-59	611	26.92
60-69	376	16.56
70-79	134	5.90
80+	60	2.64
Education level		
None	357	15.73
Primary (Std 1-7)	1394	61.41
Secondary (Form I – VI)	409	18.02
VT or Uni	95	4.19
Marital Status		
Never Married	70	3.08
Separated & Divorced	196	8.63
Widowed	395	17.40
Currently Married/ Cohabitating	1596	70.31
Religion		
Christian	1035	45.59
Muslim	1224	53.92
Employment Status		
Employed (Part/Full/Self employed)	1035	45.59
Not Working	439	19.34
Homemaker	763	33.61
Other	14	0.62
Wealth quintiles		
WQ 1	506	22.96
WQ 2	392	17.79
WQ 3	427	19.37
WQ 4	443	20.10
WQ 5	436	19.78

Table A 4.4: Results of testing assumptions

	N Non-missing	N Violations	No Design Effect I	No Design Effect II	Joint test
Drank alcohol	1494	3	1.0000	0.0997	0.0000
Drank 6+ alcoholic drinks	1482	1	1.0000	0.4281	0.0000
Ever used drugs	2218	1	0.6719	0.2025	0.0152
Physical IPV	2215	10	1.0000	0.2924	0.0000
Sexual IPV	2215	5	1.0000	0.7374	0.0000
Sex with same gender	2218	4	1.0000	0.7795	0.0005
Transactional sex	2216	15	1.0000	0.3400	0.0000
Sex with multiple partners	2214	0	1.0000	0.4952	0.0000
STI	2214	1	1.0000	0.4822	0.0001
HIV	2146	0	1.0000	0.4890	0.0000

Note: Columns 3, 4, and 5 report p-values. STI = sexually transmitted infections; HIV = human immunodeficiency virus; IPV = intimate partner violence.

Table A 4.5: Outcome distribution in control groups

	N	Means	0	1	2	3	4
Drank alcohol	749	1.73	5.57%	27.77%	49.30%	15.76%	1.48%
Drank 6+ alcoholic drinks	749	1.73	5.13%	28.29%	48.41%	16.34%	1.62%
Ever used drugs	1103	1.73	1.85%	30.22%	57.24%	10.06%	0.63%
Physical IPV	1085	1.42	6.76%	39.77%	44.31%	9.02%	0.14%
Sexual IPV	1124	1.41	6.56%	43.17%	44.71%	5.48%	0.09%
Sex with same gender	1099	1.09	20.60%	49.10%	27.64%	2.62%	0.05%
Transactional sex	1123	1.46	8.36%	40.57%	42.53%	8.27%	0.27%
Sex with multiple partners	1110	1.83	1.40%	22.00%	63.14%	12.87%	0.59%
STI	1123	1.63	3.30%	30.55%	57.57%	8.45%	0.14%
HIV	1084	1.53	5.64%	39.47%	45.20%	9.32%	0.37%

Note: STI = sexually transmitted infections; HIV = human immunodeficiency virus; IPV = intimate partner violence.

Table A 4.6: Non-response to direct questions vs. list experiment

	NR (list)	NR (direct)	NR % (list)	NR % (direct)	P-Value
Drank alcohol	24	25	1.57	1.64	0.8186
Drank 6+ alcoholic drinks	21	16	1.39	1.06	0.1968
Ever used drugs	33	42	1.45	1.85	0.0947
Physical IPV	35	43	1.54	1.89	0.1573
Sexual IPV	35	43	1.54	1.89	0.1573
Sex with same gender	34	40	1.50	1.76	0.2734
Transactional sex	35	41	1.54	1.81	0.2889
Sex with multiple partners	35	44	1.54	1.94	0.1172
STI	34	45	1.50	1.98	0.0555
HIV	35	107	1.54	4.71	0.0000

Note: Column 5 reports p-values of the two-sided t-test comparing non-response in the list and direct question. NR = Non-response; STI = sexually transmitted infections; HIV = human immunodeficiency virus; IPV = intimate partner violence.

Table A 4.7: Non-response to list experiment in control vs. treatment groups

	N (T=0)	N (T=1)	NR % (T=0)	NR % (T=1)	P-Value
Drank alcohol	766	762	1.70	1.44	0.6904
Drank 6+ alcoholic drinks	766	742	1.70	1.08	0.3055
Ever used drugs	1133	1137	1.77	1.14	0.2160
Physical IPV	1112	1158	1.35	1.73	0.4650
Sexual IPV	1152	1118	1.56	1.52	0.9354
Sex with same gender	1128	1142	1.77	1.23	0.2835
Transactional sex	1153	1117	1.65	1.43	0.6772
Sex with multiple partners	1141	1129	1.58	1.51	0.8896
STI	1150	1120	1.65	1.34	0.5397
HIV	1136	1134	1.23	1.85	0.2312

Note: Column 5 reports p-values of the two-sided t-test comparing non-response in control and treatment group. NR = Non-response; STI = sexually transmitted infections; HIV = human immunodeficiency virus; IPV = intimate partner violence.

Table A 4.8: Non-response to direct question in control vs. treatment groups

	N (T=0)	N (T=1)	NR % (T=0)	NR % (T=1)	P-Value
Drank alcohol	766	762	1.70	1.57	0.8506
Drank 6+ alcoholic drinks	766	742	1.70	0.40	0.0143
Ever used drugs	1133	1137	2.03	1.67	0.5259
Physical IPV	1112	1158	1.80	1.99	0.7432
Sexual IPV	1152	1118	2.00	1.79	0.7169
Sex with same gender	1128	1142	1.95	1.58	0.4983
Transactional sex	1153	1117	1.99	1.61	0.4932
Sex with multiple partners	1141	1129	2.45	1.42	0.0733
STI	1150	1120	1.83	2.14	0.5885
HIV	1136	1134	3.87	5.56	0.0587

Note: Column 5 reports p-values of the two-sided t-test comparing non-response in control and treatment group. NR = Non-response; STI = sexually transmitted infections; HIV = human immunodeficiency virus; IPV = intimate partner violence.

Chapter 5

Formal vs. informal mathematics: Assessing numeracy with school and market items in 5,997 school children in North-West Nigeria

with: Ann-Charline Weber and Sebastian Vollmer

Abstract

While school-aged children in resource-poor settings often perform poorly on standardized tests in mathematics, they can frequently be seen engaging in market activities, conducting monetary transactions. This suggests that children in these settings actually have much more advanced skills in basic mathematics than what is assessed at school. In this descriptive study, we provide evidence of a considerable skill gap between formal mathematics and informal mathematics in a sample of 5,997 school-aged children in North-West Nigeria. We explore several potential explanations for this skill gap. Current teaching strategies seem to miss out on existing skills in mathematics that could be built on in order to improve children's performance in school mathematics.

5.1 Introduction

Large standardized tests of learning outcomes, such as the International Common Assessment of Numeracy, the Early Grade Math Assessment, the Annual Status of Education Report in South Asian countries, and UWEZO in East African countries, show that school-aged children in many resource-poor settings across the globe have rather poor skills in mathematics (ASER Centre, 2020; PAL Network, 2020; Uwezo, 2019a, 2019b). Most recent results from Tanzania, for example, showed that in primary grade 3, only 59 percent of children achieved numeracy at subtraction level according to the national curriculum for grade 2 (Uwezo, 2019a). In Uganda, the percentage of children in primary grade 3 to 7 who could perform division at the level of grade 2 reached only 45 percent in 2018 (Uwezo, 2019b). In contrast, in some of these settings, school-aged children can often be seen engaging in market transactions, buying and selling goods on their own. This suggests a gap between arithmetic skills assessed in school and arithmetic skills existing in the daily life of the children. In fact, children seem to develop distinct concepts for numbers and calculations based on context and their understanding of these concepts may vary with their socioeconomic background (Khan, 2004; Sitabkhan, 2011; Spinillo, 2018). The literature therefore distinguishes between formal and informal mathematical knowledge (Sitabkhan et al., 2018). Formal mathematics refers to concepts that are the focus of school instruction while informal mathematical knowledge is developed in everyday life, outside school, and is sometimes referred to as street mathematics. A few studies have investigated how children use different strategies to solve calculations presented as abstract school mathematics and calculations set in real-life context and that they generally perform better in the latter (Carraher et al., 1985, 1987; Nunes et al., 1993). However, gaining skills in one concept of mathematics is not automatically transferred to another, according to one study among pre-school children in India (Dillon et al., 2017). Standardized tests such as the Early Grade Mathematics Assessment only capture children’s performance in formal mathematics and overlook children’s existing skills in informal mathematics. One study in Kolkata, India, showed a considerable gap between skills in formal and informal mathematics in a sample of about 200 working children (Banerjee et al., 2017). These children, surveyed in informal markets, were able to routinely solve market transactions while they were much less able to solve similar tasks framed as formal mathematics.

We included a market simulation game in the design of a school based learning assessment in North-West Nigeria and find a considerable skill gap between formal and informal mathematics in a large sample of school-aged children in North-West Nigeria. In contrast to the study by Banerjee et al. (Banerjee et al., 2017), these children were sampled and assessed at their schools. We explore several potential explanations for this finding and suggest that being engaged in market activities helps children to learn how to do calculations but that these skills are not transferred to formal mathematics. To our knowledge, this study is the first study providing quantitative evidence of this skill gap between formal and informal mathematics among school-aged children in a low- and middle-income country and, with a total sample of 5,997 children, the largest study.

The rest of the paper is structured as follows. Section 2 presents background information on the study setting and sample as well as descriptions of the learning assessment designed for this study and data collection. Section 3 presents results, starting with a thorough description of the skill gap, followed by an analysis of potential explanatory mechanisms. Section 4 discusses the findings and concludes.

5.2 Background and data

5.2.1 Study setting and sample schools

The study was conducted in Sokoto, a state in North-West Nigeria, bordering Niger. The state is one of the poorest states in Nigeria. According to the Demographic and Health Survey 2018, 52 percent of households belonged to the lowest wealth quintile of Nigeria (National Population Commission (NPC) [Nigeria] and ICF, 2019). Some indicators for high poverty levels are the rate of only 34.4 percent of households with an improved source of drinking water and an under-5 mortality rate of 197 deaths per 1000 live births compared to 132 in Nigeria as a whole (National Population Commission (NPC) [Nigeria] and ICF, 2019). 54.8 percent of children were stunted, an indication for cumulative growth deficits pointing at poor nutrition (National Population Commission (NPC) [Nigeria] and ICF, 2019). The primary school net attendance ratio was 28.9 percent (National Population Commission (NPC) [Nigeria] and ICF, 2019).

The study was developed and conducted during the impact evaluation of the Nigerian Partnership for Education Project (NIPEP) and thus the setting and sample is determined by this evaluation. NIPEP, a large-scale programme to improve access to and quality of basic education, was implemented between 2015 and 2019 in five north western states by federal and state ministries and education boards. In 2018 and 2019 components of NIPEP were evaluated using a randomised controlled trial in primary schools in nine of the 23 Local Government Areas of Sokoto State. As the evaluation took place in the last phase of programme implementation, a sample of only those schools was left, which met eligibility criteria for NIPEP but had not yet been included in the project nor received any similar intervention. The schools that were included last in the project were those that had not been able to meet eligibility criteria before, such as setting up a bank account and an active school-based management committee. Schools with fewer than 35 or more than 160 children enrolled in grade 2 according to official school registers were excluded from the evaluation. Schools in very distant and hard-to-reach local government areas were also excluded for safety reasons.

The results of the evaluation of NIPEP are described elsewhere (Ochmann et al., 2021). As the intervention was not successful in improving learning outcomes among children, we disregard the intervention in the following analysis. Nevertheless, the evaluation study determined the setting and sample for this study and the main sections of the learning assessment were initially designed by us for the purpose of the programme evaluation of NIPEP.

5.2.2 Design of learning assessment

The learning assessment included a numeracy and a literacy section, followed by a brief interview capturing child characteristics. The numeracy section consisted of understanding numbers, basic calculations, and a market simulation game. The literacy section consisted of items testing letter recognition, reading in various difficulties from letters to a paragraph, reading comprehension, listening comprehension, and writing. The assessment used similar tasks as internationally comparable tests such as EGMA, EGRA, and ASER, capturing different levels of numeracy and literacy. While its main purpose was not a comparable assessment of learning outcomes across population groups, it was designed to detect small differences in skill levels in a short assessment time of about 30 minutes and a maximum of 40 minutes. The assessment was adjusted to the local context, based on the official school curriculum but oriented towards the actual skill level observed during a pilot phase. Children in grade 2, 3 and 4 were also presented with tasks for grade 1 as the observed numeracy and literacy was lower than expected given the curriculum.

This study focuses on the numeracy part of the learning assessment, but a description of the full assessment can be found in the supplementary material. All assessment items were presented on flash cards in one-on-one interactions between an enumerator and a child. The items could therefore be seen in written form, but were answered orally. Children were given tasks with increasing levels of difficulty, starting with counting and number recognition in the ranges up to 10 and up to 20, and ending with addition and subtraction in the number range up to 100. For each task, an item of the first level of difficulty was presented. If the item was answered correctly, an item of the next level of difficulty was presented. If the item was not answered correctly, a second item of the same level of difficulty was presented. If this second item was answered correctly, the child was given an item of the next level of difficulty. If the second item was not answered correctly, the child was not given any further item on this task, but proceeded to a different section.

Tasks on basic calculations included addition and subtraction for three levels of difficulty, i.e. in the number ranges up to 10, up to 20, and up to 100. The items in the number range up to 100 included numbers divisible by 5 (i.e. last digit being either 0 or 5) to design them most similarly to calculations in the market simulation game where money in the local denomination Naira was used.

The market simulation game mimicked transactions of buying and selling goods in a market. The first part focused on buying. The child was given token money (laminated copies of Nigerian bank notes) and flash cards with pictures of a pencil and a book were laid down between the enumerator and the child. Prices for each item were announced. The child was asked to buy one pencil using the token money and to give the exact amount. Next, the child was asked to buy two books, again giving the exact amount. This required an addition in the number range up to 100 and identifying the correct bills. The second part focused on selling. The child took on the role of a seller, flash cards of mangoes and oranges were put on display, new bills were given to the child, prices were announced. The enumerator then requested to buy one item, paying with a bill that required change to be given. The child had to use his or her token money to return the exact change. This required a subtraction in the number range up to 100 and an addition as the change required at least two bills. Next, the enumerator bought two goods, paying with a bill that required change to be given. Again, the child was asked to return the exact change. This required first an addition of the prices of the two goods, followed by a subtraction in the number range up to 100, and another addition for identifying the sum of the required bills. The child was given a defined set of bills both for the buying and the selling part. The given set of bills ensured that the child always had the necessary bills to hand over the exact amount, but would never give the correct answer by simply handing back all bills remaining in his or her hands. All prices were chosen to be realistic for the goods at hand, but still arbitrary to avoid children knowing prices and potential answers by heart.

5.2.3 Data collection

Data used in this analysis was collected between November and December 2019 in 128 primary schools in 9 local government areas of Sokoto State. The data collection included an observational survey of all schools, interviews with school staff, and learning assessments with children. Data was collected on tablets by enumerators hired by a Nigerian survey firm. All enumerators were trained in the survey tools. A team of research assistants accompanied the training and data collection.

Data collection was planned to take place during school hours. Upon arrival at a school, one enumerator was supposed to start with the observational survey of the school. This tool captured attendance and current activities at the moment of observation, the condition of school buildings, the availability of facilities such as sanitation and power supply, and copies of school records to assess the number of teachers and enrolled children. At the same time, the remaining team of enumerators would conduct interviews with school staff and learning assessments with children. Inclusion of 25 children each from grades 2, 3, and 4 in the learning assessments was intended.

