Georg-August Universität Göttingen

Doktorarbeit

Essays in Empirical Development and Education Economics

Dissertation zur Erlangung des Doktorgrades
der Wirtschaftswissenschaftlichen Fakultät
an der Universität Göttingen

vorgelegt von
Simon Lange
aus Gehrden
September 16, 2015

Erstgutachter: Professor Stephan Klasen, Ph.D.

Zweitgutachter: Professor Dr. Thomas Kneib

Acknowledgements

The present thesis was written under the supervision of Stephan Klasen while I was working as a research associate at his chair. His advice and guidance have proven indispensable over time and he has supported me in my various endeavors for which I am very grateful. Having first worked at the chair as a student assistant in 2006, it has always been impressive to see what kind of research environment he manages to create and entertain for researchers in our field in Göttingen. I would also like to thank my co-supervisor Thomas Kneib as well as Inmaculada Martínez-Zarzoso for agreeing to be on my committee. Both have been very supportive.

Previous work on the topic of the first essay was supported financially by the Kreditanstalt für Wiederaufbau (KfW) in cooperation with the Federal Ministry for Economic Cooperation and Development (BMZ), for which I am grateful. The second essay results from a cooperation with staff at Unidad de Análisis de Políticas Económicas y Sociales (UDAPE) and the Inter-American Development Bank (IADB) in La Paz, Bolivia. In particular, I would like to thank Julia Johannsen of the IADB and Fernando Landa, Roland Pardo, and Cecilia Vidal of UDAPE. Single chapters have been presented in seminars in Göttingen and Berlin and at conferences in Passau, Lusaka, Boston, Oxford, and Ghent. I would like to thank participants at these events for many valuable comments we received.

My colleagues in Göttingen have been exceptional; they have made my life as a doctoral student very enjoyable. Some of these people are (in no particular order) Sebastian Vollmer, Aris Boukouras, Emre Aytimur, Sophia Kan, Riva Darmawan, Caro Dotter, Marion Krämer, Jana Kuhnt, Merle Kreibaum, Steffen Lohmann, Björn Jahnke, Bumi Camara, Atika Pasha, Nathalie Scholl, Juliane Zenker, Anna Minasyan, Kai Gehring, Jisu Yoon, Rahul Lahoti, Nicole Rippin, Konstantin Wacker, Jan Priebe, Isis Gaddis, Alex Rudolph, Andreas Fuchs, Ken Harttgen, Feli Nowak, Robert Rudolf, Tobias Lechtenfeld, Felix Povel, Dimitrios Minos, Elena Groß, Stephan Dietrich, and Friederike Greb.

Special thanks go to Malte Reimers and Marten von Werder, co-authors of the fourth and fifth chapters of this thesis, respectively. I have benefited in many ways from working with them and have very much enjoyed our discussions. I was particularly lucky to have started work on my thesis at the same time as Marica Lo Bue. This way, I have had the great pleasure to work with her in a joint project and to share an office with her ever since. I would like to thank her for making my workdays so enjoyable.

Ramona Rischke, my long-time fellow student and, lately, quasi-room mate, for her friendship in general and her encouragement and numerous thoughtful comments on my work in particular. She is an exceptional person and a dear friend. I would also like to thank Rafael Aigner and Patrick Hürttgen for their friendship, their enthusiasm for the dismal science, their many insights that they generously shared (and their rigor in doing so). I have learnt a lot from them.

Finally, I am grateful to my family, my parents Ortrud and Ulrich and my brother Birger, for their immutable confidence in me and their constant support. Most of all I would like to thank my wife Julia for her love and emotional support and for enduring the ups and downs that at times accompanied the writing of this thesis. I am looking forward to many more wonderful and exciting years with you.

Contents

Acknowledgements **i**

Contents **ii**

List of Figures **iv**

List of Tables **v**

1 Introduction **1**

2 How the New International Goal for Child Mortality is Unfair to Africa (Again) **7**

2.1 Introduction 8
2.2 Recent performance and the trouble with the MDGs 10
2.3 Mean reversion models 13
2.4 Projections and development targets 16
2.5 Future challenges 22
2.6 Conclusion and interpretations 25

Appendices

2.A List of countries in the analysis 26
2.B Results from dynamic panel data (DPD) estimators 26
2.C Alternative projections 28

3 Targeting Performance and Poverty Effects of Proxy Means-Tested Transfers: Trade-offs and Challenges **30**

3.1 Introduction 30
3.2 Targeting in developing countries 33
3.3 Datasets and welfare indicator 35
3.4 Proxy Means Tests 36
3.5 Assessing accuracy 39
3.6 Poverty simulations 46
3.7 Conclusion 50

Appendices

3.A Calculation of the welfare indicator 52

4 Livestock as an Imperfect Buffer Stock in Poorly Integrated Markets **56**

4.1 Introduction 56
4.2 Agricultural production in Burkina Faso 58
4.3 Datasets and descriptives 60
4.4 Impact of the 2004 drought 65

4.5 Rainfall and livestock trading 69
4.6 Saving out of transitory income 73
4.7 Rainfall, prices, and quantities 81
4.8 Conclusion 83

Appendices

4.A Imputation of prices 84
4.B Levels of rainfall across Burkina Faso, 1970–2009 86
4.C Testing exclusion restrictions 87

5 Tracking and the Intergenerational Transmission of Education: Evidence from a Natural Experiment 91
5.1 Introduction 91
5.2 Background: Germany's education system 94
5.3 Conceptual framework 97
5.4 Estimating the impact of the reform on educational outcomes 98
5.5 Results 103
5.6 Robustness 107
5.7 Effect of the reform on parental involvement 111
5.8 Discussion 112
5.9 Conclusion 113

Appendices

5.A Inferring the state of schooling at age ten 114

Bibliography 115

List of Figures

1.1 Growth of real GDP per capita and annual rainfall in Burkina Faso, 1970–2002. . 4

2.1 Ratio of under-five mortality rate in 2013 to 1990 against rate in 1990. 11
2.2 Estimated constant and slope coefficient from a regression of ARRs over ten years against initial levels, 1960–2003. 12
2.3 Projections of under-five mortality based on persistence model against projections based on mean reversion-model. 18
2.4 Under-five mortality rates as required by a 're-newed' MDG4 (two-thirds reduction relative to levels in 2005) against projections based on mean reversion-model. . . 21

3.1 ROC-curves for PMTs M1–M4 based on 2008 and 2011 calibration sample. . . . 41
3.2 Local polynomial regression of beneficiary status against percentile of per-adult real expenditure. 45
3.3 Simulated poverty effects against inclusion rate. 48

4.1 Average annual precipitation in millimeters, 2001–2012. 59
4.2 Timing of surveys relative to average levels of rainfall and the agricultural cycle. 61
4.3 Annual precipitation in 45 provinces, 2001–2007. 64
4.4 Shortfall in precipitation relative to long-term mean, 2004. 65
4.5 Shortfall in crop output relative to 2004–2007 average, 2004. 66
4.6 Reported cuts in food consumption, 2004 and 2005. 66
4.7 Daily per capita consumption of staple food, 2004 and 2005. 67
4.8 Net sales of livestock as share of total livestock holdings, 2004–2005. 69
4.9 Motive for livestock sales, 2004–2007. 69
4.10 Predicted vs. actual prices for sorghum, millet, and maize. 84
4.11 Actual, imputed, and nearest large city-, and capital city-prices for sorghum. . . 85
4.12 Actual, imputed, and nearest large city-, and capital city-prices for millet. 85
4.13 Predicted vs. actual prices for rice, groundnut, niébé, wandzou, sesame, and fonio. 86
4.14 Rainfall levels recorded at eight stations across Burkina Faso, 1970–2009. 87

5.1 Treatment and control states. 96
5.2 OS share against birth year in Lower Saxony. 100
5.3 Pre-school capacity per 100 children between the age of three and six. 109

List of Tables

2.1 Selected proposals for post-2015 global health targets. 8
2.2 ARRs and initial levels of under-five mortality in developing countries, 1960–2010. 11
2.3 Results from POLS regressions: testing for mean reversion in quinquennial ARRs, all developing countries. 15
2.4 Results from POLS regressions: testing for mean reversion in quinquennial ARRs, only countries in Sub-Saharan Africa. 16
2.5 Number of countries by region that are projected to attain an under-five mortality rate of less than 20 per 1,000 by 2030 and 2035. 19
2.6 Number of countries projected to attain a 're-newed' MDG4 by 2030 by region. . 22
2.7 Levels and changes in health service utilization and health behavior in Sub-Saharan Africa, 2000 and 2010. 24
2.8 Results from DPD estimators: testing for mean reversion in quinquennial ARRs. 27
2.9 Percentage of countries projected to achieve an under-five mortality rate of no more than 20 per 1,000 by 2030. 28

3.1 Regression results for log per adult expenditure. 38
3.2 Classification matrix. 39
3.3 TPRs at varying levels of inclusion rates and AUCs. 43
3.4 Selected results from poverty simulations. 49
3.5 Distribution of expenditure and expenditure patterns. 54

4.1 Descriptive statistics: means and standard deviations by year. 62
4.2 Livestock balance for cattle, sheep, and goats, 2004–2007. 68
4.3 Rainfall elasticities of sales and purchases of cattle and sheep/goat, 2004–2007. . 70
4.4 Results from conditional (fixed effects) Poisson models for the number of sales of cattle, sheep, and goat, 2005–2007. 72
4.5 Estimates of the effect of transitory crop income on consumption expenditure. . . 76
4.6 Estimates of the effect of transitory crop income on subsequent changes in grain stocks. 78
4.7 Estimates of the effect of transitory crop income on subsequent net purchases of livestock. 79
4.8 Rainfall elasticities of prices for livestock. 82
4.9 Analysis of station-level rainfall data, 1970–2009. 88
4.10 Tests of over-identifying restrictions: net purchases of livestock. 89

5.1 Means of key variables by state. 99
5.2 Impact of the reform on years of education. 104
5.3 Impact of the reform on the probability of being eligible for university. 105
5.4 Impact of the reform on the probability of graduating from university. 106
5.5 DD estimates of the reform's effect on educational gradients based on grouped data.107
5.6 Robustness to alternative sample definitions. 110
5.7 Impact of the reform on parental involvement. 111
5.8 Provenance of information on state of school visit at the age of ten. 114

Abbreviations

2SLS two stage-least squares

ARI acute respiratory infection

ARR annual rate of reduction

AUC area under curve

BMZ Federal Ministry for Economic Cooperation and Development

CFA Franc de la Communauté Financière d'Afrique

CPI consumer price index

DD difference-in-differences

DPD dynamic panel data

EPA Enquête Permanente Agricole

FEWS NET Famine Early Warning Systems Network

FGT Foster-Greer-Thorbecke

FE fixed effects

FEP fixed effects Poisson

FPR false positive rate

GSOEP German Socio-Economic Panel

GDP Gross Domestic Product

GMM general method of moments

ICRISAT International Crop Research Institute for the Semi-Arid Tropics

IADB Inter-American Development Bank

ITB insecticide-treated bed net

IV Instrumental Variables

KfW Kreditanstalt für Wiederaufbau

MDG Millennium Development Goal

OECD Organisation for Economic Cooperation and Development

OLS ordinary least squares

OS orientation stage

PMT Proxy Means Test

PNGT Deuxième Programme National de Gestion des Terroirs

POLS pooled ordinary least squares

PPP purchasing power parity

PSU primary sampling unit

ROC receiver operating characteristic

SDG Sustainable Development Goal

TER total error rate

TLU tropical livestock unit

TPR true positive rate

TPR true positive rate

UDAPE Unidad de Análisis de Políticas Económicas y Sociales

WASAT West African Semi-arid Tropics

Für Julia.

Chapter 1

Introduction

The four essays that constitute this thesis are, respectively, on (1) realism in international goal setting-exercises, (2) a strategy to tell the poor from the non-poor when data are scarce, (3) the poor's use of livestock as a means to self-insure, and (4) the effects of separating schoolchildren by ability on educational outcomes and equality of opportunity. While the four essays are only loosely connected in terms of their subjects—the first three are chiefly in the field of development economics—, they share two core characteristics: all four take a firmly empirical approach and all four have important policy implications. I will discuss the main ideas, arguments, and implications in what follows.

Essay 1: How the New International Goal for Child Mortality is Unfair to Africa (Again)

The Millennium Development Goals (MDGs) are a list of eight broad goals derived from the Millennium Declaration, a document adopted by an unprecedented number of heads of states at the United Nations' General Assembly in late 2000. The Goals are, in principal, a call to action against human deprivation in some of its most severe forms. But they also came in most instances with specific numerical targets—often multiple targets per goal—and a fixed deadline, the year 2015, by which the targets were to be attained. For instance, MDG1A calls for halving between 1990 and 2015 the proportion of people whose income is less than one dollar per day. Other goals address different forms of human deprivation such as lack of access to primary education (MDG2), gender inequality in education (MDG3), and pre-mature child death (MDG4).

In the first essay in this thesis (chapter 2), Stephan Klasen and I assess the feasibility of current proposals to replace MDG4. The title is a reference to Bill Easterly's (2009) article "How the Millennium Development Goals are Unfair to Africa", in which he shows that most numerical targets were harder to attain for countries in Sub-Saharan Africa than for countries elsewhere. The architects of the MDGs had thus set up Africa for failure. While Easterly is aware that the MDGs are seen by many mainly as a tool for development advocacy, he makes a forceful argument that a greater focus on the challenges Africa faces does not trump concerns over fairness and, in fact, realism when it comes to measuring performance. Framing Africa's progress as a failure, he argues, may eventually result in 'aid-fatigue' and unwarranted Afro-pessimism in donor countries and demoralizes African leaders and activists.

How were the MDGs unfair to Africa? Clearly, level-end goals such as ensuring universal access to a full course of primary education (MDG1) are harder to attain for countries starting from low levels of school enrollment. This is also true for numerical targets defined in relative terms, such as MDG4 on under-five mortality [Klasen and Lange, 2012]. This particular target calls for a reduction by two-thirds in the mortality rate between 1990 and 2015, progress that few countries had managed to bring about in the past. For two reasons, this was an even taller order for high-mortality countries in Africa: first, the 1990s were arguably Africa's "lost decade" in terms of human development and child health was no exception. Civil strife, economic stagnation, and the onset of the HIV/AIDS epidemic had led to an increase in under-five mortality in many African countries between 1990 and 2001. The 1990 baseline thus constituted a severe liability.

The second reason was more subtle: for much of the second half of the last century, there was a negative relationship between relative rates of reduction and mortality levels: the higher under-five mortality, the slower the pace at of countries' transition towards lower levels of mortality. Hence, requiring a uniform reduction by two-thirds seemed ambitious for all countries but even more so for high-mortality countries [Clemens et al., 2007, Easterly, 2009, Klasen and Lange, 2012].

Remarkably though, this correlation had given way by the mid-1990s, an empirical fact noted only recently by McArthur [2014]. The main reason for this was rapid progress in many high-mortality countries, particularly in Africa [Demombynes and Trommlerová, 2012, Rajaratnam et al., 2010], a development that has led to much optimism among commentators. Michael Clemens, for instance, referred to the rapid decline in child mortality rates as "the biggest, best story in development."[1] Given that progress in most African countries was spectacular over the last decade and that, as a consequence, the previous empirical pattern of "unconditional relative divergence" may no longer hold, it now seems that a new target defined in relative terms would no longer be "unfair" to Africa.

However, most stakeholders seem to have taken from praising recent progress to adopting a vision for the post-2015 development goals that entails "ending preventable child death in a generation" [e.g. Glass et al., 2012]—by which they usually mean a rate of no more than 20 deaths per 1,000 live births by 2030 or 2035. While this would clearly be more ambitious for high-mortality countries (of which a large majority are still in Africa), the question remains just how much more ambitious it is. Recent studies have assessed the feasibility of this new proposal by extrapolating from recent rates of progress [McArthur, 2014, Verguet et al., 2014].

In our study, we show that this is misleading. annual rates of reduction (ARRs) show a strong tendency to revert to the mean: above-average ARRs have usually been followed by decelerations. Our projections, based on econometric models that account for mean-reversion, suggest that only two out of 43 countries in Africa will attain the new target. Even if all countries in Africa would sustain high growth rates in Gross Domestic Product (GDP) per capita, a large majority would still fail to attain the new target. And there are more reasons to believe that recent years were exceptional: progress is easier to bring about when public health basics are not in place, which was arguably the case at the end of the 1990s in Africa. Further reductions in child mortality will have to come from more involved interventions that African governments may be in no position to implement.

[1] See `http://www.economist.com/node/21555571`.

The relevant policy question is whether the benefits of a re-newed focus on Africa's challenges will outweigh the fallout from re-casting Africa as a development failure. While this is clearly beyond the scope of our contribution, I think it important to keep in mind that future development goals will affect the way we think about progress. On the other hand, advocacy, no matter how well-intentioned it may be, hardly justifies methodological inaccuracy.

Essay 2: Targeting Performance and Poverty Effects of Proxy Means-Tested Transfers: Trade-offs and Challenges

In the second essay (chapter 3), also co-authored by Stephan Klasen, we turn to a rather specific challenge in alleviating poverty through government programs: how to tell the poor from the non-poor. Clearly, it pays to be the beneficiary of a an anti-poverty program and not only so for the poor. At the same time, resources for such programs are limited and "leakage" of benefits to the non-poor may be costly in both monetary and political terms.

In developed economies, the problem is usually a lesser one as administrators often have access to ample records such as pay stubs and tax records. In developing countries, however, it is common to find that the informal sector accounts for a large share of production. We study a widely-used method to target the poor, so-called Proxy Means Tests (PMTs). PMTs are based on household expenditure surveys that contain information on variables that can be used to predict consumption expenditure. One first stipulates an empirical model of the relationship between these *proxies* and consumption expenditure. Next, a shortened questionnaire is administered to all potential program participants that collects information on the proxies. Finally, consumption is predicted based on the fitted model. Ideally, PMTs are based on a set of indicators that are readily-observable and difficult to manipulate, conditions that we pay due attention to in our study.

Of course, the targeting performance of PMTs has been studied elsewhere [e.g. Grosh and Baker, 1995, Johannsen, 2006, 2008, Kidd and Wylde, 2011, Landau et al., 2012]. Our contribution is two-fold: first, we investigate situations in which the share of households targeted based on a PMT differs from the share that is poor. We use *receiver operating characteristic* (ROC)-analysis to this end [Wodon, 1997], a statistical tool that makes explicit the trade-off between increasing the proportion of poor included in the program and limiting the extent of leakage, i.e., the proportion of non-poor included. Second, we acknowledge recent empirical work that challenges the conventional wisdom that better targeting implies larger poverty impacts or higher cost-effectiveness [e.g. Ravallion, 2007]. To this end, we simulate the (static) poverty effect of PMT-based transfers that target varying proportions of the population under a fixed budget.

We find that in terms of targeting accuracy, PMTs perform reasonably well when it comes to screening out the rich, say, when only the top quintile of the population is to be excluded. They are much less accurate when it comes to narrowly targeting a transfer to the poorest segments of the population (say, the poorest decile). On the other hand, we find that the poverty effect is maximized when only a very small portion of the population is targeted, usually less than ten percent. Our findings thus suggest that among the many trade-offs faced by policy-makers in targeting the poor, the trade-off between targeting accuracy and poverty impact may have received too little attention. It also suggests that while PMT-based targeting may be an attractive option for programs targeting large parts of the population or in combination with additional criteria, alternative strategies—such as self-targeting through the addition of a work requirement [Besley and Coate, 1992b]—may be more appropriate when the goal is to channel transfers to the poorest.

Essay 3: Livestock as an Imperfect Buffer Stock in Poorly Integrated Markets

In the third essay (chapter 4) of this thesis, Malte Reimers and I re-visit a fundamental question in development economics: how do the poor deal with shocks when markets are imperfect? In particular, we investigate whether farmers in rural Burkina Faso, a country frequently hit by droughts, rely on adjustments to livestock holdings in order to smooth consumption, that is, whether livestock is accumulated during good years and sold during bad years.

The importance of this question becomes clear once we consider the economic activities of the world's poor: a majority lives in rural areas and this will likely be the case for many decades to come [Ravallion et al., 2007]. The rural poor, however, mostly rely on rain-fed agriculture to generate incomes—be it directly (e.g. subsistence farmers) or indirectly (e.g. landless agricultural laborers). This is also the case in Burkina Faso, a country in which more than two-thirds of the population live on less than \$2-a-day and agriculture employs about 80 percent of the workforce [World Bank, 2013]. This results in a close link between rainfall and GDP per capita, as is evident from figure 1.1.

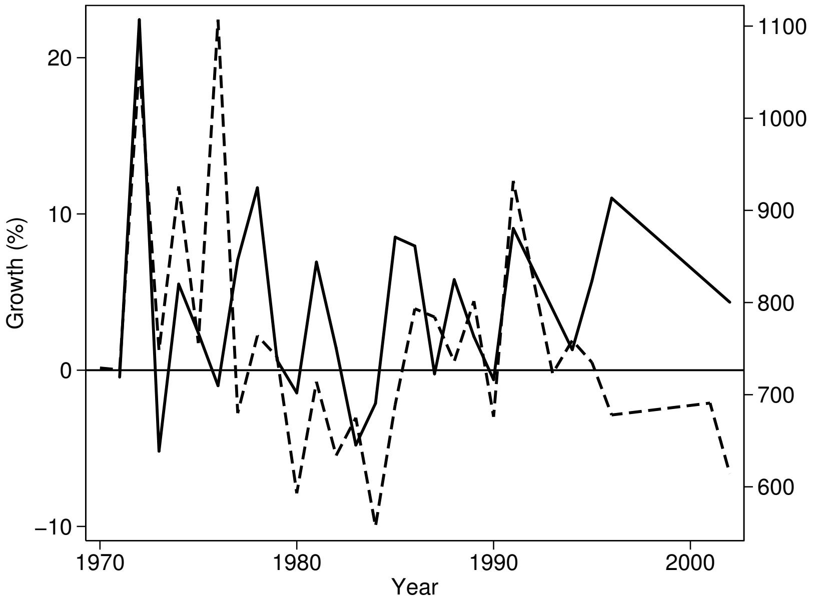

Figure 1.1: Growth of real GDP per capita (solid line) in constant 2005 dollars and annual rainfall (dashed line) in Ouagadougou, Burkina Faso, in millimeters, 1970–2002. Based on data from the Penn World Tables Mark 8.0 [Feenstra et al., 2013] and from station-level data provided by FAO [2014b].

Since rainfall variability is high in in most parts of the developing world and households often lack access to formal insurance instruments, they will try to take precautions in case bad rainfall lowers agriculture output. The overarching question “How do the poor deal with shocks?” has been studied before, both theoretically [e.g. Deaton, 1991, Zeldes, 1989, Zimmerman and Carter, 2003] and empirically [e.g. Carter and Lybbert, 2012, Deaton, 1992, Fafchamps et al., 1998, Kazianga and Udry, 2006, Rosenzweig and Wolpin, 1993]. However, studies that have investigated the role of livestock empirically have come to very different conclusions. Based on panel data from Burkina Faso from the early 1980s, Fafchamps et al. [1998], for instance, find “only very limited evidence that livestock inventories serve as buffer stock against large variations in crop income induced by severe rainfall shocks.” Using the same data, Kazianga and Udry [2006] conclude that grain storage is one way households achieve minimal consumption smoothing while changes in livestock holdings “[...] were not used to smooth consumption.”

Rosenzweig and Wolpin [1993], on the other hand, find for bullocks in rural India that “purchases are significantly more likely to occur when income is high than when income is low, consistent with what appears to be an implication of a consumption-smoothing motive” and Verpoorten [2009] argues that Rwandan households’ likelihood to sell cattle increases “upon the occurrence of a covariant adverse income shock”.

Our paper is thus a clarification: what role do livestock sales play in times of covariant economic shocks if any? We highlight that the conclusions these authors arrive at are driven by the particular choice of empirical approach they use to tackle the question: while Fafchamps et al., Kazianga and Udry, and Carter and Lybbert [2012] all base their analysis on a (net-)savings-regression and thus focus on the revenues generated through livestock sales, Rosenzweig and Wolpin and Verpoorten focus on the number of animals sold.

Using both empirical approaches in two large panel datasets that are representative of rural Burkina Faso, we find that households respond to adverse rainfall by selling livestock, particularly cattle. At the same time, however, there is no evidence that they also manage to increase revenues from sales in the wake of shocks. We argue that price adjustments explain this puzzle: cattle prices decline considerably in response to adverse, covariant rainfall shocks. Hence, households seem to face a delicate trade-off between increasing sales and safeguarding assets that will very likely earn higher returns in the future.

Our findings suggest that households indeed try to smooth consumption yet this come at a very high cost. Prices decline likely because of a lack of market integration. Therefore, improving rural infrastructure may be a way forward as may be safety net programs that stabilize incomes during the lean season (for instance, public works programs). Alternatively, formal insurance instruments may help households in smoothing consumption, particularly index-based weather insurance [see Barnett et al., 2008].

Essay 4: Tracking and the Intergenerational Transmission of Education: Evidence from a Natural Experiment

Finally, in the fourth essay (chapter 5), Marten von Werder and I investigate the effects of an education reform that took place not in the developing world but, in fact, in our home state of Lower Saxony, Germany: in the early 1970s, the state government of Lower Saxony started to introduce a new school type, the *orientation stage*[2] (henceforth, OS), a middle school that

[2] In German: *Orientierungsstufe.*

comprised grades five and six. The most significant change this reform brought about was an increase in the age at which students were separated by academic achievement, what is referred to as *tracking* in the economics of education. Rather than until the end of grade four (by which students are about ten years old), joint teaching of all students would continue for an additional two years in Lower Saxony after the reform but not in most other West German states. We exploit this quasi-experiment by comparing changes in educational outcomes across cohorts and across states between individuals with educated parents and those with uneducated parents.

What effects are to be expected from such a reform? Tracking is often credited with increasing the efficiency of teaching as it allows teachers to tailor lessons more closely to the needs of students [e.g. Duflo et al., 2011]. On the other hand, tracking may lead to higher inequality of opportunity in the sense that the importance of students' social background may play a greater role for their advancement when tracking occurs at an early age [e.g. Hanushek and Woessmann, 2006, Schütz et al., 2008]. In other words, policy makers may face a delicate trade-off between increasing the efficiency of educational production and ensuring equality of opportunity.

Our findings can be summarized as follows: the reform increased educational attainment for individuals with uneducated parents but lowered attainment for individuals with educated parents. Average educational attainment was not altered by the reform. Hence, we conclude that the reform increased equality of opportunity without affecting efficiency, a finding that is in line with previous studies [e.g. Hanushek and Woessmann, 2006]. The effect we find is entirely driven by males; there is no evidence for any kind of effect on females. While this seems plausible in the light of recent evidence on gender differences in cognitive development, it certainly calls for further work on this topic.

The fourth essay thus contributes to a small but growing literature on the long-term effects of tracking reforms. Studies in that literature and closely related to ours include Meghir and Palme [2005], Kerr et al. [2013], and Pekkarinen et al. [2009]. These authors study the effects of de-tracking in Sweden and Finland on educational outcomes and wages, cognitive skills, and the inter-generational transmission of earning potential, respectively. However, in contrast to the reform we study, the Scandinavian reforms comprised other important components that may be driving their results.

Finally, our essay is of particular relevance in the German context today: Germany stands out among countries in the Organisation for Economic Cooperation and Development (OECD) for early tracking. At the same time, parents' socio-economic status is more closely related to own educational achievement than in other OECD countries [e.g. Baumert and Schümer, 2001, Baumert et al., 2003, Dustmann, 2004, OECD, 2003, Schütz et al., 2008].[3] This has led to an ongoing public debate about equity and efficiency within Germany's education system. The OS in Lower Saxony was abolished in 2003, a step that was controversial at the time. A recent reform proposal to extend primary school from four to six years in the city-state of Hamburg faced fierce resistance from a movement of well-to-do parents called "We Want to Learn" (and often referred to as "Gucci protesters" by critics). The proposal was ultimately defeated in a referendum.[4] Our research shows that the protesters may have had a reason to fear for their children's academic success.

[3] Equity in outcomes has improved somewhat recently but the strength of the relationship between socio-economic factors and test scores is still above the OECD average. Recent improvements have been attributed to increasing participation in early childhood education and care [OECD, 2015].

[4] See `http://www.economist.com/node/15073990`.

Chapter 2

How the New International Goal for Child Mortality is Unfair to Africa (Again)

> Even if the health MDGs will not be met in all countries by 2015, the gains point the way to further dramatic reductions in the number of deaths [...]. The MDG health targets need to be retained, updated, and expanded. Preventable child deaths [...] should be ended by 2030.
>
> Sustainable Development Solutions Network [2014]

> [...] with the right investments, the stark differences in [...] child death rates between countries of differing income levels could be brought to an end within our lifetimes. Economic growth in many low-income and middle-income countries and the increasing availability of high-impact health technologies make a grand convergence in health achievable by 2035.
>
> Lancet Commission on Investing in Health [2013]

Abstract Despite unprecedented progress towards lower under-five mortality in high-mortality countries in recent years, a large fraction of these countries will not attain the numerical target under Millennium Development Goal (MDG) 4, a reduction of the mortality rate by two-thirds compared to levels in 1990. Nevertheless, many stakeholders have argued that the post-2015 agenda should contain a level-end goal for under-five mortality and recent accelerations in the rate of reduction in under-five mortality have been cited as a cause for optimism. We argue in this paper that one key fact about relative changes in mortality rates is a lack of persistence. We find robust evidence for substantial mean reversion in the data. Hence, recent accelerations observed for countries in Sub-Saharan Africa are an overly optimistic estimate of future reductions. At the same time, progress as required by the old MDG4 coincides very much with our projections for Sub-Saharan Africa and other regions. Thus, while MDG4 has been rightly criticized as overly ambitious and unfair to Africa for the 1990–2015 period, such a goal seems

The chapter is based on joint work with Stephan Klasen. It has benefited from comments by Michael Clemens and Todd Moss, as well as seminar participants in Göttingen.

more appropriate for the 2005–2030 period. We also offer a discussion of likely drivers of future reductions in child deaths.

2.1 Introduction

The Millennium Development Goals (MDGs) have been a great success in encouraging development. The Millennium Declaration, from which the MDGs derived, represents an unprecedented consensus among world leaders. As a comprehensive list of goals and targets, most of which were supplemented with indicators and a target date for completion, the MDGs have helped focus debates on some of the most pressing challenges facing humankind and have served as a highly effective advocacy tool for those concerned with improving the lives of the poor [Klasen, 2012, Manning, 2010].

The fourth MDG, a call to reduce the rate of under-five mortality by two-thirds relative to levels in 1990 until the end of this year, has received much attention and many countries, particularly in Sub-Saharan Africa, have made exceptional progress towards this goal. Nevertheless, a large fraction of these countries are unlikely to attain this goal and some commentators have argued that MDG4, among other goals, was biased against developing countries and Sub-Saharan Africa in particular [Clemens, 2004, Clemens et al., 2007, Easterly, 2009, Klasen and Lange, 2012].

Table 2.1: Selected proposals for post-2015 global health targets.

	The Lancet Commission on Investing in Health[1]	Global Investment Framework for Women's and Children's Health[2]	UNICEF[3]	Sustainable Development Solutions Network[4]	High-Level Panel of Eminent Persons on the post-2015 Development Agenda[5]
Timeframe	2035	2035	2035	2030	2030
Under-five mortality deaths	16 (interim target of 20 by 2030)	39 in low-, 22 in low-to-middle-income countries.	≤ 20	≤ 20	≤ 20

Adopted from Verguet et al. [2014].

[1] Lancet Commission on Investing in Health [2013].

[2] Global Investment Framework for Women's and Children's Health [2014].

[3] UNICEF [2013].

[4] United Nations Sustainable Development Solutions Network [2014].

[5] High-Level Panel of Emminent Persons [2013].

Nevertheless, recent proposals for the Sustainable Development Goals (SDGs) call for replacing the target defined in relative terms and turn to a global minimum standard, a level-end goal to be attained in 15 to 20 years from now—see table 2.1 which we adopted from Verguet et al. [2014]. Most prominently, the UN Secretary General's High-Level Panel of Emminent Persons [2013, p. 30] has called for an end to preventable infant and under-five deaths by 2030 and has defined this as an under-five mortality rate below 20 per 1,000. The UN's Open Working Group [2014] that negotiated in 2014 a new system of SDGs calls for ending preventable deaths of newborns and children under five years of age with indicators yet to be set. However, those are likely to be similar to the rates reported in table 2.1. Both the United Nations' Sustainable Development Solutions Network [2014] and the Lancet Commission on Investing in Health [2013] have adopted the same target, the latter claiming that "a 'grand convergence' in health is achievable within our lifetimes"—by which they mean the year 2035. Others have been slightly more careful but have also set level-end targets that seem ambitious for high-mortality countries [e.g. Global Investment Framework for Women's and Children's Health, 2014, UNICEF, 2013].

Claims that the new goal is feasible—such as the epigraphs above—have been backed sometimes by costing studies that try to assess what kind of resources it would take to attain the new health target [Boyle et al., 2014, Global Investment Framework for Women's and Children's Health, 2014]. Costing studies have also been conducted following the birth of the MDGs [e.g. Devarajan et al., 2002, High-level Panel on Financing Development, 2001]. However, it is widely understood that these studies are very crude, abstract from institutional constraints in developing countries, and are easily misinterpreted and sometimes misused [Clemens et al., 2007, Klasen, 2012].

In other cases, backing came through empirical assessments. Verguet et al. [2014] project under-five mortality rates in 2030 based on current annual rates of reduction (ARRs) over five years or 'aspirational' ARRs observed in recent years. The latter were calculated as the 90th percentile for all countries of ARRs, which turned out to be between five and 8.3 percent. They estimate that between 50 and 64 percent of countries in their sample would achieve this target by 2030. McArthur [2014] bases projections on twelve-year ARRs and finds that 135 countries are on-track to achieve an under-five mortality rate of 30 or lower by 2030 while 38 are not.

In this paper, we show that the above-cited empirical assessments are flawed. Our main point is that ARRs in under-five mortality rates exhibit considerable mean reversion. Therefore, current ARRs are not adequate estimates for future rates. The resulting geographical distribution of projections until 2030 changes considerably with a clear disadvantage for countries in Sub-Saharan Africa. At the same time, a 're-newed' MDG4—a call for a reduction in under-five mortality rates by two-thirds between 2005 and 2030—results in target rates that are ambitious yet well within reach for nearly all countries.

We also highlight future challenges that need to be overcome in order to further bring down under-five mortality in Sub-Saharan Africa. Economic growth plays a role and high growth rates over the last decade have helped. However, it seems questionable whether high levels of growth can be sustained over the next years. Finally, we will argue that future health interventions will differ considerably from those that were required to bring about the progress we have seen in recent years. Making further inroads will become increasingly more difficult as the focus shifts from low-cost, readily-implementable interventions to policies to improve service delivery and up-take more broadly.

Finding the right trade-off between realism, simplicity, and ambition in setting development goals is of great importance. Unrealistic targets may jeopardize the broad public support the MDG process has received in the past and may cause 'aid-fatigue' in donor countries once it becomes clear that many developing countries will not attain them [Easterly, 2009]. There is also some evidence that realistic targets induce effort from governments, at least if incentives are in place [Öhler et al., 2012]. Simplicity, on the other hand, has been important for the MDGs' success as a communication tool and the same is true for ambition. Finally, targets for the post-2015 period should be applicable to all countries in order to ensure broad buy-in.

Particular care should be applied in deriving an international development goal for under-five mortality. The child mortality-goal has received much attention in the past, most likely because it is of obvious relevance and easy to understand. Conceptually, it has several advantages over alternatives: first, survival is, for all practical purposes, an actual 'end of development' and a necessary requirement for achieving other capabilities that we have reason to value [Sen, 1998]. Second, mortality rates complement records on economic growth as they are responsive to changes in inequality in access to basic services and commodities (ibid.)[1] Finally, under-five mortality is easily measured and the data are widely believed to be of reasonable quality,[2] While this may seem trivial, it is not the case for many other MDG indicators [Attaran, 2005], particularly for those related to health: indicators of food insecurity and undernutrition have been widely criticized [de Haen et al., 2011, Svedberg, 2002]. Data on maternal mortality are estimated from regression-based models that rely on auxiliary variables such as fertility rates and GNI per capita. Different models have led to stark differences, particularly in trends [AbouZahr, 2011].