Actual data collection deviated from the plan. The teams of enumerators often found schools to be closed at the time of the visit, with no adult or child present during school hours. In these cases, and when only a small number of children was present at the school, children were called to school from the surrounding village in order to conduct the learning assessments. The final sample of children covered might therefore include children who do not regularly attend school. For robustness checks, we restricted the sample to schools where learning was ongoing at the time of the visit.

5.2.4 Ethical considerations

The data was collected in the framework of the evaluation of NIPEP. The evaluation study was conducted in cooperation with the Nigerian Federal Ministry of Education.

5.3 Results

5.3.1 Sample characteristics

A total of 5,997 school-aged children from 128 schools completed the learning assessment, 2,803 from grade 2, 2,048 from grade 3, and 1,123 from grade 4 (47 schools did not have grade 4). The share of female children was 43 percent, and the mean age was 10.1 years. Table 5.1 contains a summary of descriptive sample characteristics. Figure A 5.1 shows the number of observations in the full sample and reasons for observations dropping from the sample used in the regression analysis.

Observational data was available from 125 schools. In 23 percent of schools, we observed teaching activities, meaning that at least one teacher and at least one child were present and learning was going on (see Table 5.1). 83 percent had any permanent buildings on the school compound, made from cement or brick. Overall, enumerators rated the general condition of school buildings and the compound as poor or very poor in 60 percent of observed schools. 6

percent of observed schools had a stable power supply, 22 percent had any water available, and 31 percent had sanitation facilities.

Table 5.1: Sample characteristics

	Full sample	Analysis sample
Student characteristics		
Mean age	10.08	9.99
Grade 2	0.47	0.44
Grade 3	0.34	0.35
Grade 4	0.19	0.20
Female	0.43	0.42
Sell in market	0.32	0.32
Buy in market	0.39	0.38
Study at home	0.40	0.42
Solve listening comprehension	0.22	0.23
Recognise numbers up to 10	0.54	0.57
Interview characteristics		
Outside school hours	0.03	0.03
Shorter 5 min	0.18	0.19
Shorter 10 min	0.58	0.59
School characteristics		
Learning activities observed	0.26	.
Any teacher present	0.54	.
Any student present	0.68	.
Water available	0.22	.
Toilets available	0.31	.
Power available	0.06	.
Permanent building	0.83	.
At least average condition	0.40	.

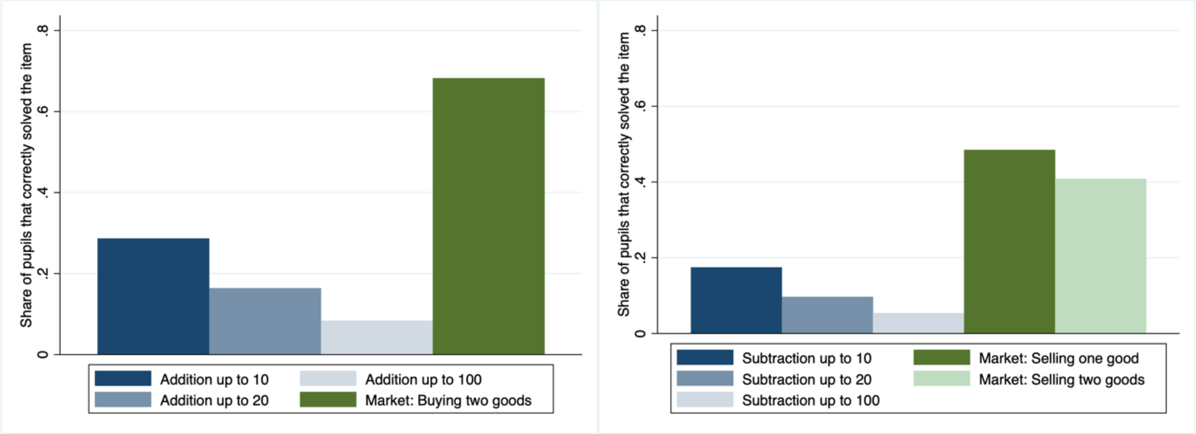
Note: Summary statistics (means or shares) of pupil characteristics, interview characteristics, and school characteristics, in the full sample and the sample used in regression analysis (analysis sample).

5.3.2 Skill gap between formal and informal mathematics

First, we make a simple comparison between the shares of children who were able to solve different calculation tasks. Results are shown in Figure 5.1. In the full sample of 5,997 children of grades 2 to 4, 28.5 percent were able to perform additions in a number range up to 10 and 8.4 percent in a number range up to 100. In contrast, 68.3 percent of the children were able to solve the market simulation game that involved buying two items simultaneously, a task comparable to addition up to 100. The result is similar for subtraction tasks. 17.5 percent, and 5.4 percent could solve subtraction tasks with numbers up to 10 and up to 100, respectively. A much larger share, 48.5 percent, was able to solve the task of selling one item in the market simulation game. This task was most similar to but slightly more complex than the subtraction up to 100. Still 40.9 percent of children were able to solve the task of giving the exact change after selling two items simultaneously, which required an even more complex calculation.

We define the share of children who were able to solve the addition (subtraction) task framed as buying (selling) but not the corresponding standard addition (subtraction) task with numbers up to 100 as the share of children with a skill gap. These children were clearly able to perform the task but could not access this skill when presented in a formal way. We find that the share of children with a skill gap is 60.4 percent for addition and 43.7 percent for subtraction.

Figure 5.1: Performance in different learning assessment items



(a) Addition *(b) Subtraction*

Note: Performance in learning assessment. N=5,997. Panel (a) contains tasks related to addition. Panel (b) contains tasks related to subtraction. The tasks "addition up to 10", "addition up to 20", "addition up to 100" in panel (a) and "subtraction up to 10", "subtraction up to 20", "subtraction up to 100" in panel (b) are standard calculation items presented numerically on flash cards. The tasks "market: buying two goods", "market: selling one good", and "market: selling two goods" mimic market transactions where the pupil has to buy goods shown on flash cards and pay (with token money) the correct amount or sell one/two goods and give the correct amount of change, respectively.

The skill gap is found in our sample across categories. Children with a skill gap exist in all schools of our sample, with almost no difference between state and Islamiyya schools[2] (addition: 60.4 vs 59.9 percent, subtraction: 44.0 vs 44.9 percent), and small differences between urban and rural locations (64.0 vs 59.2, 48.9 vs 41.9 percent), and schools where learning activities were observed during the data collection and those where no learning activities were observed (56.2 vs 61.7, 39.3 vs 45.4 percent). The skill gap is similar between boys and girls (59.7 vs 61.4, 43.3 vs 44.1 percent). While girls performed worse in the standard tasks, they performed only slightly worse in the tasks framed as market activities. We also find evidence of the skill gap in a previous survey in the same schools one year earlier. Of all 5,717

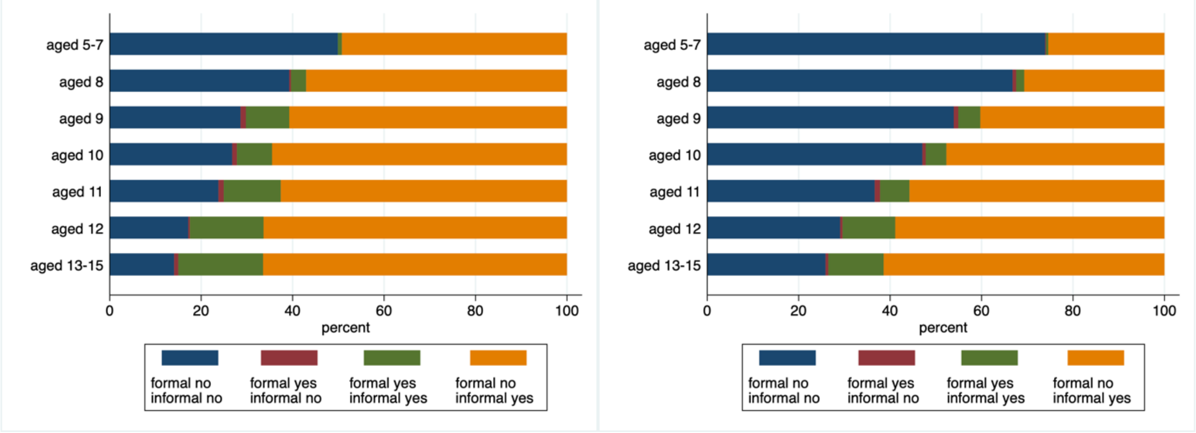
[2] Islamiyya schools are Islamic schools which are integrated into the public school system. While these schools have a focus on Islamic education subjects, they also teach primary school children reading and mathematics (Aminu Musa Yabo, 2017).

children in that earlier sample, 49.7 percent showed a skill gap in addition and 42.2 percent a skill gap in subtraction.

Looking at the performance of children across different age groups shows that both the share of children able to solve the standard task and the share of children able to solve the task framed as market transaction increase with age. For addition (subtraction) the share of children able to solve the standard task is 0.9 percent (0.7 percent) for children between 5 and seven years of age and 19.5 percent (12.7 percent) for children aged between 13 and 15. The share able to solve the corresponding task framed as buying two goods (selling one good) is 50.1 percent (25.9 percent) in the youngest age group and 85 percent (73.5 percent) in the oldest age group. As more children learn to perform market tasks than standard tasks, the share of children with a skill gap increases with age. This is especially pronounced for subtraction, where the share of children showing a skill gap is 25.4 percent in the youngest and 61.4 percent in the oldest age group. This is illustrated in Figure 5.2. The pattern is very similar across grades (see Figure A 5.2), even if we look at children of the same age across grades (see Figure A 5.3).

We thus find that a skill gap between formal and informal mathematics exists within children.

Figure 5.2: Performance in learning assessment items over age groups



(a) Addition

(b) Subtraction

Note: Performance in learning assessment for each age group, age in years. N = 4,345. Panel (a) contains tasks related to addition. Panel (b) contains tasks related to subtraction. “formal” refers to standard calculation tasks of addition (subtraction) up to 100, presented as formal mathematics. “informal” refers to calculations up to 100 presented as market transactions. Bars show shares of pupils in the following categories: (1) pupils who could solve neither standard tasks nor the market simulation; (2) pupils who could solve standard tasks, but not the market simulation; (3) pupils who could solve both standard tasks and the market simulation; and (4) pupils who could not solve standard tasks, but could solve the tasks of market simulation.

5.3.3 Ruling out design effects

All tasks of the learning assessment were presented in the same way using flashcards in one-on-one interactions. However, there were several differences in the design of tasks of formal mathematics and the tasks in the market simulation game. These differences relate to the need to give a verbal answer in the formal tasks, the different possibilities of guessing answers, and the need to recognise numbers in the formal tasks. In the following section, we explore whether effects of the particular design of the learning assessment tasks could explain the skill gap.

In the standard tasks, children were asked to give the answer verbally whereas in the market simulation game children had to hand over bills and were not required to speak. This suggests another explanation for the skill gap. Children could have been shy to speak out in the one-on-one interaction with the previously unknown adult enumerator. To explore this possibility, we added an item to the numeracy section. If a child did not solve the item of addition with numbers up to 10, he or she was given a new item with a similar task. Four flash cards with different numbers were laid out in front of the child and he or she was asked to pick the card with the correct answer instead of saying the answer out loud. Only an additional 267 children were able to solve the item in this fashion, 7.8 percent of those who received the item. Part of these children could have picked the correct flash card out of the four simply due to guessing.

A further difference lies in the probability of guessing answers correctly. In theory, children had a higher chance of guessing the correct answer in the market simulation game than the standard calculation tasks. The set of bills available to them gave them a limited number of possible combinations that they could hand over to the enumerator. For the task of buying two goods, a child had six bills with four different values in his/her hands, and thus a choice out of four assuming that he/she would hand over only one bill and is aware that some bills have the same value. For the selling task, the child had nine bills with three different values in his/her hands. For the correct answer, the child had to combine two or three bills. If the child was simply guessing, it is implausible that he/she would choose two or three. Nevertheless, including single bills as well as combinations of two (and three) as options, the child has a chance of 1 in 8 (2 in 15) to guess correctly. In the standard mathematics tasks, the chance of guessing the right answer out of the full range of numbers is very low. However, we assume that children would stick to the number range of the calculation task presented to them, drastically limiting the choice set for the standard tasks. Especially when comparing the standard subtraction task in the number range up to 10 and the task of selling two goods, the difference in the chance of guessing correctly is small and unlikely to explain the skill gap of 34 percentage points. Moreover, the share of children picking the correct card out of four in the task of addition up to 10 was much less than 25 percent (7.5 percent), indicating that not many children attempted a guess to solve the task. It therefore seems implausible that many more would have guessed in the market simulation game.

An alternative explanation for the skill gap based on the assessment design could be that children were not able to solve standard mathematics tasks, because they were unable to recognize numbers. In fact, 45.6 percent of children in the sample did not recognize numbers up to 10 whereas 84.1 percent of children were able to solve the task of buying one good for 10 Naira, i.e. recognising the 10 Naira bill. In the market simulation game, it was not strictly necessary to be able to read numbers as the bills have other features to tell them apart (i.e. colour, pictures, size) and all prices were announced orally. We therefore instructed enumerators to read out the standard calculation items from the flash card, if the child did not answer the item on the first attempt. This increases the share of children able to solve the math items from 8.4 percent to 14.8 percent for addition and from 5.4 percent to 10.1 percent for subtraction up to 100. The observed skill gap decreases by 6.0 percentage points and by 4.1 percentage points, respectively (see Figure A 5.4). Hence, it is possible that the difficulty in reading numbers could explain part of the finding. Alternatively, the increase could also be due to children simply needing a second attempt for answering the calculation task.

One could possibly argue that the calculations in the tasks framed as market transactions do not truly resemble a task in the number range up to 100 but that children break up these tasks into separate steps comparable to a lower number range. Firstly, this already is a cognitive effort that could potentially be translated to a standard task and therefore would indicate a skill gap if the same is not used for the standard tasks. Secondly, we do still see a significant gap between standard tasks in the number range up to 10 and the tasks framed as market transactions. 43 percent of all children were able to solve the task framed as buying two goods but not able to solve standard addition tasks in the number range up to 10, 34 percent solved the task of selling two goods but not the standard subtraction task up to 10. Again, the observed skill gaps for these number ranges are found across categories in our sample.

We conclude that these differences in the design of the test items could at most explain why some individual children were able to solve tasks in the market simulation game and not able to solve the standard mathematics tasks. Nevertheless, they do not provide a sufficient explanation for the size of the main finding. The skill gap is much larger than what design effects could explain.

5.3.4 Robustness of skill gap to data quality

One concern regarding our findings is limited trust in the data. While many enumerators knew the learning assessment from a previous survey, the training on the tool prior to this data collection was short. Moreover, we observed that several instructions regarding implementation of data collection were not followed. While instructions specified that male enumerators conduct learning assessments with male children and female enumerators with female children, this was not strictly followed. Some few learning assessments were conducted outside school hours. In addition, there are inconsistencies in the GPS locations that were automatically recorded while filling the form of the learning assessment. 22.5 percent of observations do not include a proper GPS location and for 3 percent the GPS location did not correspond with the school premises. This may be explained with technical issues, but we cannot rule out data fabrication for these observations. Most worryingly, learning assessments were often of surprisingly short duration. The duration of each learning assessment was captured with automatic time stamps. The test was designed to take 30 minutes if a child managed to solve all levels of difficulty. Even for a child performing poorly, some sections had to be completed in any case and a minimum of 21 tasks and 13 interview questions were presented. The tasks included one listening comprehension exercise which required a short story to be told. Nevertheless, 18 percent of learning assessments were completed in less than five minutes, suggesting that enumerators did not conduct the assessments with sufficient care. 58 percent were completed in less than 10 minutes.

Despite our concerns, the finding is robust. We observe the skill gap for all enumerators. If we restrict the sample to observations taken within school hours with proper GPS location, or a duration of at least 10 minutes, or both, the size of the skill gaps increases slightly, to 48 percent for subtraction and 73 for addition when applying both criteria.

Related to data quality is the fact that many children could not be randomly sampled in school from their classes, but were called from the surrounding village to participate in the test. Children who were selected at school may be systematically different from children who had to be called from the village. The schools of the latter group may be closed more often than the schools of the former group or attendance of children might be less regular. These children are then likely to have lower skills in formal mathematics compared to those who benefit from regular teaching. In schools where teaching was going on at the time of the visit, the share of children showing the skill gaps is slightly smaller than in schools where no teaching was going on but it is still considerable with 39 percent for subtraction and 56 percent for addition.

5.3.5 Potential mechanism: Engagement in market activities

Two factors drive the skill gap: the children performed well in the market game and they performed poorly in the standard tasks. Based on the initial observation of children being engaged in market activities, we phrased a hypothesis about a potential explanation for the first driver of the skill gap. We hypothesised that children who sometimes buy or sell things in the market on their own or who help someone in selling are more able to solve tasks in the market simulation game but not necessarily more able to solve standard calculation tasks. Using a linear regression, we explore if engagement in market activities is associated with the ability of children, who cannot solve a specific calculation when presented in the standard form, to solve it when the same task is presented as a market transaction:

$$Y_{i,s} = \alpha + \beta_1 market\ activity_{i,s} + \beta_2 X_{i,s} + \beta_3 E_{i,s} + \gamma_s + \varepsilon_{i,s}$$

with $Y_{i,s}$ being a dummy variable equal to 1 if child i in school s was able to solve the market simulation task but not the standard mathematics task. The main explanatory variable, $market\ activity_{i,s}$, is a dummy variable indicating whether or not the child is engaged in market activities. For the skill gap in addition, we use $buys\ in\ market_{i,s}$, a dummy variable equal to 1 if child i answered Yes to the question “Do you sometimes buy things in the market on your own?”. For the skill gap in subtraction, we use the variable $sells\ in\ market_{i,s}$, a dummy variable equal to 1 if child i answered Yes to the question “Do you sometimes sell things in the market or help someone to sell?”. $X_{i,s}$ is a set of child characteristics such as gender and age. $E_{i,s}$ captures the share of children interviewed by the same enumerator who correctly solved the standard math task excluding child i to control for enumerator effects. We further include school fixed effects γ_s and use robust standard errors.

In this analysis, we exclude children who were able to solve the formal item. We therefore only try to explain the ability to solve the market item among those not able to solve the formal item. The inclusion of control variables reduces the sample by 30 percent, especially driven by missing information on age or age reported outside the range 5 and 15 (see Figure A 5.1 for details). However, the remaining sample remains similar to the full sample, including the shares of children showing skill gaps in addition and subtraction. We therefore first estimate the association between engagement in market activities and ability to solve the market simulation task for the full sample, including only enumerator controls and school fixed effects. We then restrict the sample to those with non-missing control variables in all other specifications to allow for a direct comparison of coefficients.

In the following, we focus on results for the task mimicking selling one item as this is the more complex task and may be influenced more by market activities of the child than the task mimicking buying two items. Results for the task of buying two goods are presented in the appendix.

Table 5.2: Regression results for skill gap in subtraction

	(1)	(2)	(3)	(4)
sells in market	0.124*** (9.15)	0.137*** (8.43)	0.0704*** (4.32)	0.0630*** (3.89)
female			-0.0131 (-0.91)	-0.0126 (-0.88)
age			0.0411*** (10.51)	0.0369*** (9.40)
grade			0.0629*** (5.24)	0.0600*** (5.04)
studies at home			0.0902*** (5.72)	0.0735*** (4.64)
listening comprehension				0.122*** (6.85)
solves after reading out				0.0829** (3.09)
points to numbers only				-0.0105 (-0.29)
Mean dep. var.	0.46	0.50	0.50	0.50
Observations	5559	3904	3904	3904
Adj. R^2	0.24	0.24	0.30	0.31

*Note: The table presents results of a linear regression with robust standard errors, for subtraction tasks. Dependent variable: Dummy equal to 1 if child was able to solve market simulation task of selling one good. The analysis only includes children that were not able to solve formal subtraction task with numbers up to 100. All columns include school fixed effects and enumerator controls. T-statistics are given in brackets below coefficients. P-values: * < 0.05, ** < 0.01, *** < 0.001.*

Table 5.2 shows results for the outcome being a dummy equal to 1 if the child was able to solve the task of selling one item in the market simulation game, despite not being able to solve the task of subtraction with numbers up to 100. The corresponding results for buying two goods are shown in the appendix (Table A 5.1). Column (1) shows the association between the main explanatory variable $sells\ in\ market_{i,s}$ and the outcome, including only enumerator controls and school fixed effects. It shows that among children engaged in market activities, 12.4 percentage point more children were able to solve the task of selling an item. In column (2), we restrict the sample to those with non-missing control variables. This changes the association to 13.7 percentage points. In the next column, basic child characteristics, gender, age in years, grade, and an indicator whether the child reports to study at home are added.

Older age and being in a higher grade increase the probability of being able to solve the task correctly as is reflected in the Figure 5.2 and Figure A 5.2 above. Gender does not seem to play a role. Surprisingly, the coefficient of the variable indicating whether a child studies at home is positive and statistically significant. This could mean that being engaged in market activities is also an indicator of belonging to a wealthier household that might also put more value on education. However, lacking information on household wealth, we can only speculate about this. The main explanatory variable remains highly significant with a coefficient of 0.07.

The specification in column (4) includes controls for potential design effects. We add an indicator for being able to solve the task of listening comprehension. This task was the other non-standard task of the learning assessment, which could also be viewed as having a similar game-like character as the market simulation. In this task, the enumerator read out a story and the child was asked to retell it by ordering flash cards with pictures describing the story in the correct order. This indicator could be a proxy for the child's general ease in solving the non-standard tasks of the learning assessment. Indeed, with a coefficient of 12.2 percentage points it is a strong predictor for the ability to solve the market game. The ability to solve these non-standard tasks could be linked to a better ability to concentrate, or the enumerator being more careful in conducting the learning assessment, including explaining the non-standard tasks. To investigate the effect of children' inability to recognize numbers and potential discomfort in speaking out loud the variables *solves after reading out* and *points to numbers only* are added. *Points to numbers only* is a dummy equal to 1 if the child was able to point correctly to the announced number out of four presented on flash cards but not able to name a presented number correctly, both in the number range up to 10. The coefficient is small and not statistically significant, indicating that being shy to speak does not explain the skill gap. *Solves after reading out* is a dummy equal to 1 if the child was able to solve the subtraction task after the task was read out by the enumerator. This coefficient is large and positive, showing that the market game picks up arithmetic skills for children that struggle with standard tasks, potentially because they had difficulties reading numbers or for any other reason needed a second attempt. However, only 284 children fall into this category. The coefficient of the main explanatory variable *sells in market* changes little, to 6.3 percentage points.

Findings for the task of buying two items relating to the skill gap in addition are very similar to those for subtraction (see Table A 5.1).

We also repeat the regression analysis with the samples restricted to observations with proper GPS information, those conducted during school hours, and those with an interview duration of at least 10 minutes. We find the coefficients change little despite the large reduction in sample size (see Table A 5.2).

The findings are very robust to the use of alternative specifications of the skill gap. First, we use a slightly stricter definition of the skill gap. We exclude from the sample those children who solved the standard task of subtraction up to 100 when the task was read out, who were therefore able to solve the standard task in a second attempt. The coefficient remains unchanged (see first row of Table A 5.3). We then repeat specification (3), both for the sample with non-missing controls and the sample of children excluding those solving the task in a second attempt, for alternative skill gaps (see Table A 5.3). These are combinations of the market games buying two goods, selling one good and selling two goods with the standard addition and subtraction tasks in the number range up to 10 and up to 100. The displayed main explanatory variable is $sells\ in\ market$ for the market games selling one or two goods and $buys\ in\ market$ for the market game buying one good. Changing the explanatory variables and using $sells\ in\ market$ to predict the outcome of the task of buying one good and $buys\ in\ market$ to predict the outcome of the selling tasks does not alter the results substantially.

We conclude that engagement in market activities seems to explain at least part of the finding that children were able to solve calculations phrased as market transactions while they were not able to solve tasks on a comparable level of difficulty phrased as formal mathematics. The finding suggests that children have a very different understanding of the numbers they regularly use in daily life and the numbers they are confronted with at school. These concepts are not linked and therefore, skills learned through engagement in market activities are not transferred to tests of formal mathematics. This mirrors findings from studies investigating conceptualisations of numbers in children (Khan, 2004; Sitabkhan, 2011; Spinillo, 2018). The full set of arithmetic skills is therefore likely to be strongly underestimated by standardised tests such as ASER as they only capture skills in formal mathematics. Sitabkhan et al. highlight this by suggesting that standardised tests should be supplemented by observations of children’s mathematical abilities in real-life settings (Sitabkhan et al., 2018).

5.3.6 Factors associated with poor performance in formal tasks

While the skill gap is considerable due to strong performance in the market simulation tasks, it is also driven by poor performance in the standard tasks. We therefore try to identify factors associated with poor performance in formal tasks for this sample of children. We estimate linear probability models for indicators that the child was able to solve different formal tasks, including basic numeracy, addition up to 10 and 100, and subtraction up to 10 and 100. Control variables include child and school characteristics. Results are presented in Table A 5.4. We find that the probability that a child is able to solve the formal tasks increases with age and grade, for the latter only in the number range up to 10, while it is lower for girls compared to boys. If a child reports to study at home, the probability of being able to solve the formal tasks is higher. This is also the case if the child is able to solve the listening comprehension task, an indicator of the child's ease to manage unusual question formats. We use an indicator of the childing naming anyone as best reader in class as a proxy of regular attendance at school. This indicator is positively associated with the probability of solving the formal tasks for three of the five outcomes. The presence of any teacher at school at the time of the unannounced visit, an indicator of a functioning school, is associated with a higher probability that a child is able to solve the formal tasks. Availability of a toilet at school is also associated with a higher probability of being able to solve the formal tasks, as is the indicator of the school having any building of permanent structure. There is no association with the availability of blackboards in the classrooms. Furthermore, we do not find any association between engagement in market activities and children's ability to solve the formal tasks. This suggests that the concept of numbers used in market transactions is not transferred to the concept of numbers used in formal mathematics at school, but also that engagement in market activities does not affect children's performance in these tasks negatively.

5.4 Discussion

Our study describes a large gap between skills in formal and informal mathematics among primary school children in Sokoto State, Nigeria. Our study contributes to the literature by illustrating this skill gap in a large sample of school-aged children, testing both skills in each child. We explore several possible explanations for this finding and conclude that engagement in market activities can partly explain this finding. As we find the skill gap also among children who do not report to engage in market activities, further mechanisms have to be at play as well. Still, buying and selling goods in the market seems to help children gain arithmetic skills that are not transferred to comparable tasks in school. A study among pre-school children in India shows that gaining informal numerical skills does not automatically improve skills in formal mathematics (Dillon et al., 2017). While our study was implemented in a specific setting, characterized by high poverty rates and poor educational infrastructure, we believe that similar results would be found in other settings, likely where children engage in some form of labour. In fact, our findings confirm evidence from smaller, qualitative studies, mostly among working children, (Carraher et al., 1985, 1987; Khan, 2004; Nunes et al., 1993; Sitabkhan, 2011) and one quantitative study conducted among working children in India (Banerjee et al., 2017).

The most recent systematic assessment of learning outcomes in Sokoto State was conducted in 2013 with the Early Grade Mathematics Assessment (Nigeria Northern Education Initiative (NEI), 2013). The schools included in our sample seemed to be slightly worse off than the schools included in EGMA 2013. However, the EGMA schools were also characterized by poor infrastructure and high absenteeism. Performance in basic mathematic tasks was slightly better in our sample than in EGMA 2013. The most recent national assessment, the Nigeria Education Data Survey (NEDS) 2015 determined numeracy competency with items of one-digit and two-digit addition and subtraction (Adeniran et al., 2020). Nationwide, 35 percent, 49 percent, and 61 percent of children in primary 2, 3, and 4, respectively, were able to solve a double-digit addition or subtraction task (USAID et al., 2015). Developing a quality of education indicator based on the NEDS 2015, Adeniran and colleagues define the pass rate as the percentage of primary 1 and primary 2 children who are able to solve a one-digit and two-digit addition or subtraction task, respectively. While this pass rate is 31 percent nationwide, it is only 8.6 percent in the North-West region, which includes Sokoto State. They point out that the pass rate is lowest for children in rural areas, those in government schools, from poorer households, and in the northern regions of Nigeria (Adeniran et al., 2020). These are exactly the characteristics of children in this study.

This study reports that there is a large share of school-aged children who are unable to solve basic calculations in a format of formal mathematics but are able to solve more complex tasks in a format of market transactions. The findings suggest that standardised tests are likely to strongly underestimate actual arithmetic skills in these children as they only focus on formal mathematics. In order to capture true arithmetic skills relevant for their daily life, standardised tests could be expanded to include tasks phrased as actual transactions in addition to abstract, formal mathematics. Moreover, current curricula and teaching strategies do not seem to sufficiently build on and utilize existing skills to teach concepts assessed in the classroom. They miss out on potential opportunities to achieve better learning outcomes. Curricula should be re-worked and adjusted more strongly to the needs and existing skills of the children. Understanding the different way of learning mathematics and including this in curricula might also provide an opportunity for children struggling with the conventional teaching of mathematics.

5.5 Appendix A: Tables and figures

Table A 5.1: Regression results for skill gap in addition

	(1)	(2)	(3)	(4)
buys in market	0.108***	0.119***	0.0620***	0.0569***
	(8.47)	(7.93)	(3.99)	(3.67)
female			-0.00802	-0.00828
			(-0.57)	(-0.59)
age			0.0279***	0.0251***
			(7.31)	(6.52)
grade			0.0433***	0.0419***
			(3.96)	(3.83)
studies at home			0.0821***	0.0712***
			(5.43)	(4.65)
listening comprehension				0.0918***
				(5.83)
solves after reading out				0.0285
				(1.52)
points to numbers only				-0.0337
				(-0.90)
Mean dep. var.	0.67	0.70	0.70	0.70
Observations	5345	3717	3717	3717
Adj. R^2	0.21	0.20	0.24	0.24

*Note: The table presents results of a linear regression with robust standard errors, for addition tasks. Dependent variable: Dummy equal to 1 if child was able to solve market simulation task of buying two goods. The analysis only includes children that were not able to solve formal addition task with numbers up to 100. All columns include school fixed effects and enumerator controls. T-statistics are given in brackets below coefficients. P-values: * < 0.05, ** < 0.01, *** < 0.001.*

Table A 5.2: Regression results of robustness checks, data quality

	(1)	(2)	(3)
Skill gap subtraction	0.0821***	0.0987***	0.0943**
	(4.35)	(3.68)	(2.89)
N	2830	1582	1148
Skill gap addition	0.0424*	0.0631*	0.0420
	(2.33)	(2.57)	(1.42)
N	2720	1506	1092
Sample	Proper GPS, 8-13.00	>= 10 min	Proper GPS, 8-13.00, >= 10 min

Note: The table presents coefficients of the variable market activity (sells in market and buys in market) from a linear regression with robust standard errors. Dependent variable: Dummy equal to 1 if child was able to solve market simulation task of selling one good (buying two goods) but was not able to solve the standard subtraction (addition) task with numbers up to 100. All specifications include controls for gender, age, grade, studies at home, listening comprehension, solves after reading out, points only, school fixed effects, and enumerator control. Column (1) restricts the sample to observations with proper GPS location and time during school opening hours; column (2) restricts the sample to observations with interview duration of 10 or more minutes; column (3) restricts sample to observations with proper GPS location, time during school opening hours, and interview duration of 10 or more minutes.