The paper is organized as follows: the next section reviews recent trends and their relationship to levels and compares these to MDG4. We relate this to the discussion surrounding the appropriateness of the MDGs. The main point is that ARRs have seen a large acceleration during the last decade in some of the high-mortality countries but many are nevertheless not on-track to attain MDG4. This is in part due to the inappropriateness of the numerical target which, however, looks more appropriate today. Section 2.3 establishes that there is a substantial degree of mean reversion in ARRs. The consequences of this are further explored in section 2.4 in which we make projections based on a model that accounts for mean reversion and compare results to projections based on current ARRs. We also compare our projections to target mortality rates that would result from a re-renewal of MDG4. Section 2.5 discusses further issues that are important in trying to gauge the prospects of high-mortality countries in Africa of attaining the new target. Section 2.6 concludes.

2.2 Recent performance and the trouble with the MDGs

We obtain average ARRs over five and ten years from ordinary least squares (OLS)-regressions of log mortality rates on years for different time periods. Note that a greater number implies more rapid progress. The data come from the World Bank's World Development Indicators 2014.

Many developing countries, particularly in Africa, have made rapid progress towards MDG4. Table 2.2 reports ten year-ARRs for under-five mortality for the past five decades and levels in 1990 and 2000. All figures are calculated from regional aggregates for developing countries from the World Development Indicators 2014. The ARRs observed during the 2000s are greater than those for the 1990s in all regions except Latin America and the Caribbean. They have more

[1] Sen [1987] and Drèze and Sen [1989] present data from the UK showing that improvements in life expectancy occurred in decades of slow growth, particularly the war decades, a consequence of improvements in the access to food and public health services and the introduction of the National Health Service in the 1940s.

[2] This is largely the result of an increasing number of comparable cross-country surveys, the Demographic and Health Surveys, that include questions directed at women of child-bearing age about their actual birth histories and thus allow the estimation of mortality rates for children even in the absence of vital registration systems.

Table 2.2: ARRs (in percent) and initial levels of under-five mortality (deaths per 1,000 live births) in developing countries, 1960–2010.

	ARRs					Initial levels	
	1960s	1970s	1980s	1990s	2000s	1990s	2000s
East Asia & Pacific		4.32	2.84	3.12	6.16	59.0	41.9
Europe & Central Asia			4.18	2.25	4.83	55.7	42.3
Latin America & Caribbean	3.20	3.31	4.41	5.12	4.73	55.1	32.8
Middle East & North Africa	1.98	4.86	6.78	3.93	4.84	67.4	44.9
South Asia	1.57	2.11	2.77	3.12	3.83	129.4	94.0
Sub-Saharan Africa		2.26	1.23	1.19	4.13	179.0	156.0

Based on region aggregates from the World Development Indicators 2014.

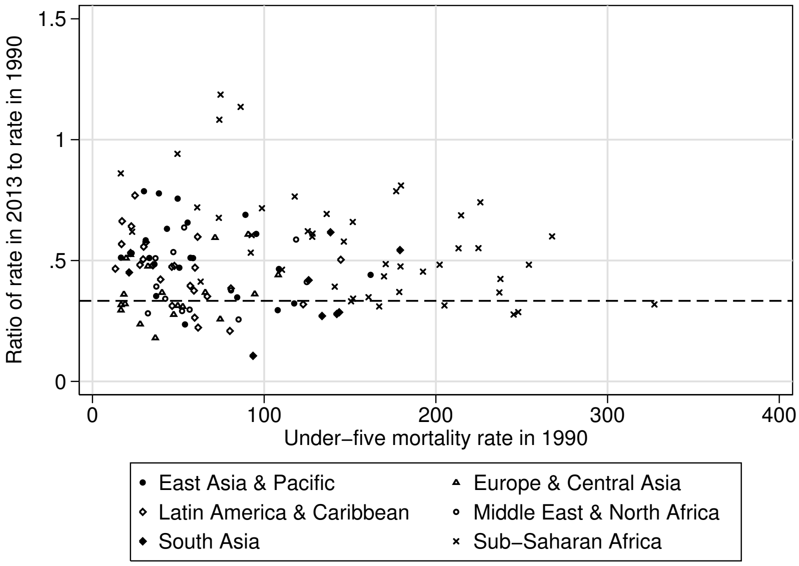

Figure 2.1: Ratio of under-five mortality rate in 2013 to 1990 against rate in 1990. Own calculations based on data from the World Development Indicators 2014.

than tripled in Africa. This has been noted elsewhere. For instance, a recent review of long-term trends in under-5 mortality synthesizing a wide variety of data [Rajaratnam et al., 2010] finds that rates of reduction have increased in 34 countries in Sub-Saharan Africa for 2000–2010 compared with 1990–2000 and have increased by one percent or more in 13 countries.

At the same time, it has been noted that many developing countries will not attain MDG4 and that it may have been unrealistic to assume they would in the first place. The first point is illustrated in figure 2.1 which plots ratios of rates in 2013 to those in 1990 against initial levels. We also indicate the 2015 MDG4-target through the horizontal dashed line that indicates a reduction by two-thirds and implies an ARRs of 4.3 percent between 1990 and 2015. While there are two years left for countries to cross the line, it is clear that many countries will fail to do so. By 2013, only three out of 22 countries in East Asia had attained the goal and only six out of 45 in Sub-Saharan Africa. The best-performing region by this measure was South Asia in which exactly half of the eight countries had attained the goal already.

It is important to note that while the MDGs are frequently interpreted as country-specific goals, they were not conceived in this way. After agreeing on 1990 as the benchmark year, the

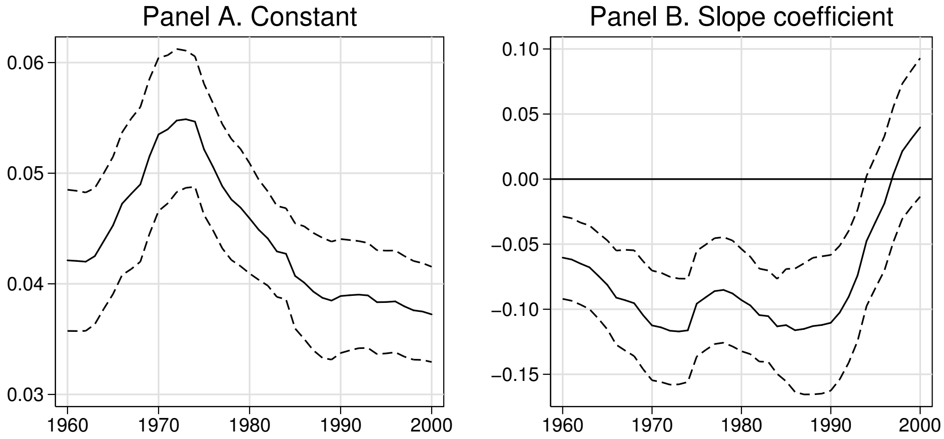

FIGURE 2.2: Estimated constant and slope coefficient from a regression of ARRs over ten years against initial levels, 1960–2003. Dashed lines indicate 95-percent-confidence intervals. Own calculations based on data from the World Development Indicators 2014.

MDG4 target and others were arrived at by linearly extrapolating from global trends [Vandemoortele, 2009]. Thus, the MDGs were to be reached at the global level rather than at the country-level and relied on data from an episode that saw large declines in under-five mortality in the developing world, particularly in Sub-Saharan Africa.

Nevertheless, for lack of further coordination between actors as to how much each country would have to contribute in order to attain MDG4, the numerical target was frequently interpreted as a national policy goal. National governments, NGOs, and even UN agencies frequently compare progress to the required rate for individual countries. The Global Monitoring Reports similarly discuss progress at the country-level [e.g. World Bank, 2011]. It seems very likely that future development goals will equally be interpreted at the country-level as there is much demand for local actors for a yardstick against which to compare their country's progress. It is thus reasonable to contend that future development goals *should explicitly be country-specific.*

If one accepts that stakeholders compared their country's progress against global targets set by the MDGs, two problems immediately become apparent. The first is that the 1990s were in many ways a 'lost decade' for mortality reductions in many developing countries [Easterly, 2009]. This is particularly true for Africa where countries experienced declining aid budgets, civil strife, the onset of high mortality associated with the HIV/AIDS epidemic, and often a combination of these factors. The result were low ARRs for these countries during the 1990s which are also evident in table 2.2. As this was clear by the end of the decade when the MDG targets were conceived, choosing 1990 as the base year dimmed Africa's prospects of attaining MDG4.

The second reason is more subtle: for most of the second part of the last century for which data are available there was a negative relationship between ARRs and initial levels in under-five mortality. High-mortality countries made less progress in relative terms (but more in absolute terms). We can illustrate this point by estimating OLS regressions of ARRs over ten years against initial under-five mortality rates (in percent) for each year in the dataset for which this is possible (i.e. 1960–2003) and including developed and developing countries,[3] e.g.

$$ARR_{i,1960-1970} = \alpha_0 + \alpha_1 u5m_{i,1960} + \epsilon_{i,1960}.$$

Figure 2.2 illustrates the evolution of the constant (panel A) and the slope coefficient (panel B) from this regression over time and 95-percent-confidence bands (dashed lines). The constant is always positive and significantly different from zero. Low-mortality countries—countries that have mortality rates close to zero—have seen ARRs of 3.5–4.5 percent on average with a slight uptick during the early 1970s. ARRs were lower for high-mortality countries for most of the time: between 1960 and 1990, a ten-percentage point higher mortality rate (e.g. going from 100 per 1,000 to 200 per 1,000) was associated with roughly a tenth of a percentage point lower ARR over the subsequent ten years and this relationship was significant at the five-percent level. The slope coefficient increased after 1990 and turned insignificant in 1995.[4]

This indicates that relative changes would have been inappropriate as a target in the past and this is also true for the MDGs which defined 1990 as the relevant baseline year. It may be appropriate now, but this option no longer seems to be on the table. We will come back to this point in section 2.4. It is unclear, however, why the link between initial mortality rates and the ARR broke down in recent years. Theory suggests that ARRs decrease in initial levels and, hence, with progress over time [Klasen and Lange, 2012] and that countries with adverse natural disease environments have lower ARRs [Strulik, 2008]. It is therefore conceivable that the 2000s have been an exceptional episode for high-mortality countries caused by relative stability, high rates of economic growth, increasing aid levels, and a re-newed focus on priority interventions.

2.3 Mean reversion models

If the 2000s were an exceptionally good decade for mortality reductions in Africa, mean reversion would imply that progress is likely to slow in the future. In this section we investigate this issue empirically. Mean reversion (or *regression toward the mean*) is the statistical phenomenon by which random variation in time series data may appear to be a meaningful empirical fact. Its manifestation are unusually large or small measurements that tend to be followed by measurements closer to the mean [e.g. Barnett et al., 2005]. Mean reversion gives rise to several fallacies such as the *Sports Illustrated Cover Jinx* (the notion that teams or athletes that appear on the cover of the *Sports Illustrated* magazine will subsequently experience worse performance) and, similarly, the *Sophomore Slump* (the notion that the "rookie of the year" does less well during the subsequent season) [Schall and Smith, 2000]. More significantly, it sometimes leads to the adaptation of detrimental behaviors and politics (e.g. the apparent empirical finding that rebukes seem to improve performance while praise seems to backfire [Kahneman, 2002]). Mean reversion is ubiquitous in other economic time series, most notably in growth rates, and has been credited with being responsible for "[m]any great economic forecasting errors of the past half century [...]" [Pritchett and Summers, 2014].

We use data on under-five mortality rates for all developing countries that we obtain from the World Development Indicators 2014. Appendix 2.A lists all 132 countries in our sample.

[3] The number of countries changes over time but the overall picture remains unchanged when we restrict the sample to 107 countries for which we have data for all years.

[4] This has also been noted recently by McArthur [2014].

We compute least squares-ARRs for all five year intervals between 1960–1964 and 2005–2009. The base model we estimate using the pooled sample is a regression of ARRs over five years on lagged ARRs (i.e. over previous five-year episodes):

$$ARR_{it} = \beta_0 + \sum_{j=1}^{L} \beta_j ARR_{it-j} + \epsilon_{it}. \tag{2.1}$$

We use OLS to estimate (2.1) and thus refer to the resulting estimator as pooled ordinary least squares (POLS).

For simplicity, consider first the case in which $L = 1$, that is, (2.1) is just a regression of ARRs on ARRs lagged once. The coefficients of interest are both β_0 and β_1: if $\beta_0 = 0$ and $\beta_1 = 1$, the ARR observed over the last five years is an unbiased estimate of the ARR over the next five years. At the other extreme, $\beta_1 = 0$ would suggest that current ARRs do not convey any information about future ARRs and that β_0 is the best predictor available for future ARRs, the most extreme version of regression to the mean. Anything in-between would suggest that current ARRs are of some use for prediction and that there is some regression to the mean. If we observe regression to the mean, i.e. $\widehat{\beta}_1 < 1$, the coefficient signals the rate at which the process reverts to its long-run mean, the estimate of which is $\widehat{\beta}_0/(1 - \widehat{\beta}_1)$.

If we include additional lags, the sum over all coefficients $\beta_1, \beta_2, ..., \beta_L$ conveys information about the extent of mean reversion in ARRs. $\sum_{j=1}^{L} \widehat{\beta}_j < 1$ would suggest mean reversion and the long-run mean is $\widehat{\beta}_0/(1 - \sum_{j=1}^{L} \widehat{\beta}_j)$. We will include up to three lags in the models we estimate.

We also consider more elaborate model specifications: first, we include on the right-hand side of (2.1) a linear time trend in order to account for the possibility of secular trends in ARRs, i.e. a trend in the long-run mean. We center this variable on 2010–2014 so that $\widehat{\beta}_0/(1 - \sum_{j=1}^{L} \widehat{\beta}_j)$ is the mean at that point in time. Second, we include the contemporaneous growth in GDP per capita.[5] Finally, we show in appendix 2.B that the results presented in this section are fairly robust to estimation methods that account for unobserved heterogeneity across countries and contemporaneous correlation. While we find some evidence for the presence of unobserved, country-specific heterogeneity, using results from pooled ordinary least squares (POLS) will only tend to downplay the role of mean reversion.

Results are reported in table 2.3 for all developing countries. Standard errors clustered at the country-level are reported in parentheses throughout. Consider first the simple model with one lag as the sole regressor reported in column (1). Note that the first observation on the outcome is for 1965–1969 as there is no observation on the lagged dependent for 1960–1964. The more lags we include, the smaller the time dimension of our sample.

The *R*-squared of this regression is one-third. Hence, there is evidence that past ARRs predict future ARRs to some extent. However, we also find evidence for mean reversion. The coefficient on the lagged ARR is significantly different from both zero and unity at the one-percent level. It suggests that ARRs revert to their long-run mean at a rate of about 0.6 every five years. The estimate for the long-run mean is an ARR of $0.014/(1 - 0.588) \approx 3.4$ percent. Note that this ARR is much lower than the 4.3 percent required to bring about a reduction in under-five mortality by two-thirds over the course of 25 years, the target under MDG4.

[5] The data come from the Penn World Tables Mark 8.0 [Feenstra et al., 2013]. We use real GDP at constant 2005 national prices since we are not interested in cross-country comparison.

Table 2.3: Results from POLS regressions: testing for mean reversion in quinquennial ARRs, all developing countries.

	1960–2005					1960–1995	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ARR_{it-1}	0.588***	0.717***	0.712***	0.746***	0.695***	0.657***	0.611***
	(0.042)	(0.067)	(0.068)	(0.082)	(0.087)	(0.085)	(0.122)
ARR_{it-2}		-0.277***	-0.273***	-0.402***	-0.339***	-0.164***	-0.253*
		(0.045)	(0.044)	(0.090)	(0.085)	(0.062)	(0.150)
ARR_{it-3}				0.173***	0.119**		0.146
				(0.065)	(0.048)		(0.105)
t (= 0 for 2010–2015)			0.001*	0.001	0.001**	-0.001	-0.001
			(0.000)	(0.000)	(0.000)	(0.000)	(0.001)
$growth_{it}$					0.125***		0.117***
					(0.038)		(0.037)
$Constant$	0.014***	0.019***	0.021***	0.019***	0.017***	0.013***	0.007**
	(0.001)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)	(0.004)
$\sum_{j=1}^{L}\widehat{\beta}_j$	0.588	0.440	0.439	0.518	0.475	0.492	0.504
$\widehat{\beta}_0/(1-\sum_{j=1}^{L}\widehat{\beta}_j)$	0.035	0.034	0.038	0.039	0.033	0.025	0.014
Observations	1,020	888	888	756	633	624	413
Countries	132	132	132	132	110	132	103
R-squared	0.33	0.37	0.37	0.38	0.44	0.33	0.37

Standard errors clustered around the country-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively.

The model reported in column (2) includes one additional lag of the ARR which also turns out highly significant. The long-run mean is very similar to the one we report above. The model reported in column (3) includes a linear time trend and two lags of the dependent. Coefficients on lagged variables are very similar to those found before and remain significant. The coefficient on the trend variable is positive and significant, albeit only at the ten-percent level. It suggest that, on average, ARRs increase over time by about one-tenth of a percentage point every five years. The long-run mean in 2010-2014 is about 3.8 percent. Adding a third lag in column (4) shows that this may also have some predictive content but the change in the R-squared is only marginal. The coefficient on the trend variable is now insignificant.

The model reported in column (5) includes contemporaneous growth rates. All coefficient estimates are now statistically significant at least at the five-percent level. Coefficients on lagged dependent variables remain fairly stable. The coefficient on the growth variable suggests that, *ceteris paribus*, a one percentage point higher growth rate increases the ARR by about one-eighth of a percentage point. The implied elasticity is somewhat lower that those reported in the literature. Pritchett and Summers [1996], for instance, find an elasticity of infant mortality rates with respect to GDP per capita of about -0.2 to -0.4 percent. Finally, in columns (6) and (7), we restrict the sample to observations that predate the MDG era, that is, we exclude all observations on ARRs recorded after 2000. We re-estimate models reported in columns (3) and (5) and find that the results are similar.

As an additional robustness check, we re-estimate all models with data only for the 45 Sub-Saharan Africa in our sample. Results are reported in table 2.4. As one would expect, we find that long-run means are somewhat lower. The sum over estimated coefficients on lagged dependent variables is lower, suggesting less mean reversion. Hence, our main finding of a strong tendency of ARRs to revert to the long-run mean remains intact.

Table 2.4: Results from POLS regressions: testing for mean reversion in quinquennial ARRs, only countries in Sub-Saharan Africa.