Table A 5.3: Regression results of robustness checks, alternative variable definition

	(1)	(2)
sell1_minus100	0.0704***	0.0754***
	(4.32)	(4.36)
N	3904	3660
sell1_minus10	0.0813***	0.0895***
	(4.43)	(4.41)
N	3304	2809
sell1_add100	0.0780***	0.0781***
	(4.57)	(4.29)
N	3738	3417
sell1_add10	0.0829***	0.0725**
	(4.00)	(3.22)
N	2776	2315
buy2_minus100	0.0289	0.0486**
	(1.73)	(2.84)
N	3884	3645
buy2_minus10	0.0547**	0.0466*
	(2.77)	(2.19)
N	3286	2797
buy2_add100	0.0620***	0.0687***
	(3.99)	(4.10)
N	3717	3402
buy2_add10	0.0735***	0.0670**
	(3.84)	(3.07)
N	2760	2303
sell2_minus100	0.0638***	0.0708***
	(3.84)	(4.02)
N	3904	3660
sell2_minus10	0.0740***	0.0846***
	(3.95)	(4.09)
N	3304	2809
sell2_add100	0.0723***	0.0740***
	(4.18)	(3.99)
N	3738	3417
sell2_add10	0.0769***	0.0672**
	(3.64)	(2.91)
N	2776	2315
Skill gap	standard	excl. reading out

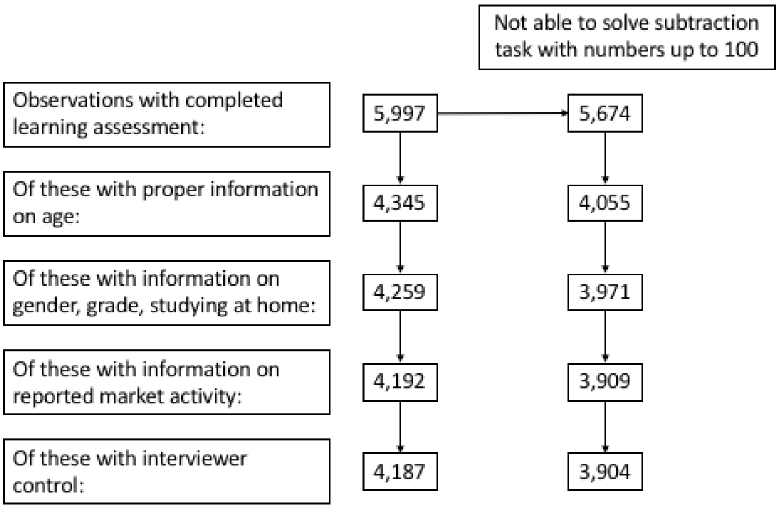
*Note: The table presents coefficients of the variable market activity (sells in market and buys in market) from a linear regression with robust standard errors. Sells in market is a dummy equal to 1 if the child reported to sell or help someone in selling on the market. Buys in market is a dummy equal to 1 if the child reported to buy on the market on his/her own. Each row shows the coefficient from a regression with a different outcome variable specified by the label. All specifications include controls for gender, age, grade, studies at home, school fixed effects and enumerator control. Column (2) excludes children who were able to solve the standard task only after it was read out to them by the enumerator. T-statistics are given in brackets below coefficients. P-values: * < 0.05, ** < 0.01, *** < 0.001.*

Table A 5.4: Regression results for learning outcomes

	(1)	(2)	(3)	(4)	(5)
	Basic numeracy	Add up to 10	Add up to 100	Minus up to 10	Minus up to 100
age	0.041***	0.028***	0.014***	0.018***	0.010***
	(13.42)	(9.13)	(6.70)	(6.71)	(5.56)
grade	0.030**	0.037***	0.001	0.024**	-0.000
	(3.21)	(3.88)	(0.14)	(2.78)	(-0.10)
female	-0.075***	-0.032**	-0.025**	-0.018	-0.016*
	(-6.09)	(-2.60)	(-2.88)	(-1.64)	(-2.21)
studies at home	0.154***	0.115***	0.049***	0.083***	0.028***
	(10.83)	(7.96)	(5.20)	(6.57)	(3.59)
sells in market	-0.017	0.005	0.014	-0.014	0.009
	(-1.19)	(0.32)	(1.38)	(-1.14)	(1.00)
listening comprehension	0.225***	0.201***	0.124***	0.181***	0.092***
	(12.90)	(12.12)	(9.39)	(11.09)	(8.08)
knows best reader	0.029	0.050**	0.021*	0.030*	0.005
	(1.96)	(3.19)	(2.32)	(2.23)	(0.63)
presence of teacher	0.056***	0.030*	0.038***	0.027*	0.021*
	(4.16)	(2.20)	(3.90)	(2.20)	(2.57)
toilet availability	0.128***	0.096***	0.042***	0.069***	0.043***
	(9.21)	(7.15)	(4.24)	(5.75)	(4.98)
any classroom with blackboard	-0.008	0.002	-0.021	-0.005	-0.001
	(-0.47)	(0.10)	(-1.77)	(-0.36)	(-0.11)
any building of permanent structure	0.048**	0.032	0.034***	0.030*	0.017*
	(2.99)	(1.85)	(3.48)	(2.05)	(2.35)
Mean dep. var.	0.30	0.34	0.11	0.21	0.07
Adj. R^2	0.28	0.33	0.20	0.28	0.13

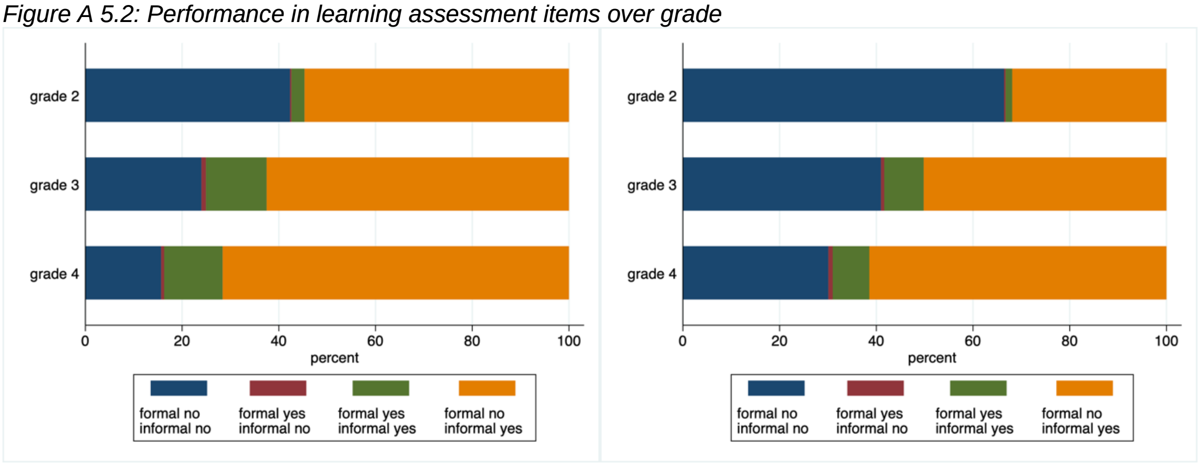
*Note: The table presents results of a linear regression with robust standard errors, for 5 different learning outcomes. N=4083. All regressions include school fixed effects and enumerator controls. T-statistics are given in brackets below coefficients. P-values: * < 0.05, ** < 0.01, *** < 0.001.*

Figure A 5.1: Sample size


Not able to solve subtraction task with numbers up to 100
Observations with completed learning assessment:
5,997
5,674
Of these with proper information on age:
4,345
4,055
Of these with information on gender, grade, studying at home:
4,259
3,971
Of these with information on reported market activity:
4,192
3,909
Of these with interviewer control:
4,187
3,904

Note: Number of observations and reasons for excluding observations. Starting point is the full sample, end point is the analysis sample used in regressions.

Figure A 5.2: Performance in learning assessment items over grade

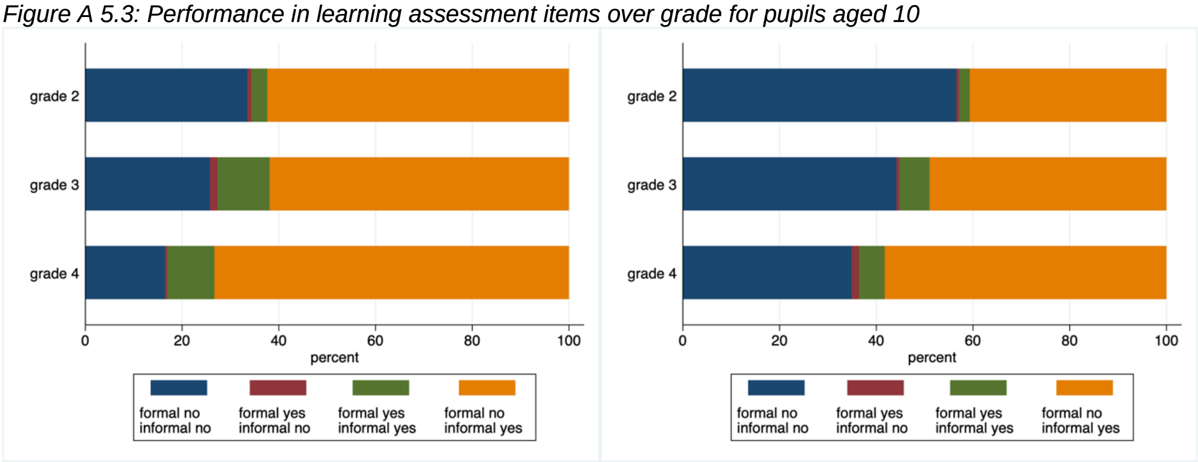


(a) Addition

(b) Subtraction

Note: Performance in learning assessment for each grade. N = 5,974. Panel (a) contains tasks related to addition. Panel (b) contains tasks related to subtraction. "formal" refers to standard calculation tasks of addition (subtraction) up to 100, presented as formal mathematics. "informal" refers to calculations up to 100 presented as market transactions. Bars show shares of pupils in the following categories: (1) pupils who could solve neither standard tasks nor the market simulation; (2) pupils who could solve standard tasks, but not the market simulation; (3) pupils who could solve both standard tasks and the market simulation; and (4) pupils who could not solve standard tasks, but could solve the tasks of market simulation.

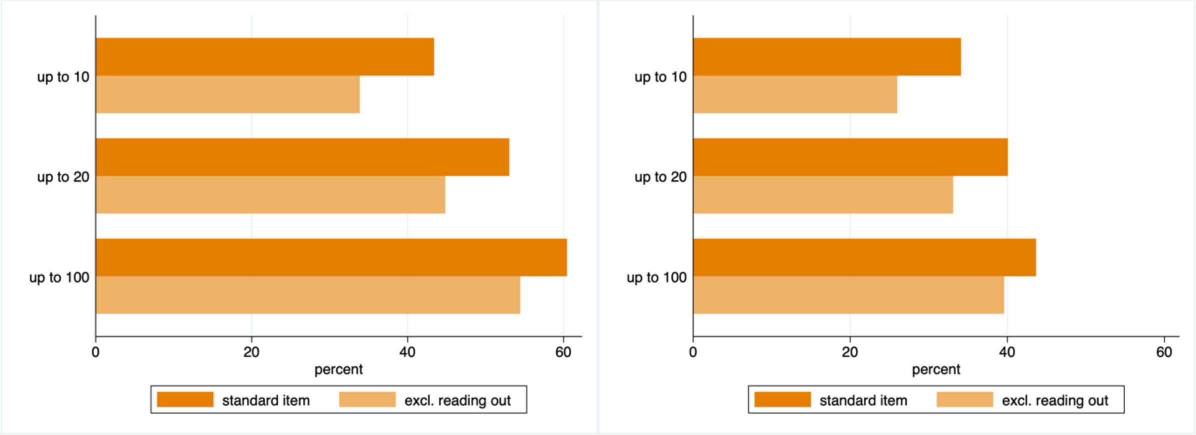
Figure A 5.3: Performance in learning assessment items over grade for pupils aged 10



(a) Addition *(b) Subtraction*

Note: Performance in learning assessment for each grade in sub-sample of pupils aged 10 years. N = 950. Panel (a) contains tasks related to addition. Panel (b) contains tasks related to subtraction. "formal" refers to standard calculation tasks of addition (subtraction) up to 100, presented as formal mathematics. "informal" refers to calculations up to 100 presented as market transactions. Bars show shares of pupils in the following categories: (1) pupils who could solve neither standard tasks nor the market simulation; (2) pupils who could solve standard tasks, but not the market simulation; (3) pupils who could solve both standard tasks and the market simulation; and (4) pupils who could not solve standard tasks, but could solve the tasks of market simulation.

Figure A 5.4: Robustness of skill gap



(a) Addition *(b) Subtraction*

Note: Share of pupils with skill gap, N = 5,997. Panel (a) contains tasks related to addition. Panel (b) contains tasks related to subtraction. "standard item" shows the share of pupils with a skill gap (i.e. able to solve the market item of buying two goods or selling one good but not the respective standard item in the number range up to 10, up to 20, up to 100) when the standard task is presented either without or with reading out the task. "excl. reading out" shows the share of pupils with a skill gap when the standard item was also read out.

5.6 Appendix B: Learning assessment

MANUAL FOR LEARNING ASSESSMENT (Version 09.07.2018)

GENERAL REMARKS

- Have your material prepared and ordered
- Place the material to your side or behind you and only place one item in front of the pupil at a time
- Be patient and friendly
- Encourage the pupil to try, praise and acknowledge the effort
- Keep your attention with the pupil

I) NUMERACY

1. Understanding of numbers

- Counters:
    - Note: for the "can you give me X counters"-task: ask the child to immediately count / put the counters in your hand (to avoid counting twice)
    - Help: encouragement only
    - Coding: correct if correct number
    - Correction:
        - ask pupil to have another look – give some time
        - give correct answer
        - do NOT count if the first item is wrong
        - if the second item is wrong, count together with the child
- Recognising numbers
    - Note:
        - present the numbers such that the number is turned for the pupil to read
        - for the task to pick a certain number out of a set of 5:
          do NOT present the flash card in a sequential order
    - Help:
        - if the pupil turns it around, turn it back
        - for two-digit numbers: if the pupil says "one seven", you can ask "and which number is this" to give a second chance
    - Coding:
        - correct if correct number
        - two-digit numbers: if the pupil says "seventeen" after given second chance, code as correct, if pupil continues saying "one seven" code as wrong
    - Correction: give correct answer
- Ordering numbers
    - Help: none
    - Coding:
        - correct if correct order
        - it does not matter whether
          - the cards are presented from the smallest to the highest number or from the highest to the smallest number
          - numbers are upside down
          - 6 and 9 are mixed up
    - Correction:
        - have another look – give some time
        - do NOT show correct order with flash cards for the first item
        - for the second item, you can order together with the pupil
- Writing numbers
    - Help: none
    - Coding:
        - correct number
        - (correct order, reverse digits; for two-digit numbers only)
        - correct number but with writing mistake / mirrored
        - wrong number
        - nothing / scribbling
    - Correction:
        - for the two-digit number ask the pupil to have a look, read it to her/him showing the order of reading / writing
        - write the correct number on the paper and show it to the pupil

2. Basic calculations

- Solving basic calculations
    - Note: give flash card such that exercise is turned for the pupil to read
    - Help:
        - at beginning remind pupil that he/she may use his/her fingers (one time)
        - DO NOT read out the exercise to the pupil
        - if the pupil shows the number with the fingers but doesn't say anything, ask him/her to say the number
    - Coding: correct if correct answer is given on the first try
    - Correction:
        - give pupil a second chance, then tell him/her the correct answer
        - if pupil takes the sum in one of the subtraction tasks, point out the minus (-) and ask to try again
        - give correct answer
        - do NOT explain how to solve the two-digits numbers for the first item

3. Money

- Pupil is buyer
    - Note:
        - be careful to give pupil the correct number of banknotes
        - lay out shopping items in front of you (display your goods)
        - push the good that the pupil is buying toward the pupil
        - make sure to repeat the price
            - "I sell pencils and exercise books. The pencils cost 10 Naira, the exercise books 25 Naira. Now you want to buy a pencil for 10 Naira"
            - "Now you want to buy two exercise books. Each one costs 25 Naira, you want to buy two."
    - Help: you may repeat the price of the good
    - Coding:
        - correct if correct amount is paid
        - it does not matter if 10 Naira are paid with one 10 Naira note or two 5 Naira notes
    - Correction: help pupil select correct banknotes
- Pupil is vendor
    - Note:
        - be careful to give pupil the correct number of banknotes
        - lay out shopping items in front of pupil (display his goods)
        - push the good that you are buying toward yourself
        - make sure you repeat the price
            - "Now you are the vendor. You are selling oranges and mangos. One orange costs 20 Naira, one Mango costs 35 Naira. Now I want to buy one mango for 35 Naira."
            - "Now I want to buy one orange for 20 Naira and a mango for 35 Naira"
        - instead of giving the note to the pupil, place it in front of her/him
        - tell the pupil how much you are giving her/him ("I give you 50 Naira and want to collect my change")
    - Help: you may repeat the price of the good
    - Coding: correct if correct change is given
    - Correction: help pupil select correct banknotes

II) LANGUAGE (Hausa)

1. Recognizing and naming letters

- Letters
    - Note: give flashcards such that the letter is turned for the pupil to read
    - Help:
        - if the pupil turns the card around, turn it back
        - if the pupil gives you a number, point out that you are now dealing with letters and not numbers
    - Coding:
        - correct if correct letter is identified correctly
        - English name, Hausa name or sound are all correct
    - Correction: say correct letter

2. Reading aloud

- Syllables
    - Note: give flashcards such that the letters are turned for the pupil to read
    - Help:
        - remind pupil once that he/she may use his/her fingers to read
        - if pupil calls letters separately, ask him/her to read the letters together, to make the sound they produce together
    - Coding:
        - correct if correct syllable, both letters read out together; both English and Hausa pronunciation is correct
        - note if correct in second attempt, after telling pupil to combine letters
    - Correction:
        - read the syllable for the pupil, using your finger
- Words
    - Note:
        - give flashcards such that the words are turned for the pupil to read
        - if the pupil reads the word in English, ask her/him to try again in Hausa, do not count the first attempt
    - Help: remind pupil once that he/she may use his/her fingers to read
    - Coding: four answer options
        - correct: this also applies if pupil reads the word in English at first but then correctly in Hausa when asked to do so
        - with mistakes: one vowel or consonant is wrong (i.e. rakomi instead of rakumi)
        - only first letter / syllable recognized
    - Correction:
        - give pupil a second chance (after coding), telling him the first syllable or point out any mistake
        - read the word for the pupil, using your finger
- Sentence
    - Note: give flashcards such that the sentences are turned for the pupil to read
    - Help:
        - remind the pupil that he/she may use his/her finger to read
        - DO NOT correct the pupil while reading
        - just encourage: ‘Go on’
    - Coding: three answer options
        - correct: One small mistake is still “correct” (e.g. yane instead of yana)
        - with mistakes: a word is read wrongly or several small mistakes
        - wrong: several words are not identified correctly / many small mistakes
    - Correction:
        - give pupil a second chance (after coding), helping with a problematic word
        - read the sentence for the pupil, using your finger
- Paragraph
    - Note: give flashcards such that the paragraph is turned for the pupil to read
    - Help:
        - none
        - DO NOT correct the pupil while reading
        - just encourage: ‘Go on’
    - Coding: four answer options
        - correct: still correct if one word is read wrongly or two small mistakes
        - with mistakes: means that several words are read wrongly or many small mistakes
        - could not finish: pupil started, but got stuck after the first sentence is finished; he/she does not continue even after encouragement
        - wrong: more than 1/3 of words have mistakes
    - Correction:
        - give pupil a second chance (after coding), helping him where he/she got stuck
        - read the paragraph for the pupil, using your finger

3. Reading comprehension

- Matching words with pictures
    - Note:
        - give flashcards such that the word is turned for the pupil to read and place the pictures around that word
        - pupil does not have to read the word out loud
    - Help:
        - remind pupil that he/she does not need to read the word out loud
        - DO NOT correct the pupil while reading
    - Coding: correct, if correct picture is picked
    - Correction:
        - Tell pupil the word that is written on the card and let the pupil give you the correct picture

4. Listening comprehension

- Story
    - Note:
        - give flashcards only after the story was read
        - give flashcards in random order and spread them unorderly in front of pupil
        - ask pupil to tell you ‘I am finished’ when he/she has finished the task
    - Help: None
    - Coding: correct, if all cards are placed in correct order
    - Correction:
        - place the pictures in the correct order and retell the story in a few words

5. Writing

- Writing name
    - Note: give pen and paper to pupil
    - Help: None
    - Coding: four answer options
        - Correct: any spelling of the name counts (e.g. Lisa and Liza)
        - Spelling mistakes
        - Fragment: fragment of the name is recognizable
        - Nothing /scribbling / random letters
    - Correction:
        - DO NOT correct, just encourage
- Writing words
    - Note:
        - first ask the pupil what he/she sees in the picture and correct him, if wrong, so that he/she writes the correct word
        - if the pupil tells the English name, ask for the Hausa name (e.g. mota, not car)
    - Help: None
    - Coding: five answer options
        - correct: small or capital letters are both correct, even a mix
        - spelling mistakes: a letter is missing or letters are in the wrong order or letters are mirrored
        - fragment: the pupil writes part of the word which already allows you to recognise the word (e.g. MT for Mota)
        - only first letter correct
        - nothing / scribbling / random letters
    - Correction:
        - show correct spelling by writing the word for the pupil

TESTING LEARNING OUTCOMES – TEST DESCRIPTION

Duration: 30 minutes per pupil

I) NUMERACY

1. Understanding of numbers

- Some counters (bottle tops) are displayed to the pupil and he / she is asked to count them.
  - First round: **6** (number in the range [1 to 10]) (if wrong: repeat, **7**)
    Second round: **14** (number in the range [10 to 20]) (if wrong: repeat, **15**)
- 20 counters are presented to the pupil (in a pile). The pupil is asked to give X counters to the enumerator.
  - First round: **8** (number in the range [1 to 10]) (if wrong: repeat, **9**)
    Second round: **12** (number in the range [10 to 20]) (if wrong: repeat **13**)
- A flash card with a number is displayed to the pupil. The pupil is asked to name the number.
  - First round: **4** (number in the range [1 to 10]) (if wrong: repeat, **5**)
    Second round: **17** (number in the range [10 to 20]) (if wrong: repeat, **18**)
- 5 flash cards with a number on each are presented to the pupil. The pupil is asked to give the card with the number X to the enumerator.
  - First round: **1, 3, 5, 7, 9** (number in the range [1 to 10]) (if wrong: repeat, **2, 4, 6, 8, 9**)
    Second round: **12, 14, 16, 18, 20** (number in the range [10 to 20]) (if wrong: repeat, **11, 13, 15, 17, 19**)
- 5 flash cards with a number on each (sequential numbers) are presented to the pupil. The pupil is asked to order the cards going from the lowest to the highest number.
  - First round: **3, 4, 5, 6, 7** (number in the range [1 to 10]) (if wrong (but not completely): repeat, **5, 6, 7, 8, 9**)
    Second round: **10, 11, 12, 13, 14** (number in the range [10 to 20]) (if wrong: repeat, **14, 15, 16, 17, 18**)
- The pupil is asked to write number X on the sheet of paper.
  - First round: **2** (number in the range [1 to 10]) (if wrong: repeat, **3**)
    Second round: **14** (number in the range [10 to 20]) (if wrong: repeat, **15**)
    Third round: **76** (number in the range [30 to 100]) (if wrong: repeat, **87**)

2. Basic calculations

- One after the other, calculation tasks are presented to the pupil and he/she is asked to solve the task and give the result.
  - a) Addition up to 10: 3 + 4 (if wrong: repeat, 5 + 2)
  - b) Subtraction up to 10: 8 – 3 (if wrong: repeat, 9 – 4)
  - c) Addition up to 20: 8 + 6 (if wrong: repeat, 7 + 8)
  - d) Subtraction up to 20: 13 – 5 (if wrong: repeat, 15 – 6)
  - e) Addition up to 100: 35 + 25 (if wrong: repeat, 45 + 15)
  - f) Subtraction up to 100: 100 – 65 (if wrong: repeat, 100 – 75)
- For each exercise: Only one item per exercise is presented. In case the item is solved correctly, the enumerator proceeds to the next exercise. In case this item is not solved correctly, another item of the same exercise is presented.
  If the pupil cannot solve the second item
  - in a) → no other exercise presented
  - in b) → move to c)
  - in c) → no other exercise presented
  - in d) → move to e)
  - in e) → no other exercise presented
  - in f) → no other exercise presented

3. Money

- A “real world” shopping problem with token money is presented to the pupil:
- Pupil receives the following banknotes: 1x50 Naira, 1x20 Naira, 2x10 Naira, 3x5 Naira (enumerator keeps 1x100 Naira, 2x10 Naira, 1x5 Naira). Flash cards with pens and exercise books are laid down.
- Enumerators says “A pen costs 10 Naira, a exercise book costs 25 Naira.”
- Enumerator says “Buy one pen and give the correct banknote” (only one note is possible).
- Enumerator says “Buy two exercise books and give the correct banknote” (two banknotes or one note are possible options).
- Now roles are switched. The pupil receives the following banknotes: 1x20 Naira, 4x10 Naira, 4x5 Naira (enumerator keeps 1x100 Naira, 1x50 Naira). Flash cards with oranges and mangoes are laid down.
- Enumerator says “A orange costs 20 Naira, a mango costs 35 Naira.”
- The enumerator buys one mango with one banknote (50) that is too high and the pupil is asked to give change.
- The enumerator buys one mango and an orange with one banknote (100) that is too high and the pupil is asked to give change.

II) LANGUAGE (Hausa)

1. Recognizing and naming letters

- One after the other, four letters are presented to the pupil on a flash card, the pupil is asked to name the letters (name in English or Hausa or just the sound is okay) **a, e, m, d.**

2. Reading aloud

- One after the other, four syllables in increasing difficulty are presented to the pupil on a flash card, the pupil is asked to read out the syllables **ka, fi, lo**, **gwa.**
- One after the other, three words in increasing difficulty are presented to the pupil on a flash card, the pupil is asked to read out the words **keke** (bicycle), **rakumi** (camel)**, ciyawa** (grass).
- One after the other, three nonsense words are presented to the pupil on a flash card, the pupil is asked to read out the words **lela, tumari, jirgoya.**
- One sentence is presented to the pupil on a flash card, the pupil is asked to read it out.
- One paragraph is presented to the pupil on a flash card, the pupil is asked to read it out.

3. Reading comprehension

- A flash card with a word and 4 cards with pictures are presented to the pupil, one of which is the picture corresponding to the words. The pupil is asked to pick the card with the picture corresponding to the word. (3 Rounds)
  kaza (chicken; bike, carrot, bucket), **ƙwallo** (ball; shoe, chair, book), **makullai** (keys; mango, butterfly, cat)

4. Listening comprehension

- A short story is read out to the pupil. Afterwards, 5 cards with pictures that somehow retell the story are given to the pupil and he/she is asked to sort them in the right order.

Aisha da Amina yan uwan junane. Sukan tafi Makaranta tare, kuma ajinsu daya. Bayan an tashi daga Makaranta a hanyar su zuwa gida Sun hadu da rakumi a kwance ga hanya.
Lokacin da suka isa gida sun samu uwayensu suna dafa abinchi. Sai a ka aikesu kasuwa su sayo tomatur, da uwayensu suka manta su sayo. Bayan an kare abinchi aka basu nasu suka ci, San nan sukaje wurin wasa. Dadare yayi sukayi kwana.

5. Writing

- The pupil is asked to write his / her name on a sheet of paper (Lisa/Liza is both okay)
- The pupil is presented 2 or 3 cards with pictures and he / she is asked to write the corresponding word on the paper underneath the name
  mota (car), **tebur** (table), **kujera** (chair)

NIPEP IMPACT EVALUATION - STUDENT QUESTIONNAIRE AND STUDENT TEST

0. Interview information I		
DateofInterview	Date of interview	_________
StartTime	Start time	_________
SchoolName	School Name	_________
SchoolID	School ID	_________
Village	Village Name	_________
LGAName	LGA Name	_________
LGAID	LGA ID	_________
	Hello, my name is [name of enumerator]. What is your name?	
PUP01	*Note: write down the full name.*	______________________________
I am part of a research team and we are trying to learn how schools in Sokoto work. For this, we have prepared a few questions and exercises for pupils like you. We would like your help in answering these questions and exercises. This is not a test, you will not be graded. Nobody in school and also not your parents will get to know your answers. We are interested to learn what pupils in Sokoto learn in school. If at any point you don’t want to continue we can stop this.		
Q_10	**Would you like to help us?**	_____ 01 Yes 02 No *(no skip was implemented here)*
I. Numeracy **1. Understanding of numbers**		
STNU01	*Present **6** counters.* How many bottletops do you count? *Note down whether the answer is correct.*	_____ 01 Yes → go to STNU03 02 No
STNU02	*Present **7** counters.* How many bottletops do you count? *Note down whether the answer is correct.*	_____ 01 Yes 02 No → go to STNU09
STNU03	*Present **14** counters.* How many bottletops do you count? *Note down whether the answer is correct.*	_____ 01 Yes → go to STNU05 02 No
STNU04	*Present **15** counters.*	_____

	How many bottletops do you count? *Note down whether the answer is correct.*	01 Yes 02 No
STNU05	*Present 20 counters.* Can you give me **8** bottletops? *Note down whether the pupil picked the right number.*	______ 01 Yes → go to STNU07 02 No
STNU06	*Return the counters so there are 20 again.* Can you give me **9** bottletops? *Note down whether the pupil picked the right number.*	______ 01 Yes 02 No → go to STNU09
STNU07	*Return the counters so there are 20 again.* Can you give me **12** bottletops? *Note down whether the pupil picked the right number.*	______ 01 Yes → go to STNU09 02 No
STNU08	*Return the counters so there are 20 again.* Can you give me **13** bottletops? *Note down whether the pupil picked the right number.*	______ 01 Yes 02 No
STNU09	*Present the flash card with the number **4**.* Can you tell me which number this is? *Note down whether the answer is correct.*	______ 01 Yes → go to STNU11 02 No
STNU10	*Present the flash card with the number **5**.* Can you tell me which number this is? *Note down whether the answer is correct.*	______ 01 Yes 02 No → go to STNU13
STNU11	*Present the flash card with the number **17**.* Can you tell me which number this is? *Note down whether the answer is correct.*	______ 01 Yes → go to STNU13 02 No