	1960–2005					1960–1995	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
ARR_{it-1}	0.548***	0.755***	0.742***	0.778***	0.680***	0.638***	0.567***
	(0.069)	(0.114)	(0.115)	(0.132)	(0.139)	(0.179)	(0.156)
ARR_{it-2}		-0.492***	-0.471***	-0.555***	-0.546***	-0.413***	-0.654***
		(0.059)	(0.057)	(0.081)	(0.081)	(0.096)	(0.186)
ARR_{it-3}				0.116	0.061		0.181
				(0.072)	(0.087)		(0.180)
t (= 0 for 2010–2015)			0.001*	0.002**	0.001*	-0.001	-0.002*
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$growth_{it}$					0.147*		0.156*
					(0.086)		(0.085)
$Constant$	0.011***	0.017***	0.022***	0.022***	0.018***	0.007	0.001
	(0.001)	(0.002)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)
$\sum_{j=1}^{L} \widehat{\beta}_j$	0.548	0.263	0.271	0.338	0.195	0.225	0.094
$\widehat{\beta}_0/(1-\sum_{j=1}^{L} \widehat{\beta}_j)$	0.025	0.022	0.030	0.033	0.023	0.009	0.001
Observations	370	325	325	280	266	235	182
Countries	45	45	45	45	42	45	42
R-squared	0.26	0.39	0.40	0.41	0.46	0.28	0.38

Standard errors clustered around the country-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively.

What does mean reversion imply for the recent discussion surrounding under-five mortality and international development goals? First, recent calculations by McArthur [2014] that suggest that as many as seven million lives may have been saved through setting MDG4 may be flawed. The above results show that mean reversion is a robust characteristic of the data. Hence, relying on the assumption that trends observed during the 1990s would have been persistent absent any goals (or any other intervention) will trivially lead to this conclusion given the dismal performance of African countries during the 1990s. In other words, the improvement in the pace of mortality reduction in the 2000s in Africa is in part due to mean reversion and that part should arguably not be attributed to the MDGs. Second, assuming persistence in currently very high ARRs will result in overly optimistic projections for under-five mortality reductions in countries in Africa. This point will be analyzed in more detail in the section that follows.

2.4 Projections and development targets

2.4.1 Persistence vs. mean reversion

We base our projections on results of the mean reversion model reported in column (3) of table 2.3. For the first set of projections, we retain the relevant parameters from the regression and compute one-period-ahead-ARRs as

$$\widetilde{\mathrm{ARR}}_{it+1}^{MR} = \widehat{\beta}_0 + \widehat{\beta}_1 \mathrm{ARR}_{it} + \widehat{\beta}_2 \mathrm{ARR}_{it-1} + \widehat{\gamma}(t+1). \tag{2.2}$$

By moving (2.2) forward through time, we obtain a series of ARRs that will converge to the long-run mean from which we can calculate the projected under-five mortality rate in 2030 and in 2035.

For comparison, we also calculate projected mortality rates in 2030 and 2035 based on the assumption that the last ARRs that we observe, those recorded between 2005 and 2009, will be persistent in each country individually:

$$\widetilde{\mathrm{ARR}}^{P}_{it+1} = \mathrm{ARR}_{it}.$$

Based on this, it is straight-forward to calculate projected levels of under-five mortality in 2030 and 2035. In principle, this is what is done by McArthur [2014] who extrapolates based on ARRs observed between 2001 and 2013.

Both projections will tend to result in conservative projections for Africa in the sense that they will tend to underestimate under-five mortality in the future. The persistence model, the model that assumes that current ARRs are also future ARRs, relies on ARRs observed between 2005 and 2009. These were higher for most African countries than in most other five-year periods for which data are available. McArthur [2014], for instance, uses twelve-year ARRs in order to "attenuate the risk that some countries might have seen unusually high rates of progress in the most recent years."

The mean reversion model is also conservative for four reasons: first, we base our projections on a model that includes a linear time trend and our results suggest that, on average, ARRs increase over time. Second, the model we rely on for projections does not include country- or region-fixed effects. Therefore, ARRs will tend to converge to a global long-run mean (subject to a positive time trend) rather than country- or region-specific means. The global mean, in turn, was higher than that for Sub-Saharan Africa: the consecutive long-run means estimated from this model for 1960–1964, 1980–1984, 2000–2004, and, out-of sample, 2020–2024 are 0.027, 0.032, 0.036, and 0.040. Comparing these numbers to those reported in table 2.2, we see that Africa's ARR was usually much lower and higher only between 2000 and 2009.[6] Third, the degree of persistence is higher than in models including country-fixed effects (see appendix 2.B). Finally, the model we use does not allow for the pattern of relative divergence that we find for previous decades (see section 2.2).

The results of these alternative projections for the year 2030 are depicted in figure 2.3 where we plot projections from the persistence model against projections from the mean reversion model. Both projections are plotted on log scales. The dashed lines represent the new target, a mortality rate of 20 per 1,000. We also add a 45-degree-line to this graph for reference.

All countries located in the bottom-right corner of the graph will attain the goal if current ARRs turn out to be persistent but will not attain the goal if they experience regression to the mean. On the other hand, countries located in the top-left corner of the graph will not attain the goal under persistence but will attain the goal if there is regression to the mean.

The first thing to note is that there is a strong linear relationship between the two sets of projections. The correlation coefficient is 0.851 and is highly significant (p-value < 0.0001). Based on an unpaired t-test (allowing for unequal variances), we cannot reject the null of equal means (p-value $= 0.884$). Results are similar for projections until 2035 (p-value $= 0.945$). There is no tendency for either approach to result in higher or lower projected mortality rates *on average*.

[6] Our estimates of fixed effects obtained from the model reported in column (9) of table 2.3 differ substantially by region: the mean fixed effect is above 0.030 for all regions except Sub-Saharan Africa where it is only 0.019.

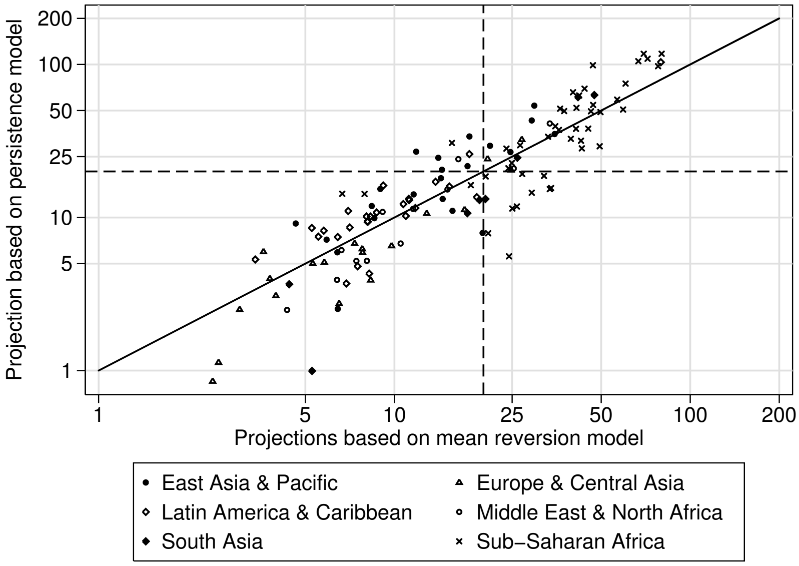

FIGURE 2.3: Projections of under-five mortality based on persistence model against projections based on mean reversion model. The dashed lines indicate the new development goal, an under-five mortality rate of 20 per 1,000. The solid line is a 45-degree-line. Both series are plotted on a log scale. Own calculations based on data from the World Development Indicators 2014.

The geographical distribution of mortality rates, however, differs considerably. Several countries in the bottom-right corner of the graph, mostly located in Africa, are projected to attain the target if current ARRs are also future ARRs but will fail to do so under regression to the mean. On the other hand, countries in the top-left corner, those that will attain the target if they experience regression to the mean but not if there current ARRs persist, are mostly located in East Asia and the Pacific region.

Table 2.5 allows a closer inspection of these regional differences. We tabulate the total number of countries by region, the number of countries that had attained a level of at most 20 per 1,000 already in 2013, and the number of countries that attain the new goal with and without taking into account mean reversion in 2030 and 2035, respectively. Note that more than one-third of all developing countries in our sample had attained the target in 2013—*even before it was agreed on.* The share of countries is particularly high in developing countries in Europe and Central Asia, where only five countries have had under-five mortality rates of more than 20 per 1,000. In Latin America and the Caribbean, more than three out of five countries had attained the target already. The share is particularly low in Africa, where only two out of 45 countries had attained the target by 2013. Thus, the new target would be of little relevance to a large number of countries outside Africa yet relevant to almost all African countries.

The two sets of projections result in a very similar number of countries that will attain the target by 2030 in Europe and Central Asia, in Latin America and the Caribbean, and in the Middle East and North Africa. No matter what method is used for projection, the prospects of most countries in these regions of attaining the target (out of those that have not done so already) seem rather good. On the other hand, large differences between projections are evident for the six countries in South Asia that have not yet attained the target. Only three out of six countries are projected to cross the line under the persistence scenario and only two under mean

Table 2.5: Number of countries by region that are projected to attain an under-five mortality rate of less than 20 per 1,000 by 2030 and 2035.

				Projections							
				2030				2035			
		≤ 20 in 2013		without mean reversion		with mean reversion		without mean reversion		with mean reversion	
	#	#	Share	#	Share	#	Share	#	Share	#	Share
East Asia & Pacific	22	7	0.32	5	0.33	10	0.67	7	0.47	12	0.80
Europe & Central Asia	19	14	0.74	2	0.40	2	0.40	3	0.60	4	0.80
Latin America & Caribbean	26	16	0.62	8	0.80	9	0.90	8	0.80	9	0.90
Middle East & North Africa	12	6	0.50	3	0.50	4	0.67	4	0.67	4	0.67
South Asia	8	2	0.25	3	0.50	2	0.33	4	0.67	3	0.50
Sub-Saharan Africa	45	2	0.04	11	0.26	2	0.05	14	0.33	7	0.16
All	132	47	0.36	32	0.38	29	0.34	40	0.47	39	0.46

Projections without mean reversion are based on the assumption of persistence in ARRs observed between 2005 and 2010. Mean reversion-projections are based on estimates reported in column (3) of table 2.3. Own calculations based on data from the World Development Indicators 2014.

reversion. Large differences between projections are also observed for East Asia and for Africa. Under the persistence scenario, only five out of 15 countries in East Asia and the Pacific are projected to attain the target while ten are projected to do so under mean reversion.[7]

In Africa, eleven out of 43 countries are projected to attain the target under persistence. This would mean that three out of five countries that will have under-five mortality rates in excess of 20 per 1,000 in 2030 are African. Under the mean reversion scenario, only two out of 43 countries are projected to attain the target. This scenario has more than 70 percent of the future 'SDG failures' hailing from Africa.

Could high rates of economic growth put the target within reach for more countries in Africa? In order to quantify the potential contribution of growth and to test the sensitivity of our projections above, we calculate a set of alternative projections based on estimates reported in column (5) of table 2.3 for which we rely on varying assumptions about future growth rates of GDP per capita. Results are reported in appendix 2.C. They suggest that if all countries in Africa manage to maintain a per capita growth rate of five percent per year, still only five percent of the countries in this region for which the target is relevant will have attained it come 2030. This percentage increases to 23 percent for a very high growth rate of ten percent annually. All but eight non-African countries that have not trivially attained the target would cross the line given this rapid pace of economic development. Remember that these projections are conservative in the sense that they will tend to overestimate future ARRs and thus underestimate future mortality rates. Models that account for country-specific long-run ARRs result in a greater degree of mean reversion and lower long-run ARRs for African countries (see the discussion and results in appendix 2.B).

We can also ask whether the general picture would change significantly if one allows for an additional five years to attain the target. Results are reported in the final two columns of table 2.5. Projections based on the mean reversion model are slightly more optimistic in this case: overall, 35 countries are projected to cross the line as opposed to just 27 countries in the mean reversion scenario. At least half of the countries in each individual region that had not attained the target in 2013 already are projected to cross the line. The only exception is Africa where only twelve percent—five countries—will have done so.

Three messages should be taken away from this: first, irrespective of the underlying assumptions, Africa is always the one region in which the smallest share of countries stands a chance at success. Second, mean reversion matters. If only a weak form of it will materialize in the future, the absolute number of countries that will be SDG success stories from Africa will be extremely limited. Third, there is a substantial number of developing countries that has attained the target even before it was sanctioned, but only two of them are in Africa. The proposed new target seems both unfair to Africa and *irrelevant* for many countries elsewhere.

What is maybe even more questionable is that just at a time when a uniform target defined in terms of a common relative reduction seems more viable, actors opt for a target that reproduces

[7] Note that the East Asia and Pacific region is characterized by a large percentage of countries that have already attained under-five mortality rates of less than 20 per 1,000—7 out of 22—and much variation in ARRs and initial levels recently. See appendix 2.A. Countries that are projected to attain an under-five mortality rate of no more than 20 per 1,000 under the mean reversion model but not under the persistence model are Fiji, Micronesia, the Marshall Islands, the Philippines, and the Solomon Islands. However, all these countries are projected to come reasonably close. Cambodia has experienced a very rapid reduction in under-five mortality recently. It is projected to attain the under-five mortality target under the persistence model but not under the mean reversion model.

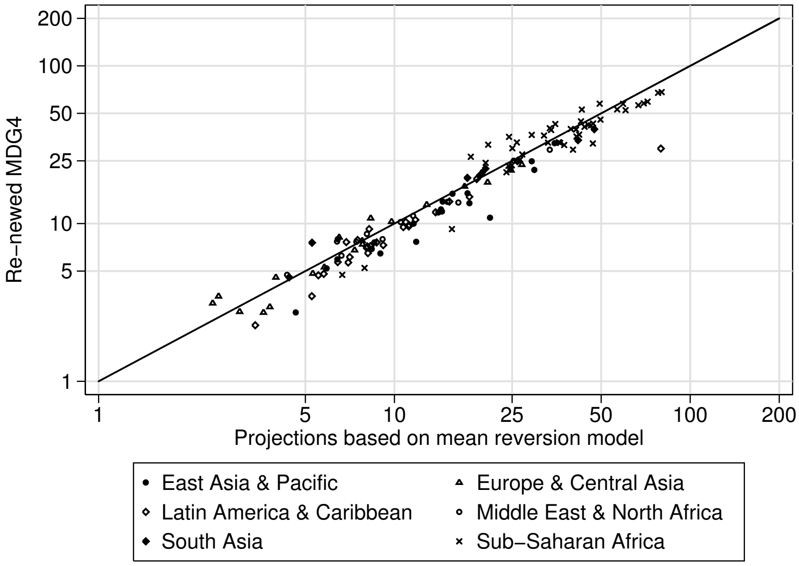

FIGURE 2.4: Under-five mortality rates as required by a 're-newed' MDG4 (two-thirds reduction relative to levels in 2005) against projections based on mean reversion-model. Both series are plotted on a log scale. Own calculations based on data from the World Development Indicators 2014.

biases in the previous child mortality goal that have been noted and criticized before [Clemens et al., 2007, Easterly, 2009, Klasen and Lange, 2012]. In what follows, we will further explore this last point by analyzing how our projections compare to a 're-newed' MDG4.

2.4.2 Old MDG = new SDG

How would our assessment of Africa's prospects change if the new under-five mortality target would be the old target? We have argued above that in 2000, setting 1990 as the base year and calling for a reduction by two-thirds was overly ambitious for most countries in Africa. This time could be different as (1) there is a tendency for ARRs to increase over time, (2) countries in Africa have an advantage in that they have seen high ARRs over the last ten years, and (3) given a degree of persistence in the data, this momentum will carry over to some extent to future ARRs. Given mean reversion, one advantage of a new target defined in relative terms would be that the focus shifts from countries with high levels of under-five mortality to countries with low ARRs in recent years. Re-newing the old target would also have the obvious advantage that it would be immediately relevant for all countries.

Replicating the old MDG would see 2005 as the base year and calling for a reduction by two-thirds until 2030. The target mortality rate is thus

$$M_i^{MDG} = (1/3)M_{i,2005}.$$

Figure 2.4 compares our projections from above (i.e. based on estimates reported in column (3) of table 2.3) to this quantity. Both series are plotted on log scales. Countries above the 45-degree-line are projected to attain the target under mean reversion—the projected mortality

Table 2.6: Number of countries projected to attain a 're-newed' MDG4 by 2030 by region.

		Attaining new MDG4	
	# of countries	#	Share
East Asia & Pacific	22	2	0.09
Europe & Central Asia	19	8	0.42
Latin America & Caribbean	26	4	0.15
Middle East & North Africa	12	4	0.33
South Asia	8	5	0.62
Sub-Saharan Africa	45	16	0.36
All	132	39	0.30

Projections are based on estimates reported in column (3) in table 2.3. Own calculations based on data from the World Development Indicators 2014.

rate is lower than the required mortality rate—while countries below the line are projected to have lower mortality rates in 2030 than would be required under a new MDG4.

Table 2.6 reports the exact numbers of countries that are projected to attain the target by region. Fewer than one-fourth of all developing countries in our sample are projected to attain the target. This suggests that a re-newed MDG4 would still be ambitious today. In comparison to the total, the success rate is higher in Europe and Central Asia, in the Middle East and North Africa, in South Asia, and also in Sub-Saharan Africa. The reason for this is that countries in Africa have seen large ARRs between 2005 and 2009 and this momentum is going to carry over to some extent into the future.

Taken together, this suggests that a re-newed MDG may be ambitious yet in reach for most countries. It would be more demanding of countries that have seen below-average progress in recent years so that the focus would shift away from levels towards recent progress. Of course, projections presented in this section are premised on the assumption that there will be no return to the pattern of relative divergence discussed in section 2.2 which we argued is still very much possible.

2.5 Future challenges

The previous sections have shown that (1) there is substantial evidence for regression to the mean and (2) that accounting for this results in projections that suggest a very different regional distribution of achievements with respect to the new child mortality goal. There are other reasons to expect ARRs to slow down over the coming 15 years which we will discuss in what follows.

2.5.1 Sustaining economic growth

Africa's growth in output is estimated to have exceeded five percent per year since the mid-1990s. This would make the region second only to emerging and developing Asia in terms of recent economic development [IMF, 2014, p. 184]. We have seen above that economic growth is robustly associated with higher ARRs, although our estimate of the elasticity is not particularly high. We have nevertheless argued that sustained growth would significantly improve Africa's prospects with respect to the new child mortality goal (see appendix 2.C). It is thus worthwhile to ask whether growth will continue its role in facilitating progress. We argue that it is unlikely that this will be the case.