STNU12	*Present the flash card with the number **18**.* Can you tell me which number this is? *Note down whether the answer is correct.*	_____ 01 Yes 02 No
STNU13	*Present the flash cards with the numbers **1, 3, 5, 7, 9**.* Can you give me the card with the number **7**? *Note down whether the pupil picked the correct card.*	_____ 01 Yes → go to STNU15 02 No
STNU14	*Present the flash cards with the numbers **2, 4, 6, 8, 9**.* Can you give me the card with the number **8**? *Note down whether the pupil picked the correct card.*	_____ 01 Yes 02 No → go to STNU17
STNU15	*Present the flash cards with the numbers **12, 14, 16, 18, 20.*** Can you give me the card with the number **16**? *Note down whether the pupil picked the correct card.*	_____ 01 Yes → go to STNU17 02 No
STNU16	*Present the flash cards with the numbers **11, 13, 15, 17, 19**.* Can you give me the card with the number **13**? *Note down whether the pupil picked the correct card.*	_____ 01 Yes 02 No
STNU17	*Present the flash cards with the numbers **3, 4, 5, 6, 7**.* Can you order the cards in the correct order from the smallest to the highest number? *Note down whether the pupil arranged the cards in the correct order.*	_____ 01 Yes → go to STNU19 02 No
STNU18	*Present the flash cards with the numbers **5, 6, 7, 8, 9**.*	_____ 01 Yes 02 No → go to STNU21

	Can you order the cards in the correct order from the smallest to the highest number? *Note down whether the pupil arranged the cards in the correct order.*	
STNU19	*Present the flash cards with the numbers **10, 11, 12, 13, 14**.* Can you order the cards in the correct order from the smallest to the highest number? *Note down whether the pupil arranged the cards in the correct order.*	_____ 01 Yes → go to STNU21 02 No
STNU20	*Present the flash cards with the numbers **14, 15, 16, 17, 18**.* Can you order the cards in the correct order from the smallest to the highest number? *Note down whether the pupil arranged the cards in the correct order.*	_____ 01 Yes 02 No
STNU21	*Hand the pupil a piece of paper.* Can you please write down number **2**? *Note down whether pupil has written* - *nothing, scribbling* - *a wrong number* - *the correct number but not correctly written (e.g. mirrored)* - *correctly written number*	_____ 01 Correct → go to STNU23 02 Correct number, writing mistakes 03 Wrong number 04 Nothing / scribbling
STNU22	*Hand the pupil a piece of paper.* Can you please write down number **3**? *Note down whether pupil has written* - *nothing, scribbling* - *a wrong number* - *the correct number but not correctly written (e.g. mirrored)* - *correctly written number*	_____ 01 Correct 02 Correct number, writing mistakes 03 Wrong number → go to STCA01 04 Nothing / scribbling → go to STCA01
STNU23	*Hand the pupil a piece of paper.* Can you please write down number **14**? *Note down whether pupil has written*	_____ 01 Correct → go to STNU25 02 Correct digits, reverse order 03 Correct number, writing mistakes

	- *nothing, scribbling* - *a wrong number* - *the correct number but not correctly written (e.g. mirrored)* - *correctly written number*	04 Wrong number 05 Nothing / scribbling → go to STCA01
STNU24	*Hand the pupil a piece of paper.* Can you please write down number **15**? *Note down whether pupil has written* - *nothing, scribbling* - *a wrong number* - *the correct number but not correctly written (e.g. mirrored)* - *reverse order* - *correctly written number*	_____ 01 Correct 02 Correct digits, reverse order 03 Correct number, writing mistakes 04 Wrong number → go to STCA01 05 Nothing / scribbling → go to STCA01
STNU25	*Hand the pupil a piece of paper.* Can you please write down number **76**? *Note down whether pupil has written* - *nothing, scribbling* - *a wrong number* - *the correct number but not correctly written (e.g. mirrored)* - *reverse order* - *correctly written number*	_____ 01 Correct → go to STCA01 02 Correct digits, reverse order 03 Correct number, writing mistakes 04 Wrong number 05 Nothing / scribbling → go to STCA01
STNU26	*Hand the pupil a piece of paper.* Can you please write down number **87**? *Note down whether pupil has written* - *nothing, scribbling* - *a wrong number* - *the correct number but not correctly written (e.g. mirrored)* - *reverse order* - *correctly written number*	_____ 01 Correct 02 Correct digits, reverse order 03 Correct number, writing mistakes 04 Wrong number 05 Nothing / scribbling
2. Basic Calculations		
STCA01	*Present flash card M1 (3+4).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes → go to STCA03 02 No

STCA02	*Present flash card M2 (5+2).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes 02 No → go to STMY00
STCA03	*Present flash card M3 (8-3).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes → go to STCA05 02 No
STCA04	*Present flash card M4 (9-4).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes 02 No
STCA05	*Present flash card M5 (8+6).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes → go to STCA07 02 No
STCA06	*Present flash card M6 (7+8).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes 02 No → go to STMY00
STCA07	*Present flash card M7 (13-5).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes → go to STCA09 02 No
STCA08	*Present flash card M8 (15-6).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes 02 No
STCA09	*Present flash card M9 (35+25).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes → go to STCA11 02 No
STCA10	*Present flash card M10 (45+15).* Can you solve this exercise?	_____ 01 Yes 02 No → go to STMY01

	Note down whether the pupil gave the correct answer.	
STCA11	*Present flash card M11 (100-65).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes → go to STMY01 02 No
STCA12	*Present flash card 12 (100-75).* Can you solve this exercise? *Note down whether the pupil gave the correct answer.*	_____ 01 Yes 02 No
3. Money		
STMY00	*Present the flash card set MY1.* *Keep the bills 100, 10, 10, 5 apart.* Now we play market. I am a vendor and sell pencils and exercise books. A Pencil costs 10 Naira, an exercise book 25 Naira. You have this money and want to buy something.	
STMY01	Now you want to buy a pen. The pen costs 10 Naira. Can you pay me the right amount of money? *Note down whether the pupil gave the correct amount.*	_____ 01 Yes 02 No
STMY02	Now you want to buy two exercise books. One exercise book costs 25 Naira. Can you pay me the right amount of money for the two? *Note down whether the pupil gave the correct amount.*	_____ 01 Yes 02 No SKIP to end if STMY01= NO and STMY02=NO
STMY03	*Take the picture cards. Present the flash card set MY2. Keep the bills 100 Naira and 50 Naira for you.*	

	Now we change roles. You are a vendor and sell mangos, oranges and bananas. I am a customer. This is your money, you can use it to give change.	
STMY04	One mango costs 35 Naira. I would like to buy one mango for 35 Naira. *Give the pupil a 50 Naira note.* Can you give me change? *Note down whether the pupil gave the correct amount.*	_____ 01 Yes 02 No
STMY05	One mango costs 35 Naira, one orange costs 20 Naira. I would like to buy one mango and one orange. *Give the pupil a 100 Naira note.* Can you give me change? *Note down whether the pupil gave the correct amount.*	_____ 01 Yes 02 No

II. Language		
Q_55	**Thank you. Now I have some exercises about reading letters and words.**	
1. Recognizing letters		
STRL01	Which letter is this? *Present flash card A1 with "a"* *If pupil is shy to say something, encourage.* *Do not assist.* *Note down whether pupil was able to identify the letter. (Hausa name, English name, sound are all correct)*	_____ 01 Yes 02 No
STRL02	And which letter is this? *Present flash card A2 with "e"*	_____ 01 Yes 02 No

	Note down whether pupil was able to identify the letter. (Hausa name, English name, sound are all correct)	
STRL03	And which letter is this? *Present flash card A3 with “m”* *Note down whether pupil was able to identify the letter. (Hausa name, English name, sound are all correct)*	_____ 01 Yes 02 No
STRL04	And which letter is this? *Present flash card A4 with “d”* *Note down whether pupil was able to identify the letter. (Hausa name, English name, sound are all correct)*	_____ 01 Yes 02 No SKIP to STRC01 if 3 out of 4 items from STRL01-04 are wrong
2. Reading aloud		
STRA01	Can you read this syllable for me? *Present flash card B1 with “ka”* *If pupil names the letters separately, ask him/her to combine it and pronounce the sound.* *Note down whether pupil was able to read the syllable.*	_____ 01 Yes 02 In second attempt after told to join the letters 03 No
STRA02	Can you read this syllable for me? *Present flash card B2 with “fi”* *If pupil names the letters separately, ask him/her to combine it and pronounce the sound.* *Note down whether pupil was able to read the syllable.*	_____ 01 Yes 02 In second attempt after told to join the letters 03 No
STRA03	Can you read this syllable for me?	_____ 01 Yes

	Present flash card B3 with "lo" *If pupil names the letters separately, ask him/her to combine it and pronounce the sound.* *Note down whether pupil was able to read the syllable.*	02 In second attempt after told to join the letters 03 No
STRA04	Can you read this syllable for me? *Present flash card B4 with "gwa"* *If pupil names the letters separately, ask him/her to combine it and pronounce the sound.* *Note down whether pupil was able to read the syllable.*	_____ 01 Yes 02 In second attempt after told to join the letters 03 No SKIP to STRC01 if 3 out of 4 items from STRA01-04 are wrong
STRA05	Can you read this word for me? If you want, you can use your finger. *Present flash card B5 with "keke"* *If pupil is shy to say something or stops while reading, encourage to try / continue. Do not assist in reading.* *Note down whether pupil was able to read the word.*	_____ 01 Yes 02 With mistakes 03 Only first letter / first syllable recognized 04 No
STRA06	Can you read this word for me? *Present flash card B6 with "rakumi"* *Note down whether pupil was able to read the word.*	_____ 01 Yes 02 With mistakes 03 Only first letter / first syllable recognized 04 No Skip to STRC01 if STRA05= NO (or Only first letter or With mistakes) AND STRA06 =NO (or Only first letter or With mistakes)
STRA07	Can you read this word for me? *Present flash card B7 with "ciyawa"*	_____ 01 Yes 02 With mistakes 03 Only first letter / first syllable recognized

	Note down whether pupil was able to read the word.	04 No
STRA08	Can you read this word for me? *Present flash card B8 with “lela”* *Note down whether pupil was able to read the word.*	_____ 01 Yes 02 With mistakes 03 Only first letter / first syllable recognized 04 No
STRA09	Can you read this word for me? *Present flash card B9 with “tumari”* *Note down whether pupil was able to read the word.*	_____ 01 Yes 02 With mistakes 03 Only first letter / first syllable recognized 04 No Skip to STRC01 if STRA08=NO (or Only first letter or With mistakes) AND STRA09 =NO (or Only first letter or With mistakes)
STRA10	Can you read this word for me? *Present flash card B10 with “jirgoya”* *Note down whether pupil was able to read the word.*	_____ 01 Yes 02 With mistakes 03 Only first letter / first syllable recognized 04 No
STRA11	Can you read this sentence for me? If you want, you can use your finger. *Present flash card B11 with full sentence.* *If pupil is shy to say something or stops while reading, encourage to try / continue. Do not assist in reading.* *Note down whether pupil was able to read the sentence.*	_____ 01 Yes 02 With mistakes 03 No → go to STRC01
STRA12	Can you read this paragraph for me? *Present flash card B12 with short paragraph.* *Note down whether pupil was able to read the paragraph.*	_____ 01 Yes 02 With mistakes 03 Read at least one sentence but could not finish 04 No

3. Reading comprehension		
STRC01	I will show you a card with a word written on it and four picture cards. One of the pictures shows the word that is written on the card. Please read the word and give me the picture card showing the same thing. *Present flash card set number C1.* *Note down whether pupil was able to show the correct card.*	_____ 01 Yes 02 No
STRC02	Now I have a new set of cards. Please read the word and give me the picture card showing the same thing. *Present flash card set number C2.* *Note down whether pupil was able to show the correct card.*	_____ 01 Yes 02 No Skip to STLC01 if STRC01= NO AND STRC02 =NO
STRC03	Now I have a new set of cards. Please read the word and give me the picture card showing the same thing. *Present flash card set number C3.* *Note down whether pupil was able to show the correct card.*	_____ 01 Yes 02 No
3. Listening comprehension		
STLC00	Now I am going to read out a short story. Please listen carefully as I will ask you to retell the story later. *Aisha da Amina yan uwan junane. Sukan tafi Makaranta tare, kuma ajinsu daya. Bayan an tashi daga Makaranta a hanyar su zuwa gida Sun hadu da rakumi a kwance ga hanya.* *Lokacin da suka isa gida sun samu uwayensu suna dafa abinchi. Sai a ka aikesu kasuwa su sayo tomatur, da uwayensu suka manta su sayo. Bayan an kare abinchi aka basu nasu suka ci, San nan sukaje wurin wasa.*	

	Dadare yayi sukayi kwana. *Aisha and Amina are sisters. They go to school together and they are in the same class.* *After school on their way home they passed a camel laying down by the road side. When they reached home they met their parents cooking. They were sent to the market to buy tomatoes that the parents had forgotten to buy. When the food was ready they ate their own and went out to play. Then when it was night they went to bed.*	
STLC01	Here I have some pictures that show situations from the story we just heard. Can you order the pictures in the correct order so that they retell the story? *Present flash card set number D.* *Note down whether pupil was able to order the cards correctly.*	_____ 01 Yes 02 No
5. Writing		
STW01	*Hand the pupil a piece of paper.* Can you please write down your name? *Note down whether pupil has written* - *correctly written name* - *name with spelling mistake / letter mirrored* - *fragment* - *nothing, scribbling, random letters*	_____ 01 Correct 02 Spelling mistakes 03 Fragment 04 Nothing / scribbling / random letters → Go to PUP00
STW02	*Present flash card E1 (car).* What do you see on the flash card? Can you please write down the word? *Note down whether pupil has written* - *correctly written name* - *name with spelling mistake / letter mirrored* - *fragment*	_____ 01 Correct 02 Spelling mistakes 03 Fragment 04 First letter correct 05 Nothing / scribbling / random letters

	- *nothing, scribbling, random letters*	
STW03	*Present flash card E2 (tebur).* What do you see on the flash card? Can you please write down the word? *Note down whether pupil has written* - *correctly written name* - *name with spelling mistake / letter mirrored* - *fragment* - *nothing, scribbling, random letters*	_____ 01 Correct 02 Spelling mistakes 03 Fragment 04 First letter correct 05 Nothing / scribbling / random letters → Go to PUP00
STW04	*Present flash card E3 (kujera).* What do you see on the flash card? Can you please write down the word? *Note down whether pupil has written* - *correctly written name* - *name with spelling mistake / letter mirrored* - *fragment* - *nothing, scribbling, random letters*	_____ 01 Correct 02 Spelling mistakes 03 Fragment 04 First letter correct 05 Nothing / scribbling / random letters
PUP00	We are at the end of this part. Next, I would like to ask you some questions.	
PUP02	What is your father’s name? *Write down the full name. If Don’t know or No answer, write DK and NA respectively*	______________________________ DK Don’t know NA No answer
PUP03	*Note down gender of respondent.*	_____ 01 Female 02 Male
PUP04	How old are you?	______ 01 Years, specify ____________ 02 Don’t know 03 No answer
PUP05	In which class are you?	______

PUP06	Did you have breakfast in the morning?	_____ 01 Yes 02 No 04 No answer
PUP07	Do you sometimes study at home?	_____ 01 Yes 02 No → go to PUP11 04 No answer → go to PUP11
PUP08	How often do you study at home? (Everyday, several days a week, once a week, less than once a week.)	_____ 01 Everyday 02 Several days a week 03 Once a week 04 Less than once a week 05 Don’t know 06 No answer
PUP09	Do you usually study alone or is somebody helping you or sitting with you?	_____ 01 Alone → go to PUP11 02 Sometimes alone, sometimes with help 03 Somebody is helping me 04 Don’t know → go to PUP11 05 No answer → go to PUP11
PUP10	Who is helping or sitting with you? Anybody else? *Multiple answers possible. Tick all that apply and write down any additional answer.*	[ ] Mother [ ] Father [ ] Grandmother [ ] Grandfather [ ] Older siblings [ ] Other relatives (older generation, e.g. uncle) [ ] Other relatives (same generation, e.g. cousin) [ ] Friends [ ] Other, specify _______________ [ ] No answer
PUP11	If you don’t want to go to school, can you stay at home or will anybody make you go?	_____ 01 I can stay at home 02 Sometimes I can stay, sometimes not 03 My parents make me go 04 My siblings / my friends make me go 05 Don’t know 06 No answer