One point about Africa's recent growth performance is that we cannot be all too certain about it as the continent's growth data are inherently unreliable [Jerven, 2010]. There have been alternative takes at growth in Africa, notably by Young [2012]. He bases his measurement of real consumption on easy-to-observe indicators such as the quality of housing and the ownership of durable assets and estimates that living standards have been growing at a rate between three and four percent per year. However, it has been questioned whether this approach is valid [Harttgen et al., 2013].

But let us assume for now that Africa's growth performance has rivaled East Asia's for the last ten to 15 years. Will it be possible to maintain these rates of growth? One point that casts doubt on this prospect is the absence of any signs of structural change in Africa in recent years towards manufacturing [McMillan and Rodrik, 2011]. While Africa's labor force has been shifting out of agriculture [McMillan and Harttgen, 2014], they have not found employment in industries associated with high productivity growth. Instead, they relocated to services. While these had above-average productivity levels at the time, they are also associated with low productivity growth [de Vries et al., 2013].

An additional point to keep in mind is that growth in Africa depends on growth elsewhere. Easterly [2001], for instance, speculates that the growth slowdown observed in developing countries during the 1980s and 1990s is largely due to a concurrent slowdown in the developed world. Growth rates in Africa today depend also on growth in East Asia, particularly growth in China, and recent studies have suggested that this region may see a deceleration in the near future [Pritchett and Summers, 2014].[8]

2.5.2 What health policies for future progress?

An important question in the context of our investigation of the usefulness of the newly proposed under-five mortality-goal is from which kind of policies and interventions future reductions in under-five mortality are going to stem from. We argue that (1) much of the low-hanging fruits may have been reaped, that (2) future progress is going to depend on an increased focus on health interventions that require greater state *capacity* (or strength) as opposed to *scope*, and that (3) such activities have historically been found to be harder to bring about in terms of the time required to implement them.

Table 2.7 tabulates levels and changes in selected health indicators between 2000 and 2010 for which aggregate data for Sub-Saharan Africa were available from the World Development Indicators. We classify indicators according to whether they are related to vaccination campaigns, infrastructure investment, or health service use and report figures in this order from top to bottom. What matters for future progress in terms of avoiding premature child deaths is arguably the change in the percentage of the population covered. Therefore, current rates of coverage are an indicator of the potential of these interventions to bring about reductions in under-five mortality in the future.[9] Table 2.7 shows that vaccination coverage in Africa is now at

[8] The reason cited for this is the same as the one that we cited for a slowdown in ARRs—growth rates show an even higher degree of mean reversion.

[9] We simplify to some extent. Even with full coverage there may be technological advances such as more effective vaccines and longer-lasting bed nets that may bring about further gains.

Table 2.7: Levels and changes in health service utilization and health behavior in Sub-Saharan Africa, 2000 and 2010.

	2000	2010	Difference
Vaccinations			
DPT (% of children ages 12–23 months)	51.63	72.01	20.37
BCG (% of one-year-old children)	68.10	82.21	14.11
Measles (% of children ages 12–23 months)	52.69	73.55	20.87
Infrastructure			
Improved water source (% of population with access)	55.03	62.61	7.58
Improved sanitation facilities (% of population with access)	25.92	29.08	3.16
Service use			
ARI treatment (% of children under five taken to health provider)	40.51	51.79	11.28
Births attended by skilled health staff (% of total)	42.08	49.72	7.64

Own calculations based on data from the World Development Indicators 2014.

comparatively high levels. Future progress may have to stem primarily from interventions aimed at improving infrastructure (e.g. access to safe drinking water) and up-take of health-related services (e.g. births delivered through skilled health staff), which are still at lower levels.

What is required to bring about improvements in these fields? Health interventions differ in terms of what kind of institutional environment has to be in place in order for them to be effective. Some interventions require resources but little government capacity. One-shot vaccines and the distribution of bed nets against malaria which do not require highly-skilled health personnel are prime examples [Fukuyama, 2004]. From table 2.7 we see that much progress has been made in vaccination coverage recently. There is also some evidence that the distribution of insecticide-treated bed nets (ITBs) accounts for large portions of the decline in infant mortality rates in some places. Demombynes and Trommlerová [2012] attribute half of the decline in this indicator to the use of ITBs which increased from eight to 60 percent between 2003 and 2008. Overall, it is plausible that high rates of economic growth in combination with readily-available, low-cost, and low-tech interventions account for much of the recent decline in under-five mortality rates in Sub-Saharan Africa.

Other treatments require substantially higher government capacity. Anti-retroviral drugs, for instance, must be taken in complex doses over lengthy time periods. Effectiveness thus requires a strong public-health infrastructure [Fukuyama, 2004]. Another example are small-scale infrastructure projects to improve water quality at the point-of-use. These interventions are generally effective in preventing diarrhea in children under five [Clasen et al., 2007] which, in turn, is the second leading cause of death in children under five [WHO, 2014]. The lack of safe drinking water also contributes to the spread of water-borne diseases. However, the sole construction of wells, a prime activity of governments and external development agencies in recent years, is only the first step. Wells require frequent maintenance and sometimes repair which, in turn, requires some technical know-how and the capability to provide such public goods. Surveys of the functionality of wells in Africa and other developing regions frequently find that a large number is in need of repair. For instance, Miguel and Gugerty [2005] use data on wells constructed with the assistance of the Finnish government in Western Kenya. As is common in such projects, foreign assistance was limited to the construction of the wells and provision of the equipment required for operation while actual maintenance was the responsibility of local well committees. Despite the importance of access to safe drinking water, they report that only 57 percent of these wells were fully functional in 2000/2001, suggesting wide-spread failures of local collective action in well maintenance. Such projects are therefore a prime example of interventions that require some capacity in order to be sustainable.

Government capacity is also key to a functioning health care system, in turn a prerequisite for high levels of up-take. Table 2.7 tabulates levels and changes in the proportion of children with acute respiratory infections (ARIs) taken to a health provider and the proportion of births attended by skilled personnel. These figures suggest that there is still some room for improvements in the future through changes in health-seeking behavior. The quality of health services in many developing countries is, however, poor. Based on data on medical care quality from four developing countries Das et al. [2008] find that not only is the competence of doctors in these countries low. Doctors also exert less effort and thus do not bring to bear what little knowledge they have. In Tanzania, for instance, doctors complete only 24 percent of the essential checklist when faced with a patient with malaria and only 38 percent when the patient is a child with diarrhea [Leonard and Masatu, 2007]. Similarly, Reyburn et al. [2004] find less than half of all patients treated for malaria in hospitals in northeast Tanzania were actually suffering from malaria. As long as these quality-issues persist, the effectiveness of improving treatments is severely limited. Solving this kind of incentive problem requires institutional change, a challenge that while surmountable, is arguably more difficult to bring about [World Bank, 2004].

2.6 Conclusion and interpretations

Compared to earlier efforts of international goal-setting, one of the major innovations of the MDG process has been the introduction of numerical targets, some of which are measurable and time-bound. This has been tainted somewhat by the insight that attaining most targets, including MDG4, was all but impossible for the poorest countries. If international goal-setting for the developing community is to remain a successful venture, both with respect to galvanizing support and bringing about policy changes, future goals should not only be ambitious and easy to communicate, they should also be realistic.

Just as targets for under-five mortality defined in relative terms are starting to look more reasonable than in the past, recent proposals have called for a recasting of MDG4 as a level-end goal, an absolute minimum standard that most commentators put at about 20 deaths per 1,000 by 2030. This paper shows that these recent proposals are overly ambitious for countries in Africa and that empirical studies that demonstrate their alleged feasibility by extrapolating from recent rates are likely misleading. At the same time, the prospective target is not relevant for all countries as many developing countries have already attained it.

It may thus be that the new goal-setters got carried away by recent progress and are now running the danger of committing the same mistake as with the MDGs: setting unrealistic mortality targets for Africa.

2.A List of countries in the analysis

We list here all 132 developing countries in our analysis along with the ARR (in percent) observed between 2005 and 2010 and the under-five mortality rate in 2013 (printed bold whenever it is no more than 20 per 1,000 in that year). All data come from the World Development Indicators 2014.

East Asia & Pacific: Cambodia (7.9, 38); China (8.4, **13**); Fiji (-0.6, 24); Indonesia (4.4, 29); Kiribati (0.7, 58); Korea, Dem. Rep. (0.3, 27); Lao PDR (3.9, 71); Malaysia (-0.4, **9**); Marshall Islands (0.7, 38); Micronesia, Fed. Sts. (3.0, 36); Mongolia (5.5, 32); Myanmar (3.6, 51); Palau (3.2, **18**); Papua New Guinea (2.1, 61); Philippines (2.2, 30); Samoa (0.9, **18**); Solomon Islands (1.4, 30); Thailand (4.3, **13**); Tonga (3.0, **12**); Tuvalu (2.8, 29); Vanuatu (2.2, **17**); Vietnam (2.9, 24).

Europe & Central Asia: Albania (4.3, **15**); Armenia (5.1, **16**); Azerbaijan (5.8, 34); Belarus (9.0, **5**); Bosnia and Herzegovina (1.1, **7**); Bulgaria (4.5, **12**); Georgia (8.2, **13**); Hungary (4.6, **6**); Kazakhstan (7.9, **16**); Kyrgyz Republic (5.0, 24); Macedonia, FYR (5.5, **7**); Moldova (5.1, **15**); Montenegro (8.2, **5**); Serbia (3.1, **7**); Tajikistan (4.5, 48); Turkey (5.9, **19**); Turkmenistan (3.1, 55); Ukraine (4.1, **10**); Uzbekistan (3.2, 43).

Latin America & Caribbean: Argentina (3.2, **13**); Belize (3.2, **17**); Bolivia (5.6, 39); Brazil (6.9, **14**); Colombia (2.9, **17**); Costa Rica (0.8, **10**); Cuba (1.2, **6**); Dominica (2.5, **11**); Dominican Republic (2.8, 28); Ecuador (3.0, 23); El Salvador (6.1, **16**); Grenada (2.2, **12**); Guatemala (3.7, 31); Guyana (2.1, 37); Haiti (2.5, 73); Honduras (4.3, 22); Jamaica (2.9, **17**); Mexico (2.8, **14**); Nicaragua (3.9, 23); Panama (2.9, **18**); Paraguay (3.3, 22); Peru (7.0, **17**); St. Lucia (1.7, **14**); St. Vincent and the Grenadines (1.2, **19**); Suriname (3.2, 23); Venezuela, RB (3.0, **15**).

Middle East & North Africa: Algeria (4.2, 25); Djibouti (3.0, 70); Egypt, Arab Rep. (5.9, 22); Iran, Islamic Rep. (6.1, **17**); Iraq (2.1, 34); Jordan (3.1, **19**); Lebanon (6.7, **9**); Libya (6.8, **14**); Morocco (3.9, 30); Syrian Arab Republic (4.3, **15**); Tunisia (5.7, **15**); Yemen, Rep. (4.9, 51).

South Asia: Afghanistan (2.5, 97); Bangladesh (6.2, 41); Bhutan (6.5, 36); India (4.3, 53); Maldives (11.4, **10**); Nepal (5.9, 40); Pakistan (2.0, 86); Sri Lanka (5.1, **10**).

Sub-Saharan Africa: Angola (2.2, 167); Benin (4.4, 85); Botswana (3.0, 47); Burkina Faso (6.5, 98); Burundi (5.0, 83); Cabo Verde (-0.5, 26); Cameroon (3.6, 94); Central African Republic (1.8, 139); Chad (1.9, 147); Comoros (2.5, 78); Congo, Rep. (9.2, 49); Cote d'Ivoire (3.4, 100); Eritrea (4.6, 50); Ethiopia (7.6, 64); Gabon (3.7, 56); Gambia, The (3.8, 74); Ghana (1.2, 78); Guinea (4.0, 101); Guinea-Bissau (2.9, 124); Kenya (4.1, 71); Lesotho (2.2, 98); Liberia (7.7, 71); Madagascar (5.6, 56); Malawi (7.8, 68); Mali (4.7, 123); Mauritania (2.2, 90); Mauritius (0.3, **14**); Mozambique (5.5, 87); Namibia (5.2, 50); Niger (6.7, 104); Nigeria (3.8, 117); Rwanda (10.8, 52); Sao Tome and Principe (4.4, 51); Senegal (7.9, 55); Seychelles (0.0, **14**); Sierra Leone (2.9, 161); Somalia (1.5, 146); South Africa (5.6, 44); Sudan (2.5, 77); Swaziland (4.2, 80); Tanzania (7.7, 52); Togo (2.8, 85); Uganda (6.7, 66); Zambia (4.7, 87); Zimbabwe (-0.1, 89).

2.B Results from dynamic panel data (DPD) estimators

Table 2.8 presents results from more sophisticated dynamic panel data (DPD) estimation methods. The models we estimate include two lags of the dependent as regressors and, in one specification, the contemporaneous growth rate of Gross Domestic Product (GDP) per capita. The reason for this choice is that we find no evidence for serial correlation in the error terms for these specifications, indicating that the dynamics of these models are not misspecified.

As a first robustness check, we replace linear trends with time-fixed effects. This accounts for contemporaneous correlation, often the most important form of cross-unit correlation in these applications. Results from estimating these specifications via POLS are presented in columns (1) and (4) of table 2.8, respectively. Coefficient estimates on lagged terms are very similar to those we saw in section 2.3 (e.g. coefficients reported in column (3) of table 2.3). p-values of

Table 2.8: Results from DPD estimators: testing for mean reversion in quinquennial ARRs, all developing countries.

	(1)	(2)	(3)	(4)	(5)	(6)
	POLS	FE	Diff-GMM	POLS	FE	Diff-GMM
ARR_{it-1}	0.712***	0.553***	0.613***	0.700***	0.534***	0.643***
	(0.067)	(0.078)	(0.110)	(0.069)	(0.079)	(0.108)
ARR_{it-2}	-0.242***	-0.381***	-0.377***	-0.226***	-0.359***	-0.305***
	(0.042)	(0.054)	(0.083)	(0.049)	(0.060)	(0.069)
$growth_{it}$				0.098***	0.112***	0.076*
				(0.033)	(0.039)	(0.042)
$\sum_{j=1}^{L} \widehat{\beta}_j$	0.471	0.173	0.236	0.474	0.175	0.338
Observations	888	888	756	736	736	626
Countries	132	132	132	110	110	110
R-squared	0.41	0.52		0.47	0.57	
Hansen-test for exogeneity of instruments:						
Exogen. instruments			10			11
p-value of Hansen test			0.29			0.75
p-values of Arellano-Bond-test for serial correlation:						
$AR(1)$	0.571	0.359	0.000	0.866	0.115	0.000
$AR(2)$	0.325	0.191	0.365	0.700	0.316	0.798
$AR(3)$	0.771	0.000	0.279	0.871	0.000	0.490
$AR(4)$	0.935	0.000	0.185	0.994	0.000	0.220

Standard errors clustered around the country-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. All models include a complete set of time-fixed effects.

the Arellano-Bond-test [Arellano and Bond, 1991] suggest that there is no evidence for serial correlation of orders one through four.

One additional concern may be unobserved heterogeneity across countries. If the error term in (2.1) includes a country-specific component, POLS will usually result in upward-biased coefficient estimates as both the dependent and its lags will be positively correlated with the composite error [Nickell, 1981]. Note, however, that upward-biased estimates are conservative in our application as they would suggest less mean reversion in the data. The standard approach to deal with unobserved, time-invariant heterogeneity when panel data is available is to include country-fixed effects. However, this is usually no recipe against dynamic panel bias [Bond, 2002, Nickell, 1981]. Including fixed effects is equivalent to running OLS on the de-meaned data and, by construction, the de-meaned error term will be negatively correlated with the de-meaned lagged dependent variables. The bias will be decreasing in T but since T is quite short in our case, we may expect the bias in fixed effects (FE)-estimates to be significant. Roodman [2009b] argues that reporting results from FE estimation is nevertheless useful as coefficient estimates on the lagged dependent from POLS and FE models usually provide a plausible range for the actual coefficient of interest. The same applies for the sum over coefficient estimates, an indicator of the extent of mean reversion. We report results from FE models in columns (2) and (5) and find that the estimate on the once-lagged dependent is indeed lower and so is the sum over coefficients. The plausible ranges implied by the POLS and FE estimates is narrow—0.173–0.471 and 0.175–0.474. Note, however, that there is evidence for higher-order serial correlation, suggesting that FE models suffer from specification problems.

Finally, we employ general method of moments (GMM)-estimators in order to obtain consistent estimates of the dynamics of ARRs. We use Arellano and Bond's (1991) Difference-GMM

Table 2.9: Percentage of countries projected to achieve an under-five mortality rate of no more than 20 per 1,000 by 2030 under alternative assumptions about economic growth.

		Growth rates (percent)				
	#	3.5	5.0	6.5	8.0	10.0
East Asia & Pacific	15	0.73	0.73	0.73	0.73	0.80
Europe & Central Asia	5	0.40	0.60	0.60	0.80	0.80
Latin America & Caribbean	10	0.90	0.90	0.90	0.90	0.90
Middle East & North Africa	6	0.67	0.67	0.67	0.67	0.83
South Asia	6	0.50	0.50	0.50	0.50	0.67
Sub-Saharan Africa	43	0.02	0.05	0.09	0.14	0.23

Projections are based on estimates reported in column (5) in table 2.3. Own calculations based on data from the World Development Indicators 2014.

estimator that relies on first-differencing for purging fixed effects and instrumenting the differenced lagged dependent with lagged levels.[10] It is often important to keep the instrument count low as too many instruments relative to observations will tend to result in finite sample-bias [Roodman, 2009a]. Therefore, we use as instruments only the 'collapsed' set (i.e. we do not employ separate instruments for each time period) and only the second, third, and fourth lag of the dependent variable. This results in ten and eleven instruments, respectively.

Results are reported in columns (3) and (6). As one would expect, the sum over both coefficient estimates on the lagged dependent variables falls between those for POLS and FE results. However, they are closer to FE results suggesting that we over-estimate the extent the persistence in ARRs in our projections in section 2.4. There is no evidence for serial correlation (except of order one which is to be expected after differencing) and a Hansen-test of overidentifying restrictions suggests that we cannot reject exogeneity of instruments.