PUP12	Do your teachers often have to use the cane to punish other pupils?	_____ 01 Yes 02 Sometimes 03 No 04 Don't know 05 No answer
PUP13	Is going to school more important for girls or for boys?	_____ 01 Girls 02 Boys 03 Both 04 Don't know 05 No answer
PUP14	Who is the best reader in your class? *Mark whether the pupil named by the respondent is a girl, a boy or the respondent himself.*	_____ 01 Girl 02 Boy 03 Respondent / self 04 Don't know 05 No answer
PUP18	How do you feel when playing with your friends? *Show and explain scale*	01 Strongly dislike 02 Somewhat dislike 03 Average 04 Somewhat like 05 Strongly like 06 No answer
PUP15	How do you feel about going to school? *Refer to scale*	01 Strongly dislike 02 Somewhat dislike 03 Average 04 Somewhat like 05 Strongly like 06 No answer
PUP16	And how about the school compound? *Refer to scale*	01 Strongly dislike 02 Somewhat dislike 03 Average 04 Somewhat like 05 Strongly like 06 No answer
PUP17	And how about your classroom? *Refer to scale*	01 Strongly dislike 02 Somewhat dislike 03 Average 04 Somewhat like 05 Strongly like 06 No answer

Thanks	**We are at the end of the questionnaire now. Thank you for your help.**	
X. Interview information II		
Endtime	End time	__________
Int_Name	Interviewer’s Name	__________
Int_Code	Interviewer’s Code	__________
Sup_Name	Supervisor’s Name	__________
Sup_Code	Supervisor’s Code	__________

References

Abay, K. A., Berhane, G., Hoddinott, J., & Tafere, K. (2021). *Assessing Response Fatigue in Phone Surveys: Experimental Evidence on Dietary Diversity in Ethiopia* [Working Paper]. World Bank. https://doi.org/10.1596/1813-9450-9636

Adeniran, A., Ishaku, J., & Akanni, L. O. (2020). Is Nigeria experiencing a learning crisis: Evidence from curriculum-matched learning assessment. *International Journal of Educational Development*, *77*, 102199. https://doi.org/10.1016/j.ijedudev.2020.102199

Agüero, J., & Frisancho, V. (2017). *Misreporting in Sensitive Health Behaviors and its Impact on Treatment Effects: An Application to Intimate Partner Violence* (SSRN Scholarly Paper ID 3103802). Social Science Research Network. https://doi.org/10.2139/ssrn.3103802

Ahlquist, J. S. (2018). List Experiment Design, Non-Strategic Respondent Error, and Item Count Technique Estimators. *Political Analysis*, *26*(1), 34–53. https://doi.org/10.1017/pan.2017.31

Airhihenbuwa, C. O., Ford, C. L., & Iwelunmor, J. I. (2014). Why Culture Matters in Health Interventions: Lessons From HIV/AIDS Stigma and NCDs. *Health Education & Behavior*, *41*(1), 78–84. https://doi.org/10.1177/1090198113487199

Alvarez, R. M., Atkeson, L. R., Levin, I., & Li, Y. (2019). Paying Attention to Inattentive Survey Respondents. *Political Analysis*, *27*(2), 145–162. https://doi.org/10.1017/pan.2018.57

Aminu Musa Yabo. (2017). *Historial foundations of education in Nigeria*. Life-Line Educational Consultants.

Aronow, P. m., Coppock, A., Crawford, F. W., & Green, D. P. (2015). Combining List Experiment and Direct Question Estimates of Sensitive Behavior Prevalence. *Journal of Survey Statistics and Methodology*, *3*(1), 43–66. https://doi.org/10.1093/jssam/smu023

ASER Centre. (2020). *ASER 2019: Annual Status of Education Report (Rural). “Early Years.”*

Asha, K. P. (2014). Efficiency of Anganwadi Centres—A Study in Thiruvananthapuram District, Kerala. *Journal of Academia and Industrial Research*, *3*(3), 132–136.

Banerjee, A., Barnhardt, S., & Duflo, E. (2018). Can iron-fortified salt control anemia? Evidence from two experiments in rural Bihar. *Journal of Development Economics*, *133*, 127–146. https://doi.org/10.1016/j.jdeveco.2017.12.004

Banerjee, A. V., Bhattacharjee, S., Chattopadhyay, R., & Ganimian, A. J. (2017). The Untapped math skills of working children in India: Evidence, possible explanations, and implications. *Unpublished Manuscript*.

Beegle, K., De Weerdt, J., Friedman, J., & Gibson, J. (2012). Methods of household consumption measurement through surveys: Experimental results from Tanzania. *Journal of Development Economics*, *98*(1), 3–18. https://doi.org/10.1016/j.jdeveco.2011.11.001

Blair, G., Coppock, A., & Moor, M. (2020). When to Worry about Sensitivity Bias: A Social Reference Theory and Evidence from 30 Years of List Experiments. *American Political Science Review*, *114*(4), 1297–1315. https://doi.org/10.1017/S0003055420000374

Blair, G., & Imai, K. (2010). *list: Statistical Methods for the Item Count Technique and List Experiment* (9.2.2). https://CRAN.R-project.org/package=list

Blair, G., & Imai, K. (2012). Statistical Analysis of List Experiments. *Political Analysis*, *20*, 47–77. https://doi.org/10.1093/pan/mpr048

Blair, G., Imai, K., & Zhou, Y.-Y. (2015). Design and Analysis of the Randomized Response Technique. *Journal of the American Statistical Association*, *110*(511), 1304–1319. https://doi.org/10.1080/01621459.2015.1050028

Blattman, C., Jamison, J., Koroknay-Palicz, T., Rodrigues, K., & Sheridan, M. (2016). Measuring the measurement error: A method to qualitatively validate survey data. *Journal of Development Economics*, *120*, 99–112. https://doi.org/10.1016/j.jdeveco.2016.01.005

Bobonis, G. J., Miguel, E., & Puri-Sharma, C. (2006). Anemia and School Participation. *Journal of Human Resources*, *XLI*(4), 692–721. https://doi.org/10.3368/jhr.XLI.4.692

Bogler, L., Bommer, C., Ebert, C., Kumar, A., Subramanian, S. V., Subramanyam, M., & Vollmer, S. (2021). *Effects of a large-scale Participatory Learning and Action Programme in Women's Groups on Health, Nutrition, Water, Sanitation, and Hygiene: A Cluster-Randomized Controlled Trial in Bihar, India*.

Brewer, M., Etheridge, B., & O'Dea, C. (2017). Why are Households that Report the Lowest Incomes So Well-off? *The Economic Journal*, *127*(605), F24–F49. https://doi.org/10.1111/ecoj.12334

Bruckmeier, K., Riphahn, R. T., & Wiemers, J. (2021). Misreporting of program take-up in survey data and its consequences for measuring non-take-up: New evidence from linked administrative and survey data. *Empirical Economics*, *61*(3), 1567–1616. https://doi.org/10.1007/s00181-020-01921-4

Brusco, N. K., & Watts, J. J. (2015). Empirical evidence of recall bias for primary health care visits. *BMC Health Services Research*, *15*(1), 381. https://doi.org/10.1186/s12913-015-1039-1

Bullock, W., Imai, K., & Shapiro, J. N. (2017). Statistical Analysis of Endorsement Experiments: Measuring Support for Militant Groups in Pakistan. *Political Analysis*, *19*(4), 363–384. https://doi.org/10.1093/pan/mpr031

Campbell, H., Arifeen, S. el, Hazir, T., O'Kelly, J., Bryce, J., Rudan, I., & Qazi, S. A. (2013). Measuring Coverage in MNCH: Challenges in Monitoring the Proportion of Young Children with Pneumonia Who Receive Antibiotic Treatment. *PLOS Medicine*, *10*(5), e1001421. https://doi.org/10.1371/journal.pmed.1001421

Carletto, G., Tiberti, M., & Zezza, A. (2022). Measure for Measure: Comparing Survey Based Estimates of Income and Consumption for Rural Households. *The World Bank Research Observer*, *37*(1), 1–38. https://doi.org/10.1093/wbro/lkab009

Carraher, T. N., Carraher, D. W., & Schliemann, A. D. (1985). Mathematics in the streets and in schools. *British Journal of Developmental Psychology*, *3*(1), 21–29.

Carraher, T. N., Carraher, D. W., & Schliemann, A. D. (1987). Written and oral mathematics. *Journal for Research in Mathematics Education*, *18*(2), 83–97.

Castilla, C., & Murphy, D. M. A. (2021). *Bidirectional Intimate Partner Violence: Evidence from a List Experiment in Kenya* (SSRN Scholarly Paper ID 3771604). Social Science Research Network. https://doi.org/10.2139/ssrn.3771604

Charles, C. V., Dewey, C. E., Daniell, W. E., & Summerlee, A. J. S. (2011). Iron-deficiency anaemia in rural Cambodia: Community trial of a novel iron supplementation technique. *European Journal of Public Health*, *21*(1), 43–48.

Charles, C. V., Dewey, C. E., Hall, A., Hak, C., Channary, S., & Summerlee, A. J. S. (2015). A randomized control trial using a fish-shaped iron ingot for the amelioration of iron deficiency anemia in rural Cambodian women. *Tropical Medicine & Surgery*, *3*(3), 195.

Chuang, E., Dupas, P., Huillery, E., & Seban, J. (2021). Sex, lies, and measurement: Consistency tests for indirect response survey methods. *Journal of Development Economics*, *148*, 102582. https://doi.org/10.1016/j.jdeveco.2020.102582

Corstange, D. (2009). Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT. *Political Analysis*, *17*(1), 45–63. https://doi.org/10.1093/pan/mpn013

Cullen, C. (2020). *Method Matters: Underreporting of Intimate Partner Violence in Nigeria and Rwanda* (SSRN Scholarly Paper ID 3624515). Social Science Research Network. https://papers.ssrn.com/abstract=3624515

Czaja, C., Crossette, L., & Metlay, J. P. (2005). Accuracy of Adult Reported Pneumococcal Vaccination Status of Children. *Annals of Epidemiology*, *15*(4), 253–256. https://doi.org/10.1016/j.annepidem.2004.07.091

Dasgupta, R., Arora, N. K., Ramji, S., Chaturvedi, S., Rewal, S., Suresh, K., Deshmukh, V., & Thakur, N. (2012). Managing Childhood Under-Nutrition: Role and Scope of Health Services. *Economic and Political Weekly*, *47*(16), 15–19.

de Nicola, F., & Giné, X. (2014). How accurate are recall data? Evidence from coastal India. *Journal of Development Economics*, *106*, 52–65. https://doi.org/10.1016/j.jdeveco.2013.08.008

Deshpande, S. (2019). *Governing Nutrition, Performing State: Workers of the Integrated Child Development Services (ICDS) Programme, India*. University of Sussex.

Desiere, S., & Costa, V. (2019). *Employment Data in Household Surveys: Taking Stock, Looking Ahead* (SSRN Scholarly Paper No. 3430490). https://papers.ssrn.com/abstract=3430490

Dillon, M. R., Kannan, H., Dean, J. T., Spelke, E. S., & Duflo, E. (2017). Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics. *Science*, *357*, 47–55.

Druckman, J. N., Gilli, M., Klar, S., & Robison, J. (2015). Measuring Drug and Alcohol Use Among College Student-Athletes*. *Social Science Quarterly*, *96*(2), 369–380. https://doi.org/10.1111/ssqu.12135

Dupas, P. (2011). Health behavior in developing countries. *Annual Review of Economics*, *3*(1), 425–449.

Ebert, C., Heesemann, E., & Vollmer, S. (2020). *Encouraging parents to invest: A randomized trial with two simple interventions in early childhood* (Working Paper No. 856). Ruhr Economic Papers. https://doi.org/10.4419/86788992

Fraker, A., Shah, N. B., & Abraham, R. (2013). *Quantitative assessment: Beneficiary nutritional status and performance of ICDS Supplementary Nutrition Programme in Bihar*.

Ganimian, A. J., Muralidharan, K., & Walters, C. R. (2021). *Augmenting State Capacity for Child Development: Experimental Evidence from India* (Working Paper No. 28780; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w28780

Gardner, W., & Kassebaum, N. (2020). Global, Regional, and National Prevalence of Anemia and Its Causes in 204 Countries and Territories, 1990–2019. *Current Developments in Nutrition*, *4*(Suppl 2), 830. https://doi.org/10.1093/cdn/nzaa053_035

Gilens, M., Sniderman, P. M., & Kuklinski, J. H. (1998). Affirmative Action and the Politics of Realignment. *British Journal of Political Science*, *28*(1), 159–183. https://doi.org/10.1017/S0007123498000143

Gilligan, D. O., Melissa, H., Jessica, L., & Heleene, T. (2021). *Using a list experiment to measure intimate partner violence: Cautionary evidence from Ethiopia*. Intl Food Policy Res Inst.

Gingerich, D. W. (2010). Understanding Off-the-Books Politics: Conducting Inference on the Determinants of Sensitive Behavior with Randomized Response Surveys. *Political Analysis*, *18*(3), 349–380. https://doi.org/10.1093/pan/mpq010

Glynn, A. N. (2013). What Can We Learn with Statistical Truth Serum? *Public Opinion Quarterly*, *77*(S1), 159–172. https://doi.org/10.1093/poq/nfs070

Gonzalez-Ocantos, E., de Jonge, C. K., Meléndez, C., Osorio, J., & Nickerson, D. W. (2012). Vote Buying and Social Desirability Bias: Experimental Evidence from Nicaragua. *American Journal of Political Science*, *56*(1), 202–217. https://doi.org/10.1111/j.1540-5907.2011.00540.x

Haber, N., Harling, G., Cohen, J., Mutevedzi, T., Tanser, F., Gareta, D., Herbst, K., Pillay, D., Bärnighausen, T., & Fink, G. (2018). List randomization for eliciting HIV status and sexual behaviors in rural KwaZulu-Natal, South Africa: A randomized experiment using known true values for validation. *BMC Medical Research Methodology*, *18*(1), 46. https://doi.org/10.1186/s12874-018-0507-9

Halterman, J. S., Kaczorowski, J. M., Aligne, C. A., Auinger, P., & Szilagyi, P. G. (2001). Iron Deficiency and Cognitive Achievement Among School-Aged Children and Adolescents in the United States. *Pediatrics*, *107*(6), 1381–1386. https://doi.org/10.1542/peds.107.6.1381

Harrison, L. H., Moursi, S., Guinena, A. H., Gadomski, A. M., El-Ansary, K. S., Khallaf, N., & Black, R. E. (1995). Maternal Reporting of Acute Respiratory Infection in Egypt. *International Journal of Epidemiology*, *24*(5), 1058–1063. https://doi.org/10.1093/ije/24.5.1058

Hazir, T., Begum, K., Arifeen, S. el, Khan, A. M., Huque, M. H., Kazmi, N., Roy, S., Abbasi, S., Rahman, Q. S., Theodoratou, E., Khorshed, M. S., Rahman, K. M., Bari, S., Kaiser, M. M. I., Saha, S. K., Ahmed, A. S. M. N. U., Rudan, I., Bryce, J., Qazi, S. A., & Campbell,

H. (2013). Measuring Coverage in MNCH: A Prospective Validation Study in Pakistan and Bangladesh on Measuring Correct Treatment of Childhood Pneumonia. *PLOS Medicine*, *10*(5), e1001422. https://doi.org/10.1371/journal.pmed.1001422

Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly*, *74*(1), 37–67. https://doi.org/10.1093/poq/nfp065

Hurst, E., Li, G., & Pugsley, B. (2014). Are Household Surveys Like Tax Forms? Evidence from Income Underreporting of the Self-Employed. *The Review of Economics and Statistics*, *96*(1), 19–33.