Overall, our results in this appendix indicate that while there is some evidence for the presence of a country-specific, time-invariant component in the error term, the resulting bias in POLS estimates still seems to be moderate. Note that using consistent estimates for projections is not an option as fixed effects are purged rather than estimated. It is important to remember that upward biased coefficient estimates and the omission of country-specific long-run means from the model only render our projections conservative in the sense that we will tend to under-estimate under-five mortality rates in Africa in the future.

2.C Alternative projections

We derive a set of additional projections that are based on regression results reported in column (5) in table 2.3, i.e. the model that includes the first, second, and third lag of the dependent variable, a linear time trend, and the average rate of growth of GDP per capita over five years. We make alternative assumptions about future growth rates of GDP per capita, 3.5, 5, 6.8, and 8 percent, respectively, and calculate the share of countries in each region that would achieve an under-five mortality rate of 20 or less in 2030. Results are reported in table 2.9. Note that the

[10] A popular alternative is System-GMM [Arellano and Bover, 1995, Blundell and Bond, 1998] but this requires that differences are orthogonal to the fixed effects. Since African countries have had both lower ARRs and have seen a large acceleration in ARRs recently, this assumption is unlikely to hold. See Roodman [2009a] for details.

number of countries in the second column refers to the number of countries that had not already achieved an under-five mortality rates of no more than 20 per 1,000.

Results indicate that even under very favorable economic circumstances until 2030, the target would be biased against countries in Africa. Even if all countries in the region managed to bring about a growth rate of ten percent per year, a scenario that seems extremely unlikely (see section 2.5.1), less than one-fourth of the countries in Africa that have not done so already today would attain an under-five mortality rate of less than 20 per 1,000 deaths. The success rate is projected to be less than ten percent if countries in Africa managed to grow at a rate of 6.5 percent per year, one percentage point higher than the growth rate recorded between 1996 and 2005 [IMF, 2014, p. 184].

Chapter 3

Targeting Performance and Poverty Effects of Proxy Means-Tested Transfers: Trade-offs and Challenges

Abstract In the absence of reliable and exhaustive income data, Proxy Means Tests (PMTs) are frequently employed as a cost-effective way to identify income-poor beneficiaries of targeted anti-poverty programs. However, their usefulness depends on whether proxies accurately identify the income poor. Based on receiver operating characteristic (ROC)-analysis, we find that PMTs perform poorly in terms of identifying poor households in Bolivian data when transfers are targeted narrowly to the poor but that the true positive rate is highly responsive to increases in the proportion of beneficiaries. Using non-parametric regression-techniques, we show that the resulting leakage can largely be confined to the non-poor close to the poverty line. However, simulating the effect on poverty measures of a uniform transfer to beneficiaries across inclusion rates suggests that the largest poverty effect is attained with very narrow targeting. Hence, we find a trade-off between targeting accuracy and poverty effect.

3.1 Introduction

The idea of targeting social transfer programs to those in need is intuitively appealing—in the words of Amartya Sen: "the more accurate a subsidy in fact is in reaching the poor, the less the wastage, and the less it costs to achieve the desired objective. It is a matter of cost-effectiveness in securing a particular benefit. [...] it is one of maximizing the poverty-removal benefits accruing from a given burden of cost" [Sen, 1995].[1] If the argument in favor of targeting takes precedent

The chapter is based on joint work with Stephan Klasen. We are grateful to Julia Johannsen, Fernando Landa, Roland Pardo, Ramona Rischke, and Cecilia Vidal for their valuable comments.

[1] Akerlof [1978] shows formally that targeting, what he refers to as 'tagging', increases the level of benefits received by the poor as it reduces marginal tax rates and thus costs associated with income redistribution.

over other concerns,[2] the question is how to identify the *targets* in this exercise, i.e. how to identify households that have insufficient *means*. While direct means tests using verifiable data on incomes (e.g. from tax records) are usually viable in developed countries, identifying the needy is much more challenging in developing countries where a larger share of total income is generated in the urban informal sector and where people in rural areas tend to rely on subsistence agriculture for a living. Verifiable data on incomes or other direct measures of economic welfare are thus often lacking for the majority of the population.

In such a setting, one is forced to choose from imperfect methods that do not require complete expenditure or income data. For instance, schemes based on *geographic targeting* channel resources to regions in which the extent of poverty seems greatest [Bigman and Fofack, 2000, Schady, 2002]. Alternatively, one may try to tap into local knowledge by delegating the selection of beneficiaries to locals, a strategy referred to as *community-based targeting* [Bardhan and Mookherjee, 2005, Conning and Kevane, 2002, Galasso and Ravallion, 2005]. Workfare schemes [e.g. Besley and Coate, 1992b, Dutta et al., 2012] and the subsidization of goods and services consumed primarily by the poor are *self-targeting mechanisms* [Alatas et al., 2013, Besley and Kanbur, 1988]. Participants incur some sort of utility loss that makes participation attractive only for the deserving. In practice we often observe that several of these methods are combined.

A further alternative is to consider a number of readily observable and verifiable household characteristics to construct a Proxy Means Tests (PMTs).[3] PMTs employ a limited number of variables in order to predict the welfare-level of households.[4] A statistical model, usually a linear model, is first calibrated with data from a representative household expenditure survey. A shortened questionnaire is then administered to a much larger set of households. In conjunction with the statistical model, data from the shortened questionnaire are finally used to derive scores for the identification of beneficiaries. Whether this strategy is suitable for the identification of the deserving poor is largely an empirical question that depends on the underlying model, the quality of the available data, the joint distribution of the target measure and the proxies, and—as will be demonstrated in the present paper—the proportions of the population in poverty and targeted to receive the program.

We analyze the potential for PMTs based on two waves of Bolivian household data, the *Encuesta de Hogares* 2008 and 2011. Based on receiver operating characteristic (ROC)-analysis, a statistical tool to assess the accuracy of classifiers, we compare regression-based Proxy Means Tests (PMTs) that differ in terms of complexity.[5] The paper thus contributes to the on-going debate about the usefulness of PMTs [e.g. Alatas et al., 2012, Grosh and Baker, 1995, Kidd and Wylde, 2011]. However, in contrast to these studies, the analysis of ROC-curves allows us to

[2] There is a general debate on whether transfers should be confined exclusively to the poor or whether universal transfers are to be preferred. Chief among the concerns about targeted transfers are incentive problems. This starts with the familiar notion of the 'iron triangle' of welfare, which holds that transfers that remove poverty in a cost-effective way will involve major incentive problems. Moreover, the political economy-literature on the subject often argues that transfers targeted to the poor exclusively will lack political support [e.g. Besley and Kanbur, 1990, Gelbach and Pritchett, 2000, 2002, Moene and Wallerstein, 2001]. Other concerns include social costs such as welfare stigma [e.g. Besley and Coate, 1992a, Moffitt, 1983] and the loss of privacy [Sen, 1995].

[3] PMTs are widely employed—especially in Latin America—with Chile the first country to base targeting of its social pension and disability scheme on a PMT [Lindert et al., 2006, e.g.]. Other examples include Colombia [Canstañeda, 2005] and Mexico [Skoufias et al., 2001].

[4] Grosh and Baker [1995] and Kidd and Wylde [2011] provide systematic assessments of PMTs. Coady et al. [2004] provide a meta-study comparing PMTs and alternative targeting strategies in the developing country-context in terms of targeting accuracy.

[5] Wodon [1997], Johannsen [2008], and Landau et al. [2012], *inter alia*, discuss ROC-analysis in the context of poverty analysis and targeting.

illustrate targeting outcomes over the entire range of options that are available to policy-makers, i.e. the typical trade-off in terms of under-coverage of the poor on the one hand and leakage of benefits to the non-poor on the other.[6] We then take the analysis one step further by going from targeting measures such as error rates—often the only criterion to evaluate the accuracy of targeting mechanisms—to simulations of poverty effects of uniform transfers, the most common form of cash transfer in developing countries.[7]

Our findings suggest that PMTs are reasonably accurate in the Bolivian setting when a large fraction of the population is considered poor and an equally large share is targeted. For instance, our estimates suggest that only 52–60 percent of the poor are covered when ten percent are poor and ten percent are targeted compared to 70–78 percent when 50 percent are poor and an equal share is targeted. While this is very much in line with the literature,[8] we also find that the proportion of poor covered is responsive to an increase in the inclusion rate in the latter scenario: when 70 percent are targeted—despite only 50 percent being poor—around 90 percent of the poor are covered. In other words, leakage in this case is confined mostly to the poor close to the poverty line. We illustrate this point further based on non-parametric regression methods.

Our main finding, however, relates to the discrepancy between targeting accuracy and poverty impacts: our simulation study suggests that significant poverty effects are only attained with very narrow targeting. This is true for both the poverty gap and the squared poverty gap and for different initial poverty headcounts. There is thus an important trade-off between targeting accuracy, in particular, coverage of the poor, and poverty effects of transfer schemes that has received very little attention in the literature. One exception is Ravallion [2007] who finds no evidence for an empirical relationship between various measures of targeting accuracy and indicators of the poverty effect and cost-effectiveness of China's *Di Bao*-transfer program.[9]

It also emerges that increasing the number of proxies (i.e. going from parsimonious models to models employing more proxies) exhibits quickly decreasing returns in terms of accuracy. More importantly, parsimonious PMTs based on few easy-to-verify proxies such as geography and demographics perform worse in terms of targeting accuracy yet outcomes do not differ much from more sophisticated PMTs when it comes to poverty effects. Moreover, calibrating PMTs based on data from 2008 in order to identify poor households in the 2011 data instead of a subset of the 2011 data has a negligible effect on both targeting accuracy and poverty effects.

The remainder of the paper is organized as follows: in section 3.2, we review different targeting strategies and compare them to PMT-based targeting. Section 3.3 describes our datasets and our indicator of economic welfare. Section 3.4 introduces the proxy sets on which the remaining analysis is based and reports results from calibrating models. Section 3.5 introduces ROC-analysis and reports results from out-of-sample evaluations of PMTs in terms of targeting

[6] For instance, Kidd and Wylde [2011] do not consider situations in which the proportion of individuals in poverty differs from the proportion targeted.

[7] We consider only the direct poverty effects of transfers that arises out of the arithmetic of adding the transfer to existing household consumption expenditure, not the ultimate impact of the transfer which would depend on behavioral responses of beneficiaries.

[8] Kidd and Wylde [2011] argue that PMTs have a large margin of error when few are poor and an equal number is targeted. In data from Bangladesh, Rwanda, Sri Lanka and Indonesia, they find that when ten percent of the population is poor and an equal share is selected for transfers based on PMTs, about 60–70 percent of beneficiaries are 'false positives.' This fraction decreases to about 30–40 percent when the poverty headcount is 40 percent and an equal share is eligible.

[9] Ravallion's study differs from ours in that he considers an actual transfer program and compares poverty effects based on regression analysis at the level of municipalities. However, the underlying assumptions with respect to the poverty effect of the program are similar to the ones we make.

accuracy and section 3.6 reports results from poverty simulations. Final remarks are offered in section 3.7.

3.2 Targeting in developing countries

Targeting poverty-alleviation programs can be based on various forms of means tests. What is required is only some sort of information on households' economic means. In developed countries, where information on incomes is usually available (e.g. through pay stubs or tax records), *verified means tests* (VMTs) can be conducted, that is, the selection of beneficiaries proceeds on the basis of these data. In developing countries, however, reliable and complete data on income is often not available. Moreover, as will be discussed in the next section, income is often not an appropriate indicator of economic welfare in developing-country settings. *Simple means tests* (SMTs), which rely on self-reports of living standards, are an alternative in the absence of any data. One example is Brazil's *Bolsa Família*, one of the largest conditional cash transfer programs with about eleven million beneficiaries, which uses targeting based on applicants' own reports administered at the level of municipalities [Veras Soares et al., 2010].[10] The drawback is that applicants will usually have a strong incentive to misreport their income in order to gain beneficiary status. Of course, they may also simply misjudge their income. Administrators may thus want to verify whether observed living standards are consistent with reported living standards but this will require government agents to pay visits to applicants. This, in turn, will usually drive up administrative costs and may lead to patronage. *Bolsa Família*'s simple means test, for instance, has been criticized on the grounds that its decentralized administration and unverified selection criterion results in patronage and leakage [Handa and Davis, 2006].

PMTs are a more sophisticated means testing device that relies on objective data rather than self-reports. They can be applied when reliable information on incomes is not available and collecting this information for all applicants is either impossible or prohibitively expensive. In a first step, data from a detailed household survey are used in order to extract easy-to-observe indicators (proxies) correlated with a pre-defined indicator of economic welfare and to fit an econometric model that captures the relationship between economic welfare and the set of proxies. In a second step, all applicants are required to fill in shortened questionnaires, administered separately, in order to gather data on proxies for all potential beneficiaries. These data can then be used to predict economic welfare for all applicants. Based on these predictions (often referred to as *PMT scores*), one decides whether applicants are eligible for transfers.

How do PMTs compare to universalistic transfers and alternative targeting strategies? Some targeting will usually result in a larger poverty effect as more resources will be available for those in need. See Akerlof [1978] for this argument and Grosh and Baker [1995] for evidence from simulations. In comparison with universalistic programs for which no data are required, PMTs are frequently found to bring about greater poverty effects [Grosh and Baker, 1995]. This is also very much in line with what we find in our poverty simulations. More importantly, PMTs are often found to perform better in terms of targeting outcomes than alternative methods that are sometimes easier to administer (e.g. geographic targeting) [Coady et al., 2004]. Finally, an

[10] According to Veras Soares et al. [2010], the application form for Brazil's program has additional questions on consumption which are used to cross-check the income figure. A deviation of 20 percent will trigger further actions.

additional advantage of PMTs as a targeting device is their cost-effectiveness. Administering detailed household surveys to all applicants will in most cases not be an option, especially when what is required are high-quality data on household expenditure rather than data on incomes.

On the other hand, PMTs have come to be criticized for several reasons: first, PMTs are still subject to manipulation if administrative capacity is limited or proxies and/or data on proxies are difficult to gather and verify. For instance, Camacho and Conover [2011] demonstrate that Colombia's SISBEN I-program,[11] a PMT-based, targeted cash transfer-scheme, has been subject to systematic manipulation, most likely by enumerators or local politicians. In addition, PMT-based targeting may induce adverse behavioral responses if details of its workings are known to applicants. Second, it is not always clear whether the indicator of welfare used in order to derive individual scores is appropriate. This is a problem with all means tests; economic means may simply not be seen as the right indicator of well-being. For instance, while Alatas et al. [2012] find that expenditure-based PMTs perform better in terms of targeting outcomes, they also report that community-based methods result in fewer complaints. This may indicate that locals' notion of poverty differs [for a similar argument see Ravallion, 2008].

There are at least two widely-used targeting strategies that try to circumvent the measurement problem. The first are community-based targeting schemes that explicitly aim to incorporate local knowledge by delegating the selection of beneficiaries to locals. While tapping into local knowledge may have advantages, this strategy is often criticized on the ground that it is prone to local capture [Bardhan and Mookherjee, 2005, Conning and Kevane, 2002]. Alternatively, self-targeted programs such as public works programs circumvent the problem by adding a requirement that results in disutility.[12] Presumably, they can be designed in such a way that only the truly needy will apply. See Besley and Coate [1992b] for the incentive argument and Drèze and Sen [1989] for evidence of the impact of such self-targeting in the case of India. However, Alatas et al. [2013] argue that such ordeal mechanisms may actually have ambiguous effects on targeting.

Finally, Kidd and Wylde [2011] argue that PMTs suffer from large margins of error when only a small share of the total population is targeted and assumed poor. In their analysis of data from Bangladesh, Rwanda, Sri Lanka and Indonesia they find that when ten percent of the population are poor and an equal share is selected for transfers based on PMTs, about 60–70 percent of beneficiaries are 'false positives.' This fraction decreases to about 30–40 percent when the poverty headcount is 40 percent and an equal share is eligible. They also show that regression-based PMTs fail to predict per adult equivalent expenditure at the tails of the actual distributions. Instead, PMT scores tend to overestimate expenditure of the poorest and underestimate expenditure of the richest.

While our results confirm this broad pattern, our findings with regard to accuracy are somewhat more optimistic. For instance, we find that about 42–50 percent of all beneficiaries are erroneously included in the first scenario (ten percent poor and ten percent targeted). We also find that accuracy improves substantially when both the poverty headcount and the proportion

[11] See Canstañeda [2005] for details on the SISBEN I-program.

[12] An example is India's Mahatma Gandhi National Rural Employment Guarantee Scheme which offers up to one hundred days of unskilled manual labor per year on public works projects for anybody willing to work at the stipulated minimum wage rate. See Dutta et al. [2012] for a recent review of this particular scheme.

targeted are higher. For instance, if the poverty headcount is 50 percent and 50 percent are targeted based on PMTs, less than 30 percent included are included erroneously.[13]

Our analysis is also broader in scope as we consider scenarios in which the proportion poor differs from the proportion targeted. In particular, we consider scenarios in which one deliberately allows for some leakage to the non-poor and investigate who benefits from leakage in section 3.5.3. Our findings suggest that if, for instance, at a headcount of 50 percent, 80 percent are targeted, we actually cover more than 90 percent of the poor. Hence, PMTs seem to be an attractive targeting device if the aim is not to identify the poorest of the poor, but to weed out the wealthiest households.

Note that this has the additional advantage of securing the necessary public support for the program. There is a long-standing literature that suggests that targeting only the poorest often results in programs that do not dispose of the resources to make a serious dent in poverty [e.g. Atkinson, 1995, Gelbach and Pritchett, 2000, 2002, Moene and Wallerstein, 2001]. Ravallion's (2007) findings can be interpreted to provide some empirical evidence for these models.

We certainly agree with Kidd and Wylde's observation that PMTs fail to accurately predict welfare of the poorest and the richest. Predictions from *ordinary least squares* (OLS)-based regressions will always have a smaller variance than the outcome variable. However, this is irrelevant if the poverty rate is known and information on all potential beneficiaries is available. Note that the actual values of scores are then no longer relevant; only the *ordering* among households induced by scores will matter.

3.3 Datasets and welfare indicator

The datasets used in this study are the *Encuesta de Hogares* (EH) 2008 and 2011, comparable household expenditure surveys that are both representative at the national level. The data were obtained by two-stage sampling with inflation factors for households. The questionnaire includes modules on households' demographic make-up, socioeconomics (education, asset ownership, dwelling characteristics, service use), as well as detailed information on household expenditure.