Imai, K. (2011). Multivariate Regression Analysis for the Item Count Technique. *Journal of the American Statistical Association*, *106*(494), 407–416. https://doi.org/10.1198/jasa.2011.ap10415

International Institute for Population Sciences (IIPS). (2021). *National Family Health Survey (NFHS-5), 2019-20: District Fact Sheet Madhepura, Bihar*. IIPS.

International Institute for Population Sciences (IIPS) and ICF. (2021a). *National Family Health Survey (NFHS-5), 2019-21: India: Volume I*. IIPS.

International Institute for Population Sciences (IIPS) and ICF. (2021b). *National Family Health Survey (NFHS-5), 2019-21: State Fact Sheet Bihar*.

Jamison, J. C., Karlan, D., & Raffler, P. (2013). *Mixed Method Evaluation of a Passive mHealth Sexual Information Texting Service in Uganda* (Working Paper No. 19107; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w19107

Jerke, J., Johann, D., Rauhut, H., & Thomas, K. (2019). Too sophisticated even for highly educated survey respondents? A qualitative assessment of indirect question formats for sensitive questions. *Survey Research Methods*, *13*(3), Article 3. https://doi.org/10.18148/srm/2019.v13i3.7453

John, A., Nisbett, N., Barnett, I., Avula, R., & Menon, P. (2020). Factors influencing the performance of community health workers: A qualitative study of Anganwadi Workers from Bihar, India. *PLOS ONE*, *15*(11), e0242460. https://doi.org/10.1371/journal.pone.0242460

Kapil, U., Kapil, R., & Gupta, A. (2019). National Iron Plus Initiative: Current status & future strategy. *The Indian Journal of Medical Research*, *150*(3), 239–247. https://doi.org/10.4103/ijmr.IJMR_1782_18

Kelly, C. A., Soler-Hampejsek, E., Mensch, B. S., & Hewett, P. C. (2013). Social Desirability Bias in Sexual Behavior Reporting: Evidence from an Interview Mode Experiment in Rural Malawi. *International Perspectives on Sexual and Reproductive Health*, *39*(1), 14. https://doi.org/10.1363/3901413

Khan, F. A. (2004). Living, learning and doing mathematics: A study of working-class children in Delhi. *Contemporary Education Dialogue*, *1*(2), 199–227.

Kilic, T., & Sohnesen, T. P. (2019). Same Question But Different Answer: Experimental Evidence on Questionnaire Design's Impact on Poverty Measured by Proxies. *Review of Income and Wealth*, *65*(1), 144–165. https://doi.org/10.1111/roiw.12343

Kovaleva, A., Beierlein, C., Kemper, C. J., & Rammstedt, B. (2012). *Eine Kurzskala zur Messung von Kontrollüberzeugung: Die Skala Internale-Externale-Kontrollüberzeugung-4 (IE-4)*. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-312096

Krämer, M., Kumar, S., & Vollmer, S. (2021). Improving Child Health and Cognition: Evidence from a School-Based Nutrition Intervention in India. *The Review of Economics and Statistics*, *103*(5), 818–834. https://doi.org/10.1162/rest_a_00950

Kramon, E., & Weghorst, K. (2019). (Mis)Measuring Sensitive Attitudes with the List Experiment. *Public Opinion Quarterly*, *83*(S1), 236–263. https://doi.org/10.1093/poq/nfz009

Kuklinski, J. H., Cobb, M. D., & Gilens, M. (1997). Racial Attitudes and the "New South." *The Journal of Politics*, *59*(2), 323–349. https://doi.org/10.1017/S0022381600053470

LaBrie, J. W., & Earleywine, M. (2000). Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique. *The Journal of Sex Research*, *37*(4), 321–326. https://doi.org/10.1080/00224490009552054

Lax, J. R., Phillips, J. H., & Stollwerk, A. F. (2016). Are Survey Respondents Lying about Their Support for Same-Sex Marriage? Lessons from a List Experiment. *Public Opinion Quarterly*, *80*(2), 510–533. https://doi.org/10.1093/poq/nfv056

Lépine, A., Treibich, C., & D'Exelle, B. (2020). Nothing but the truth: Consistency and efficiency of the list experiment method for the measurement of sensitive health behaviours. *Social Science & Medicine*, *266*, 113326. https://doi.org/10.1016/j.socscimed.2020.113326

Lépine, A., Treibich, C., Ndour, C. T., Gueye, K., & Vickerman, P. (2020). HIV infection risk and condom use among sex workers in Senegal: Evidence from the list experiment method. *Health Policy and Planning*, *35*(4), 408–415. https://doi.org/10.1093/heapol/czz155

Leyna, G. H., Berkman, L. F., Njelekela, M. A., Kazonda, P., Irema, K., & Killewo, J. (2017). Profile: The Dar Es Salaam Health and Demographic Surveillance System (Dar es Salaam HDSS). *International Journal of Epidemiology*, *46*(3), 801–808.

Liu, G., Liao, Z., Xu, X., Liang, Y., Xiong, Y., & Ni, J. (2017). Accuracy of parent-reported measles-containing vaccination status of children with measles. *Public Health*, *144*, 92–95. https://doi.org/10.1016/j.puhe.2016.12.013

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

Lyall, J., Blair, G., & Imai, K. (2013). Explaining Support for Combatants during Wartime: A Survey Experiment in Afghanistan. *American Political Science Review*, *107*(4), 679–705. https://doi.org/10.1017/S0003055413000403

Maity, B. (2016). Interstate Differences in the Performance of "Anganwadi" Centres under ICDS Scheme. *Economic and Political Weekly*, *51*(51), 59–66.

Martinelli, C., & Parker, S. W. (2009). Deception and Misreporting in a Social Program. *Journal of the European Economic Association*, *7*(4), 886–908. https://doi.org/10.1162/JEEA.2009.7.4.886

McKenzie, D., & Siegel, M. (2013). Eliciting illegal migration rates through list randomization. *Migration Studies*, *1*(3), 276–291. https://doi.org/10.1093/migration/mnt018

Meyer, B. D., & Mittag, N. (2019). Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness, and Holes in the Safety Net. *American Economic Journal: Applied Economics*, *11*(2), 176–204. https://doi.org/10.1257/app.20170478

Mikkelsen, L., Phillips, D. E., AbouZahr, C., Setel, P. W., de Savigny, D., Lozano, R., & Lopez, A. D. (2015). A global assessment of civil registration and vital statistics systems: Monitoring data quality and progress. *The Lancet*, *386*(10001), 1395–1406. https://doi.org/10.1016/S0140-6736(15)60171-4

Miles, M., Ryman, T. K., Dietz, V., Zell, E., & Luman, E. T. (2013). Validity of vaccination cards and parental recall to estimate vaccination coverage: A systematic review of the literature. *Vaccine*, *31*(12), 1560–1568. https://doi.org/10.1016/j.vaccine.2012.10.089

Ministry of Women and Child Development, India. (2020, September 17). Major Welfare Schemes of the Ministry [Press Release]. *Press Release*. https://pib.gov.in/PressReleasePage.aspx?PRID=1655615

Mittal, N., & Meenakshi, J. V. (2019). Does the ICDS Improve Children's Diets? Some Evidence from Rural Bihar. *The Journal of Development Studies*, *55*(11), 2424–2439. https://doi.org/10.1080/00220388.2018.1487054

Moseson, H., Treleaven, E., Gerdts, C., & Diamond-Smith, N. (2017). The List Experiment for Measuring Abortion: What We Know and What We Need. *Studies in Family Planning*, *48*(4), 397–405. https://doi.org/10.1111/sifp.12042

National Population Commission (NPC) [Nigeria] and ICF. (2019). *Nigeria. Demographic and Health Survey 2018*. NPC and ICF.

Nigeria Northern Education Initiative (NEI). (2013). *Results of the 2013 Early Grade Reading and Early Grade Mathematics Assessments (EGRA & EGMA) in Sokoto State*.

Nunes, T., Schliemann, A. D., & Carraher, D. W. (1993). *Street mathematics and school mathematics*. Cambridge University Press.

Ochmann, S., Owolabi, K. E., Olatunji-David, F., Okunlola, N., & Vollmer, S. (2021). The impact of grants in combination with school-based management trainings on primary education: A cluster-randomized trial in Northern Nigeria. *Journal of Development Effectiveness*, *0*(0), 1–20. https://doi.org/10.1080/19439342.2021.2007980

PAL Network. (2020). *ICAN: International Common Assessment of Numeracy. Background, Features and Large-scale Implementation*. People's Action for Learning Network.

Parmar, M., Patel, S., Rathod, S., Patel, N., & Ninama, K. (2015). Knowledge or Anganwadi Worker about Integrated Child Development Services (ICDS): A Study of Urban Blocks in Ahmedabad District of Gujarat. *International Journal of Multidisciplinary Research and Development*, *2*(8), 170–174.

Peterman, A., Palermo, T. M., Handa, S., Seidenfeld, D., & Team, on behalf of the Z. C. G. P. E. (2018). List randomization for soliciting experience of intimate partner violence: Application to the evaluation of Zambia's unconditional child grant program. *Health Economics*, *27*(3), 622–628. https://doi.org/10.1002/hec.3588

Phillips, A. E., Gomez, G. B., Boily, M.-C., & Garnett, G. P. (2010). A systematic review and meta-analysis of quantitative interviewing tools to investigate self-reported HIV and STI associated behaviours in low- and middle-income countries. *International Journal of Epidemiology*, *39*(6), 1541–1555. https://doi.org/10.1093/ije/dyq114

Porter, C., Favara, M., Sánchez, A., & Scott, D. (2021). The impact of COVID-19 lockdowns on physical domestic violence: Evidence from a list randomization experiment. *SSM - Population Health*, *14*, 100792. https://doi.org/10.1016/j.ssmph.2021.100792

Ramakrishnan, R., Venkata Rao, T., Sundaramoorthy, L., & Joshua, V. (1999). Magnitude of recall bias in the estimation of immunization coverage and its determinants. *Indian Pediatrics*, *36*, 881–885.

Roe-Sepowitz, D., Bontrager, S., Pickett, J. T., & Kosloski, A. E. (2019). Estimating the sex buying behavior of adult males in the United States: List experiment and direct question estimates. *Journal of Criminal Justice*, *63*, 41–48. https://doi.org/10.1016/j.jcrimjus.2019.04.005

Rosenfeld, B., Imai, K., & Shapiro, J. N. (2016). An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions. *American Journal of Political Science*, *60*(3), 783–802. https://doi.org/10.1111/ajps.12205

Sawadogo-Lewis, T., Keita, Y., Wilson, E., Sawadogo, S., Téréra, I., Sangho, H., & Munos, M. (2021). Can We Use Routine Data for Strategic Decision Making? A Time Trend Comparison Between Survey and Routine Data in Mali. *Global Health: Science and Practice*, *9*(4), 869–880. https://doi.org/10.9745/GHSP-D-21-00281

Sitabkhan, Y. (2011). The Use of Convenient Value Strategies among Young Train Vendors in Mumbai, India. *Proceedings of EpiSTEME 3*.

Sitabkhan, Y., Platas, L. M., & Geller, L. R. K. (2018). Capturing Children's Mathematical Knowledge: An Assessment Framework. *Global Education Review*, *5*(3), 106–124.

Spinillo, A. G. (2018). Number sense in elementary school children: The uses and meanings given to numbers in different investigative situations. In *Invited Lectures from the 13th international Congress on Mathematical Education* (pp. 639–650). Springer.

Stoebenau, K., Heise, L., Wamoyi, J., & Bobrova, N. (2016). Revisiting the understanding of "transactional sex" in sub-Saharan Africa: A review and synthesis of the literature. *Social Science & Medicine*, *168*, 186–197. https://doi.org/10.1016/j.socscimed.2016.09.023

USAID, Federal Ministry of Education, Nigeria, National Population Commission, & Universal Basic Education Commission. (2015). *Nigeria: 2015 Nigeria Education Data Survey (NEDS)*. National Population Commission. https://ierc-publicfiles.s3.amazonaws.com/public/resources/2015-NEDS-National-011716.pdf

Uwezo. (2019a). *Are our children learning? Uwezo Tanzania Learning Assessment Report 2019*. Twaweza East Africa.

Uwezo. (2019b). *Are our children learning? Uwezo Uganda Eighth Learning Assessment Report 2019*. Twaweza East Africa.

Verma, R., Gupta, S., & Birner, R. (2018). Can vigilance-focused governance reforms improve service delivery? The case of Integrated Child Development Services (ICDS) in Bihar, India. *Development Policy Review*, *36*, O786–O802. https://doi.org/10.1111/dpr.12344

von Grafenstein, L., Kumar, A., Kumar, S., & Vollmer, S. (2021). *Impacts of Double-Fortified Salt on Anemia and Cognition: Four-Year Follow-Up Evidence from a School-Based Nutrition Intervention in India*. http://dx.doi.org/10.2139/ssrn.3905062

Wamoyi, J., Stobeanau, K., Bobrova, N., Abramsky, T., & Watts, C. (2016). Transactional sex and risk for HIV infection in sub-Saharan Africa: A systematic review and meta-analysis. *Journal of the International AIDS Society*, *19*(1), 20992. https://doi.org/10.7448/IAS.19.1.20992

Weisberg, H. F. (2009). *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. University of Chicago Press.

Wollburg, P., Tiberti, M., & Zezza, A. (2021). Recall length and measurement error in agricultural surveys. *Food Policy*, *100*, 102003. https://doi.org/10.1016/j.foodpol.2020.102003

World Health Organization. (2016). *Guideline: Daily iron supplementation in infants and children*. World Health Organization.

Zezza, A., Carletto, C., Fiedler, J. L., Gennari, P., & Jolliffe, D. (2017). Food counts. Measuring food consumption and expenditures in household consumption and expenditure surveys (HCES). Introduction to the special issue. *Food Policy*, *72*, 1–6. https://doi.org/10.1016/j.foodpol.2017.08.007

Declarations

Declaration of author contributions

All research papers that are included in this dissertation are co-authored as indicated on the respective chapter title pages and the contributions are as follows.

Essay 1: Joint work with Ann-Charline Weber (AW), Abhijeet Kumar (AK), and Sebastian Vollmer (SV)

AW, SV, and LB jointly developed the idea. AW and LB conceptualized the study. SV and AK provided input to the conceptualization of the study. AW, AK, and LB prepared and monitored data collection in the field. AW and LB conducted the analysis and wrote the draft. SV provided input to analysis and interpretation of results.

Essay 2: Joint work with Ann-Charline Weber (AW), Abhijeet Kumar (AK), and Sebastian Vollmer (SV)

AW, SV, and LB jointly developed the idea. AW and LB conceptualized the study. SV and AK provided input to the conceptualization of the study. AW, AK, and LB prepared and monitored data collection and conducted the implementation of the intervention in the field. AW and LB conducted the analysis and wrote the draft. SV provided input to changes during implementation of the intervention, analysis, and interpretation of results.

Essay 3: Joint work with Sebastian Vollmer (SV) and Till Bärnighausen (TB)

TB and SV conceptualized the study. LB conducted the formal analysis and wrote the draft. TB and SV commented on the analysis. SV provided input, comments, and edits on the manuscript.

Essay 4: Joint work with Ann-Charline Weber and Sebastian Vollmer

AW and LB jointly developed the idea and conceptualized the study. AW and LB developed the learning assessment, prepared and monitored data collection and implementation of the learning assessment in the field. AW and LB conducted the analysis and wrote the draft. SV provided input to the analysis, interpretation of results, and writing.