The first step in the design of a PMT is the choice of an indicator of well-being. Sen [1995], among others, argues that one should use an indicator that is directly related to the program in question. Hence, if the program is mainly concerned with monetary poverty, the lack of adequate command over goods and services, households' economic means are an obvious choice. This section explains the construction of our indicator of 'economic welfare' that we will use throughout the remainder of this paper.

We base our indicator of economic welfare on household expenditure data from the *Encuesta de Hogares*. We include expenditure on food, education, rents, services, service flows from durable goods, and other non-food items. Some items are excluded in order to avoid double counting. We also exclude expenditure associated with catastrophic events. Imputations are used sparingly since they would be largely self-defeating. We also make appropriate adjustments for price differences between localities and years as well as differences in households' demographic make-up. Details are provided in appendix 3.A. Despite these adjustments, we will refer to the resulting indicator simply as *real expenditure* in what follows.

[13] Note that if these proportions are always equal, the inclusion error always equals the exclusion error.

Note that expenditure is more appropriate than incomes in our setting [Deaton and Zaidi, 2002]. Incomes tend to have an important seasonal component, especially for households involved in agricultural production. Reported incomes may also be problematic when the informal economy is a major source of incomes (as is the case in Bolivia) and when a large portion of the food produced is consumed by households. On the other hand, even poor households are often found to have the means to smooth-out consumption expenditure to some degree.

3.4 Proxy Means Tests

3.4.1 Regression-based PMTs and out-of-sample prediction

We employ OLS-regressions of log real expenditure against proxies in order to evaluate the usefulness of PMTs in the present data. In order to mimic a real-life PMT exercise and to avoid 'overfitting' the data, all investigations are carried out using out-of-sample predictions. In particular, we test the predictive power of models trained using (i) the 2008 data and (ii) a subsample of the 2011 data as *calibration samples*. The 2008 survey provides data on 3,937 households that have information on all relevant variables and household expenditure surveys of this size are frequent in Bolivia. We will therefore randomly set aside a subsample of comparable size—3,932 households—from the 2011 data, from which we retain 8,842 household records in total. The remaining 4,910 2011 records are then employed to analyze the performance of PMTs, that is, they serve as the *validation sample*.[14,15]

In sampling households for calibration, one has to take into account the original design of the survey. Households were sampled in a two-stage procedure. First, *primary sampling units* (PSU) were randomly selected.[16] In a second step, households within PSUs were sampled. We thus also randomly sample PSUs in order to obtain the 2011 calibration sample. It is also evident that urban households are over-sampled both in the 2008 and the 2011 survey rounds. While about half of Bolivia's population lives in rural areas, the corresponding proportions are 59 and 67 percent in the data, respectively. We therefore ensure that this proportion carries over to the 2011 calibration sample. This results in a calibration sample with 2,661 (68 percent) urban households.

Finally, an important question when dealing with complex survey data is whether or not weights should be used. Since we are ultimately interested in poverty and targeting rates at the individual-level, we use individual inflation factors in what follows. These are defined as household inflation factors, available in the datasets, multiplied with the number of household members.

3.4.2 Proxy sets

We investigate the targeting performance of four models that differ in approach and complexity. To do so, we first group proxies extracted from the data into five mutually exclusive sets:

[14] While this is not exactly what would be done in practice where households in the calibration sample may also be beneficiaries, we imagine that the number of households in the calibration sample is small in practice compared to the households that potentially qualify.

[15] We also tested for differences in means of key variables in our analysis based on t-tests across 2011 samples and found mostly no statistically significant differences.

[16] On average, there are 11.2 households per PSU in our data with a median of 12 households. The minimum is five households and the maximum 14.

geographics, demographics, dwelling characteristics, electricity bill information, and assets. The most basic model (called M1) includes only geographic and demographic proxies on the right hand-side. Geography is covered by coding binary variables for all of Bolivia's nine departments. We then interact these variables with a binary variable indicating households residing in rural areas. In total, this adds 17 parameters to the model. Demography is covered by the number of household members by sex and age: the number of females and males below the age of five, between five and 15, between 16 and 64, and 65 and older, respectively. This results in eight regressors. The total number of parameters of this first benchmark models is thus 26 (including a constant).

The second model (M2) includes all of the above proxies and adds variables capturing dwelling characteristics and service use. We code dummy variables that indicate whether superior materials were used in the construction of walls, roof, and floors of the dwelling.[17] We also include the number of rooms, dummies indicating whether the dwelling's kitchen is located in a separate room, a separate water connection, and two exclusive dummies indicating in-house or shared toilet. Two further dummies indicate access to waste removal-services and electricity. We believe that these variables are easy to verify during a personal visit to the household. Information on service use may in addition be available from official records.

The third model (M3) adds information on (the log of) spending on electricity services. If the household is not served or information is missing, we impute the median value in the entire sample and mark these households by including dummies for missing information and zeros. The final model (M4) does not rely on expenditure on electricity services but includes a set of dummy variables indicating ownership of the following durables: refrigerator, personal computer, TV set, microwave, washing machine, air conditioner, heater, car, landline phone, and cell phone.

Our choice of proxies is guided by concerns over verifiability, incentive-compatibility and legality. The different sets are progressively more difficult to verify. For instance, many existing PMTs employ information on the ownership of durable goods—similar to our set M4. These tend to be highly correlated with economic welfare yet are easily concealed from government agents. Educational attainment may be a good proxy in the sense that it is highly correlated with earnings and thus with economic welfare. However, it seems questionable whether information on educational attainment could be verified. Other household characteristics that are good indicators of poverty may induce perverse behavior when employed in a targeting scheme. The number of children out of school or undernourished, for instance, falls into this category [e.g. Morris et al., 2004]. Similar problems potentially arise with proxies used in our PMTs M1 through M3 but, as we would claim, to a lesser degree.[18]

3.4.3 Results

Table 3.1 reports results from the above regression models based on the 2008 and 2011 calibration samples, respectively. However, since individual coefficients are hardly decisive in this exercise,

[17] Improved materials are cement, bricks, or concrete for walls; bricks or armored concrete for roofs; and anything except dirt and wood planks for floors.

[18] In an extreme case, for instance, households may alter their demographic composition in response to the targeting mechanism. The most contentious issue may be the use of expenditure on electricity. This information should be employed such that incentive-problems are minimized. Ideally, one would use data directly obtained from providers and consider expenditure over the course of a year in order to avoid variation due to seasonality. The survey underlying our data asks for the amount usually spent per month but all interviews were conducted within one month.

Table 3.1: Regression results for log per adult expenditure based on 2011 and 2008 calibration samples (3,932 and 3,937 observations, respectively).

	M1		M2		M3		M4	
	2011	2008	2011	2008	2011	2008	2011	2008
Number of parameters	26	26	36	36	39	39	46	46
R^2 (within-sample)	0.36	0.44	0.52	0.57	0.56	0.60	0.62	0.65
ρ^2 (out-of-sample)	0.38	0.38	0.52	0.51	0.57	0.56	0.62	0.61

Based on weighted OLS regressions using individual inflation factors. Authors' own calculations based on data from the *Encuesta de Hogares* 2008 and 2011.

we report only indicators of the regression fit, namely the number of parameters, R^2-statistics from regressions based on calibration samples, and, for comparison, the squared Pearson correlation coefficient between predicted values (or *PMT scores*, in the terminology of this method) and actual values in the validation sample, i.e. the 2011 households not included in the calibration sample. All quantities are estimated using individual inflation factors as weights.

Several observations can be made from table 3.1: as can be seen from the R^2-statistics, the regression fit with the 2008 calibration sample is much better than the fit obtained with the 2011 calibration sample for M1, the most parsimonious model. The R^2 is 0.44 for the model fitted to 2008 data compared to 0.36 for the model fitted to 2011 data. The difference is less pronounced for more sophisticated models but still noticeable. This implies that households' geographic and demographic characteristics were better predictors in 2008 as compared to 2011, a somewhat surprising finding.

As one would expect, the squared correlation coefficient, denoted ρ^2 in table 3.1, is lower than the R^2 when the 2008 data are used to obtain PMT scores.[19] This is particularly true for PMT scores obtained from the models calibrated on 2008 data. There are virtually no differences in the case of scores obtained from models calibrated on 2011 data.

While initially the fit of the models is improved substantially by adding more proxies, the changes in the R^2-statistic quickly level off. In particular, differences between M2 and M3, which adds information on the electricity use and payments for the service, are small. The highest correlation out-of-sample is obtained with M4, the model including information on asset ownership. In this case, the model fitted to the 2011 calibration data accounts for more than three-fifths of the variation in real expenditure in the validation sample.

What accounts for the remaining variation in the data? Measurement error in the dependent variable could be a culprit. This would point to a fundamental weakness of PMTs: they are only as good as the underlying data and improving the quality of the process through which the data are obtained (e.g. by asking household to keep diaries in order to obtain more reliable records of consumption expenditure) would seem to be the only solution to this problem.

Another explanation are short-lived fluctuations in consumption expenditure. It may be the case that many of the proxies we consider will not adjust to economic shocks such as job loss, accidents, and illnesses as quickly as will consumption expenditure. A household will not immediately sell off all its assets once it learns that it has to cover high hospitalization costs. While one could likely find proxies for such shocks in the dataset at hand, it seems clear that

[19] Remember that if the model includes an intercept, the R^2 equals the correlation coefficient between PMT scores and actual responses.

these would be hard to verify. It therefore seems questionable whether the explanatory power of these models could be improved substantially by finding and adding more proxies if proxies are to be verifiable.

Overall, the model fit we obtain based on the most sophisticated model with 46 parameters (37 variables, eight interaction terms, and a constant) still compares favorably to other results in the literature. For instance, a recent study that investigated models created from 340 candidate variables in the Indonesian Family Life Survey finds that even the best models based on 40 variables do not attain an R^2 higher than 60 percent [Bah, 2013].

In this present section we introduced our empirical models and evaluated model fits. We now turn to assessing the accuracy of the predictions that can be obtained. R^2-statistics are informative in terms of model fits to some degree, but leave much to be asked when it comes to evaluating the targeting accuracy of PMTs.

3.5 Assessing accuracy

3.5.1 ROC-analysis

We now turn to assessing how well PMTs are suited to distinguish between the poor and the non-poor. What we need are indicators that convey information about the proportion that is correctly identified. A first approach to measuring the usefulness of regression results in identifying the poor is to introduce a poverty line in the space of economic welfare and calculate the *total error rate* (TER), the proportion of households correctly classified based on PMT scores. For instance, if the poverty headcount is 20 percent, we would identify the bottom quintile from the distribution of scores and define them as poor. The calculation of the TER is then straight-forward: we would simply add the number of non-poor classified as poor and the number of poor classified as non-poor and divide by the total population.

The TER can be a misleading measure of accuracy as it depends on the proportion of poor, however. For instance, if only a small proportion of households is poor, identifying nobody as beneficiary would still result in a TER close to zero. ROC-analysis [Johannsen, 2008, Landau et al., 2012, Thompson and Zucchini, 1989, Wodon, 1997] was specifically designed to overcome this problem.

Table 3.2: Classification matrix: taxonomy of targeting errors

	Beneficiary Status		
Poverty Status	Beneficiary	Non-beneficiary	Total
Poor	n_1	n_2	N^p
Non-poor	n_3	n_4	N^{np}
Total	N^b	N^{nb}	N

Consider table 3.2, a *confusion matrix*: each cell contains information about the number of households in a specific group. For instance, n_1 is the number of poor that are classified as beneficiaries and n_3 is the number of non-poor classified as beneficiaries. Granted we have data as in table 3.2, we can define the *true positive rate* (TPR), the proportion of poor (or *positives*) identified as beneficiaries, and the *false positive rate* (FPR), the proportion of non-poor identified as beneficiaries as

$$\text{TPR} = n_1/N^p \quad \text{and} \quad \text{FPR} = n_3/N^{np},$$

respectively. The FPR is also often referred to as the *leakage rate*, whereas n_2/N^P, the share of the poor that are not beneficiaries, is also often referred to as the *undercoverage* rate [Grosh and Baker, 1995].

TPRs and FPRs depend on the proportion admitted to the program which may vary from zero to one. Applied to targeting of poverty-alleviation programs, ROC-curves are plots of the TPR against the FPR for variations in the cut-off used to identify the poor in the space of PMT scores.

Since the number of poor and non-poor is fixed, both the TPR and FPR will increase as more people are admitted into the program. However, if the classifier is better than random classification, that is, if PMT scores convey some information about who is poor and who is not, the TPR will increase at a faster rate initially giving rise to a concave ROC-curve. With perfect targeting, i.e. if there was a perfect linear relationship between predicted welfare and actual welfare, only the TPR would increase initially until all the poor were covered (the TPR is unity). The FPR, on the other hand, would be zero over this range. Only when all poor are beneficiaries will the FPR increase until, finally, all household are covered (both TPR and FPR are unity). A random classifier, on the other hand, would see both TPR and FPR increase at the same rate. Hence, the ratio of the two would be unity in expectation. The farther the ROC-curve 'bends' towards the top left corner of the plot, the more accurate the targeting.

A summary measure of the information provided by ROC-curves is the the area under curve (AUC). Note that the AUC is bounded between zero and unity, where the later would result from a perfect classifier; the larger the area, the better the model. A random classifier would, however, still generate an area that is one-half in expectation. AUCs focus on the entire curve and are thus subject to the criticism that they also focus on areas that are of little interest in practice [e.g. Thompson and Zucchini, 1989]. For instance, in our application, very low TPRs would often seem unacceptable. Alternatives include focusing on fixed values of the TPR (or the FPR) or on partial areas [see Thompson and Zucchini, 1989] as well as on specific coverage rates, the proportion of households admitted to the program. Below, we will focus primarily on the latter approach.

While a classifier is strictly superior whenever it produces greater TPRs for all FPRs, some of the curves may intersect. In other words, one model may perform better at low levels of the FPR, but worse at higher levels. At the same time, both models may generate ROC-curves that yield (almost) identical (partial) AUCs. Wodon [1997] shows that one can still find a criterion that allows selection of one model as the superior model as long as society's (or a policy-maker's) preferences can be represented by a strictly quasi-concave utility function that takes the TPR and the FPR as arguments. However, we find that PMTs that involve more data are superior with only minor exceptions.

In what follows, we therefore compare PMTs based on ROC-curves, AUCs, and at several values of the coverage rate. However, we will do so for different poverty headcounts of ten, 25, and 50 percent. Poverty is hardly a discrete condition and PMT-based targeting may be an option for programs designed for those in extreme poverty as well as those in moderate poverty or even those above the official poverty line. In our case, we are interested in learning about the usefulness of PMTs in different situations—including different levels of poverty.

3.5.2 Results

The ROC-curves generated by our PMTs are plotted in figure 3.1. As described above, ROC-curves are generated by varying the proportion of beneficiaries based on predicted real expenditure and comparing classification matrices. We define as poor successively ten, 25, and 50 percent of the population, where the latter two figures correspond roughly to the incidence of extreme and moderate poverty in Bolivia at the time. Recall that we employ individual-level frequency weights, defined as household expansion factors multiplied with the number of household members, in order to obtain results representative of the entire Bolivian population in 2011.

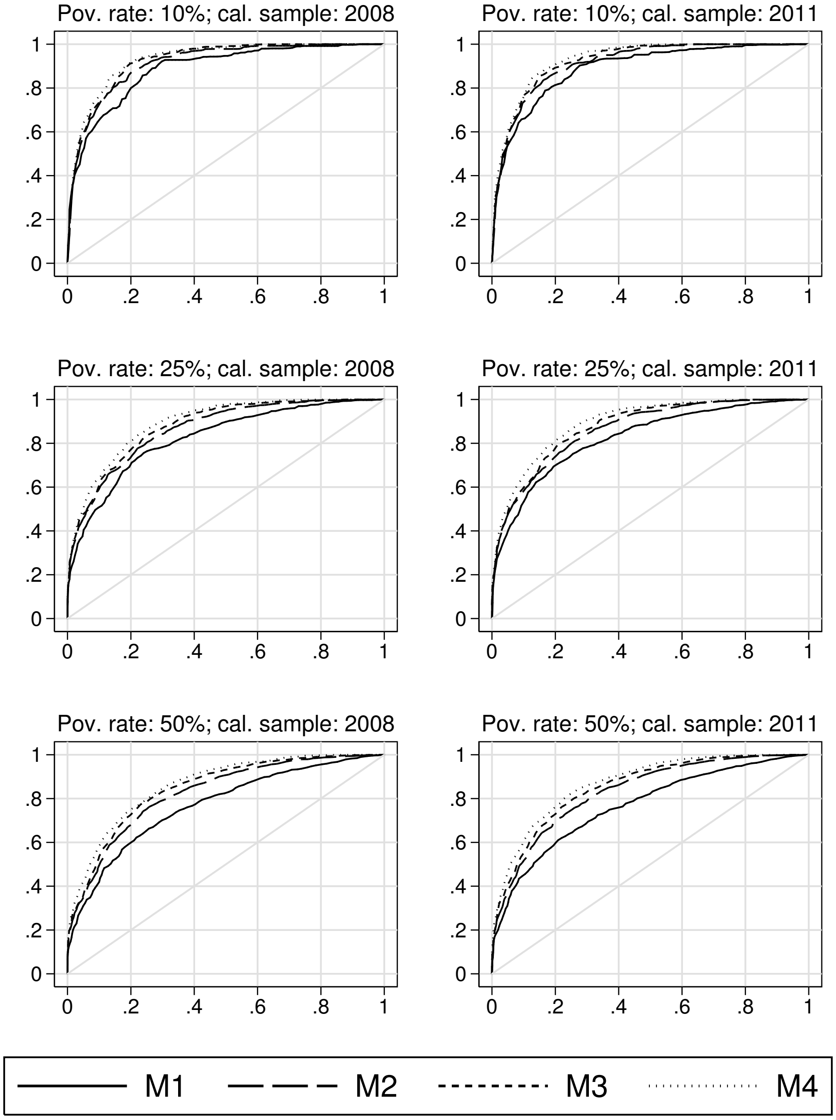

Figure 3.1: ROC-curves for PMTs M1–M4 based on 2008 and 2011 calibration sample.

Part of the information conveyed in figure 3.1 is tabulated in table 3.3 where we report TPRs at different inclusion rates for each PMT as well as AUCs for poverty rates of ten (panel A), 25 (B), and 50 percent (C), respectively. For instance, the first entry in panel A suggests that if ten percent are poor and ten percent are targeted, 52.6 percent of those receiving transfers will actually be poor. Note that we do not have to report in addition FPRs as they are functions of TPRs, the poverty rate, and the inclusion rate.[20] Therefore, from a targeting accuracy-perspective and conditional on inclusion and poverty rates, PMTs with higher TPRs are strictly preferred.

Reading the table from left to right we can infer the effect of increasing the complexity of PMTs for a given poverty headcount and beneficiary proportion. For instance, 57.1 percent of those targeted are poor under M2 and close to 60 percent under M3 and M4. However, the differences are small, probably smaller that would be suggested by differences in R^2-statistics reported in section 3.4.3. Given the caveats of using more complex models such as higher costs in generating the data and, possibly, higher susceptibility to manipulation and deceit (e.g. in the case of assets), it seems likely that policy-makers would want to opt for parsimonious PMTs.

If we compare the left half with the right half of the table, we can see whether using up-to-date data improves accuracy. While models trained on the 2011 data perform slightly better, differences are small, usually between zero and two percentage points. We conclude that all of the PMTs we stipulate are fairly robust over short time periods.

We also find that if the proportion receiving benefits is equal to the proportion in poverty, PMTs produce the highest TPRs when half the population is poor. Under M1, for instance, the TPR is only 54.1 percent when ten percent are poor and the same number is marked for inclusion. However, when half the population is poor and half receive transfers, almost 70 percent of beneficiaries will be poor. We conclude that PMTs are less accurate in identifying the poor when only a small percentage of the population is poor and an equally small share is targeted. The appropriateness of PMTs when the goal is to reach the poorest with a very limited program is thus questionable.

This reaffirms findings from the literature, both from simulations and real world-experience. For instance, in their simulations based on data from Bangladesh, Rwanda, Sri Lanka, and Indonesia, Kidd and Wylde [2011] find that TPRs between 29 and 43 percent when ten percent are poor and ten percent are targeted. Our results suggest that targeting at this level would be more accurate in Bolivia. For Mexico's *Opportunidades* and Brazil's *Bolsa Família* programs which target around 25 and six percent of the population, respectively, Veras Soares et al. [2010] report TPRs of 20 and 41 percent, respectively.[21] Even though our results look somewhat better, it is not clear whether one would want to accept that only about half of the poorest ten percent are covered by the program.

PMTs perform much better when a larger portion of the population is considered poor and the program channels transfers to an equally larger share of beneficiaries. When half of the

[20] Given the headcount, $P_0 = N^p/N$, the beneficiary share, $B_0 = N^b/N$, and the TPR, $\text{TPR} = n_1/N^p$, the FPR can be written as

$$\text{FPR} = \frac{B_0 - \text{TPR} \cdot P_0}{1 - P_0},$$

which depends only on known quantities. Moreover, given P_0 and B_0, the FPR is a strictly decreasing function of the TPR.

[21] Brazil does not have an official poverty line but the cut-off point for program eligibility in 2004 was R$100. Veras Soares et al. [2010] use this cut-off to arrive at their results. Such a poverty line would then result in a poverty rate of roughly seven percent.

Table 3.3: TPRs at varying levels of inclusion rates and AUCs.

	2008				2011			
Inclusion rate (%)	M1	M2	M3	M4	M1	M2	M3	M4
Panel A. Poverty headcount of 10 percent.								
10	0.526	0.571	0.591	0.598	0.541	0.548	0.569	0.577
25	0.775	0.832	0.878	0.893	0.805	0.855	0.880	0.893
50	0.943	0.978	0.988	0.982	0.952	0.973	0.988	0.993
60	0.962	0.988	0.996	0.996	0.964	0.993	0.994	0.999
70	0.981	0.993	0.999	1.000	0.979	1.000	1.000	1.000
80	0.992	0.999	1.000	1.000	0.990	1.000	1.000	1.000
90	0.997	1.000	1.000	1.000	0.998	1.000	1.000	1.000
AUC	0.870	0.899	0.910	0.916	0.874	0.898	0.908	0.918
Panel B. Poverty headcount of 25 percent.								
10	0.313	0.340	0.347	0.347	0.316	0.339	0.345	0.347
25	0.580	0.635	0.652	0.668	0.605	0.630	0.634	0.674
50	0.840	0.893	0.917	0.933	0.839	0.890	0.917	0.936
60	0.898	0.945	0.961	0.975	0.897	0.943	0.958	0.966
70	0.935	0.974	0.982	0.987	0.936	0.974	0.980	0.985
80	0.970	0.989	0.992	0.998	0.963	0.993	0.992	0.998
90	0.993	0.997	0.999	1.000	0.989	0.999	1.000	1.000
AUC	0.805	0.848	0.862	0.878	0.805	0.847	0.860	0.878
Panel C. Poverty headcount of 50 percent.								
10	0.185	0.194	0.195	0.195	0.186	0.190	0.192	0.195
25	0.403	0.424	0.430	0.442	0.418	0.432	0.439	0.452
50	0.701	0.751	0.771	0.773	0.695	0.748	0.770	0.782
60	0.786	0.836	0.857	0.871	0.778	0.843	0.857	0.873
70	0.856	0.906	0.922	0.935	0.853	0.912	0.926	0.934
80	0.919	0.956	0.969	0.970	0.915	0.960	0.971	0.982
90	0.964	0.988	0.992	0.998	0.963	0.991	0.995	0.997
AUC	0.744	0.803	0.823	0.840	0.742	0.809	0.828	0.848

Authors' own calculations based on data from the *Encuesta de Hogares* 2008 and 2011.

population is poor and half is targeted, TPRs range from 69 to 78 percent. Between 80 and 95 percent of the poor would be included if 80 percent of the population would be targeted. Thus, if leakage is not a major concern, broad targeting can be achieved with reasonably high TPRs with these models.

3.5.3 Assessing leakage

As we have seen, one has to allow for considerable leakage when a high true positive rate (say, 90 percent) is required but the percentage of 'deserving poor' is small. But how much of a problem is leakage? Clearly, one would be less concerned about benefits reaching the non-poor but vulnerable—households close to the poverty line that have a high likelihood of becoming poor in the future. Also, the actual real expenditure of the non-poor close to the poverty line will not differ much from those directly below it. Poverty lines are helpful in that they allow for the calculation of poverty measures and poverty comparisons but, after all, poverty is hardly a discrete condition and one has to allow for the fact that poverty lines are essentially arbitrary.

In our analysis, where will we find those erroneously identified as poor? Much can already be inferred from table 3.3 where we find that true positive rates increase when the poverty rate is increased together with the percentage targeted. This implies that leakage tends to benefit the 'poor among the non-poor.' This is further illustrated in figure 3.2, where we investigate the distribution of benefits when 25 and 50 are targeted. Again, this is done for our out-of-sample scores in order to mimic the real world situation where models are fitted with one dataset and the PMT is then applied to fresh data.

Figure 3.2 is generated by first defining dummy variables identifying beneficiaries for all PMTs, i.e. the poorest 25 and 50 percent based on PMT scores. These indicators are then regressed against percentiles of the log real expenditure distribution. We use *kernel-weighted local polynomial regression*, a non-parametric method that allows us to recover the non-linear relationship between the beneficiary dummy and real expenditure percentiles.[22] The resulting regression lines can be interpreted as the probability of receiving the transfer and, by the law of large numbers, as the proportion of beneficiaries *conditional on real expenditure*. The black lines depict the result from using the four PMTs with a targeting cut-off of 25 percent and the blue lines the results for a cut-off of 50 percent.

The steeper the slope of the regression line, the greater the (negative) effect of real expenditure on the probability of receiving the transfer and the better the associated PMT performs in terms of accuracy. Note that random classifiers would result in horizontal lines; the probability of receiving the transfer would be unrelated to real expenditure. Regression lines pertaining to a situation in which 50 percent are admitted to the program (blue lines) are located to the right of regression lines pertaining to situations in which 25 percent (black lines) of the population are admitted as the probability of receiving the transfer is greater for a given level of real expenditure.

We know from table 3.3 that when 50 percent of the population is targeted, more than 90 percent of the individuals in the bottom decile of real expenditure will be covered. In line with these high TPRs reported in the table, we find that even with the most parsimonious model, the probability of receiving the transfer is greater than 90 percent everywhere below a poverty line that renders the bottom ten percent poor. At the same time, leakage will be considerable as more than four out of five households receiving the transfer will not be poor. The question remains, however, where in the distribution of real expenditures these 'false positives' will be located. We find that the probability of receiving the transfer drops to about 50-55 percent at the median of the distribution and to about ten percent by the time we get to the richest ten percent with all PMTs except M1 for which it is still about 20 percent. With 25 percent of the population admitted to the program, we find that the probability of receiving the transfer is only between 70 and 80 percent for an individual at the tenth percentile. The probability drops to only 20 percent for an individual at the median of the distribution and to less than ten percent for an individual at the 90th percentile the top decile.

Finally, note that we see again some differences in targeting performance across models. In particular, the most parsimonious model, M1, performs worse as the probability of receiving the transfer decreases only slowly. On the other hand, there seem to be few differences between more sophisticated PMTs. This is in line with our findings above that changes in the regression fit, as measured by the R^2-statistic, quickly level off as we go to moderately complex to very complex models.

[22] We use pre-defined rule-of-thumb bandwidths for all regressions.

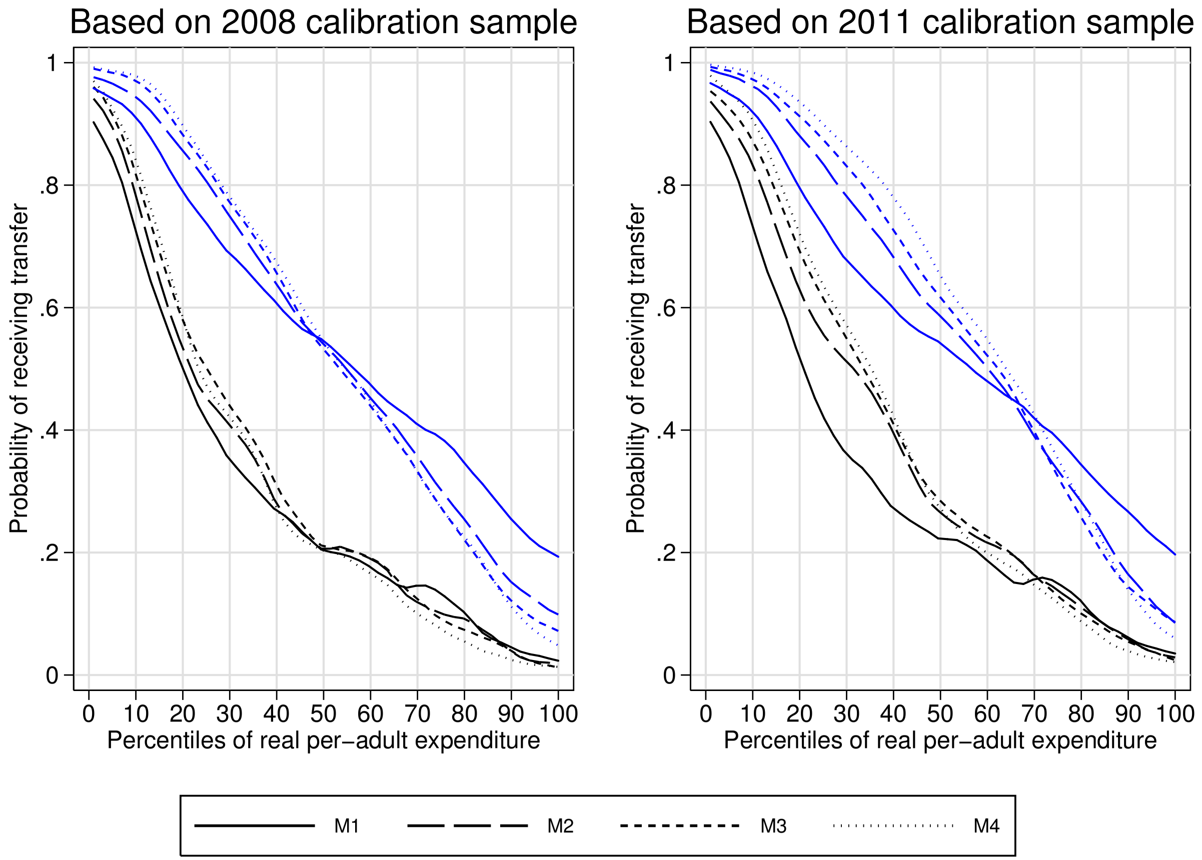

FIGURE 3.2: Local polynomial regression of beneficiary status against percentile of per-adult real expenditure. Based on PMTs trained on the 2008 (left panel) and 2011 calibration samples (right panel) and evaluated with the validation sample. The solid, long-dashed, short-dashed, and dotted regression lines indicate models M1–M4; the black and blue regression lines correspond to targeting the poorest 25 and 50 percent of the population, respectively, when an equal share is poor.

Our findings in this section can be summarized as follows: first, returns to increasing the number of proxies are rapidly decreasing. Second, the potential of PMTs in Bolivia seems to be higher than in other developing countries. Third, our PMTs still result in low TPRs (between 52 and 60 perent) when used to target only the very poor. However, we also find that leakage is confined mostly to individuals located just above the poverty line. Finally, TPRs are much higher (about 69 to 78 percent) when PMTs are used to target half the population and half the population is poor. Again, increasing the inclusion rate in this situation results in a steep increase in TPRs.

3.6 Poverty simulations

3.6.1 Set-up

This final section offers simple simulations that illustrate the potential poverty effects of PMT-based monetary transfers and compare these to the case of ideal targeting, i.e. targeting based on the actual ranking generated by our welfare indicator. Ravallion [2007] shows that programs can be well-targeted and still have little effect on poverty. It is therefore paramount to also simulate poverty effects of transfers based on the PMT instruments discussed above.

Our set-up is simple: we assume a fixed budget b for a transfer scheme is available. We consider a total budget that amounts to one percent of total household expenditure or 97bs. million per month. We use household inflation factors in order to arrive at total expenditure and the total number of households in Bolivia. Since the exercise is again conducted only for the validation sample, we also calculate the sum of households weights in the calibration sample and multiply the budget with the ratio of the number of households represented in the validation sample to the all households (55 percent). We thus arrive at a budget $b^* = 0.55 \times b$ available for individuals in the population corresponding to the validation sample of 53.3bs. million. Each individual beneficiary receives a fixed transfer t that depends on the number of households targeted and the average household size of households targeted. Hence, if p percent of this population is targeted, each targeted individual receives a lump-sum transfer $t = b^*/pN$.[23]

As above, we consider three different poverty lines in the space of real expenditures that identify ten, 25 and 50 percent of the population as poor, respectively. As noted above, the later two headcounts roughly correspond to official estimates of the incidence of extreme and moderate poverty in Bolivia. However, we do not attach much significance to the actual value of these poverty lines. Rather, we imagine policy-makers that are concerned with the welfare of the bottom quintile and bottom half of the population, respectively.

We evaluate poverty effects by computing poverty measures of the Foster-Greer-Thorbecke (FGT)-class [Foster et al., 1984] for all combinations of transfers and proportions targeted. The simulations will therefore provide an illustration of the argument often made in favor of targeting, *viz.* that more narrowly targeted transfers will achieve higher poverty effects. We will be concerned mainly with how quickly the poverty effects level off as the inclusion rate of the program increases.

[23] Note that real expenditure per adult equivalent does not vary within households. Hence, all individuals in a household with real expenditure below the cut-off level will be targeted.

Several limitations of this exercise should be noted. First, we do not take into account any behavioral responses of households. The implicit assumption is that household attributes are perfectly observable at zero cost and that households do not change their attributes in order to gain beneficiary status. At the same time, we believe that these behavioral responses are unlikely to differ greatly between targeting mechanisms so that our comparative assessment is still likely to be valid. Second, we do not consider how the funds used for the transfer scheme are generated. In particular, households are not taxed. Third, we do not consider any politico-economic constraints policy-makers are likely to face [Gelbach and Pritchett, 2000, 2002, Moene and Wallerstein, 2001]. Political constraints would likely make the budget available an increasing, non-linear function of the proportion of the population targeted as more broadly targeted schemes will command more political support. Fourth, note that despite assuming a fixed budget for transfers, the analysis is not a full-fledged cost-benefit analysis in which poverty effects are compared to administrative costs. In order to interpret the exercise that way, one would have to assume that the costs of implementing schemes vary neither across different PMTs nor across the proportion of the population targeted. Finally, we fully add transfers to household total expenditure and then re-calculate per adult real expenditure. We thus assume implicitly that the entire transfer is spent. While illustrative in terms of poverty effects, our results are informative primarily in comparing different targeting approaches with different levels of targeting accuracy.

3.6.2 Results

The results of this simulation for an assumed pre-transfer headcount of ten, 25 and 50 percent (from left to right) are depicted in figure 3.3, where we plot relative changes in post-transfer poverty measures (headcount, poverty gap, and squared poverty gap from top to bottom) against the share of the population targeted. The dashed gray lines depict outcomes under perfect targeting, i.e. targeting based on actual real expenditure. In order to avoid over-populating graphs, we plot only results for PMTs trained on the 2011 calibration sample. As one would expect given our findings above, corresponding plots for PMTs trained on 2008 data are very similar. Part of the information conveyed in figure 3.3 is also tabulated in table 3.4, where we focus on the maximum poverty effect attained under ideal targeting as well as PMTs M1 and M4 and the inclusion rate at which this maximum occurs.[24]

To put these results into perspective, they can be compared to a universal equal transfer. If all individuals would receive an equal share of the budget, the transfer would amount to roughly 9bs. Such a modest transfer would reduce an initial poverty headcount of ten (25) [50] percent by 6.79 (4.41) [1.71] percent. The pre-transfer poverty gap is 0.03 (0.08) [0.19] and the squared gap is 0.01 (0.04) [0.10]. These would be reduced by 10.59 (5.80) [3.44] and 14.74 (8.08) [4.97] percent, respectively. Unsurprisingly, the effect on poverty measures is inversely related to the pre-transfer headcount for a fixed budget.

The first key result from this exercise is that all four PMTs perform very similar in terms of the poverty impacts they bring about. This becomes clear from the data in table 3.4: the difference in poverty impacts between the two PMTs is barely greater than one percentage point. Only when ten percent of the population are poor and the focus is on the distribution-sensitive FGT2 measure is the difference more pronounced. Remember that we saw some differences in

[24] Since our PMTs perform similarly, we report these figures only for ideal targeting and PMTs M1 and M4.

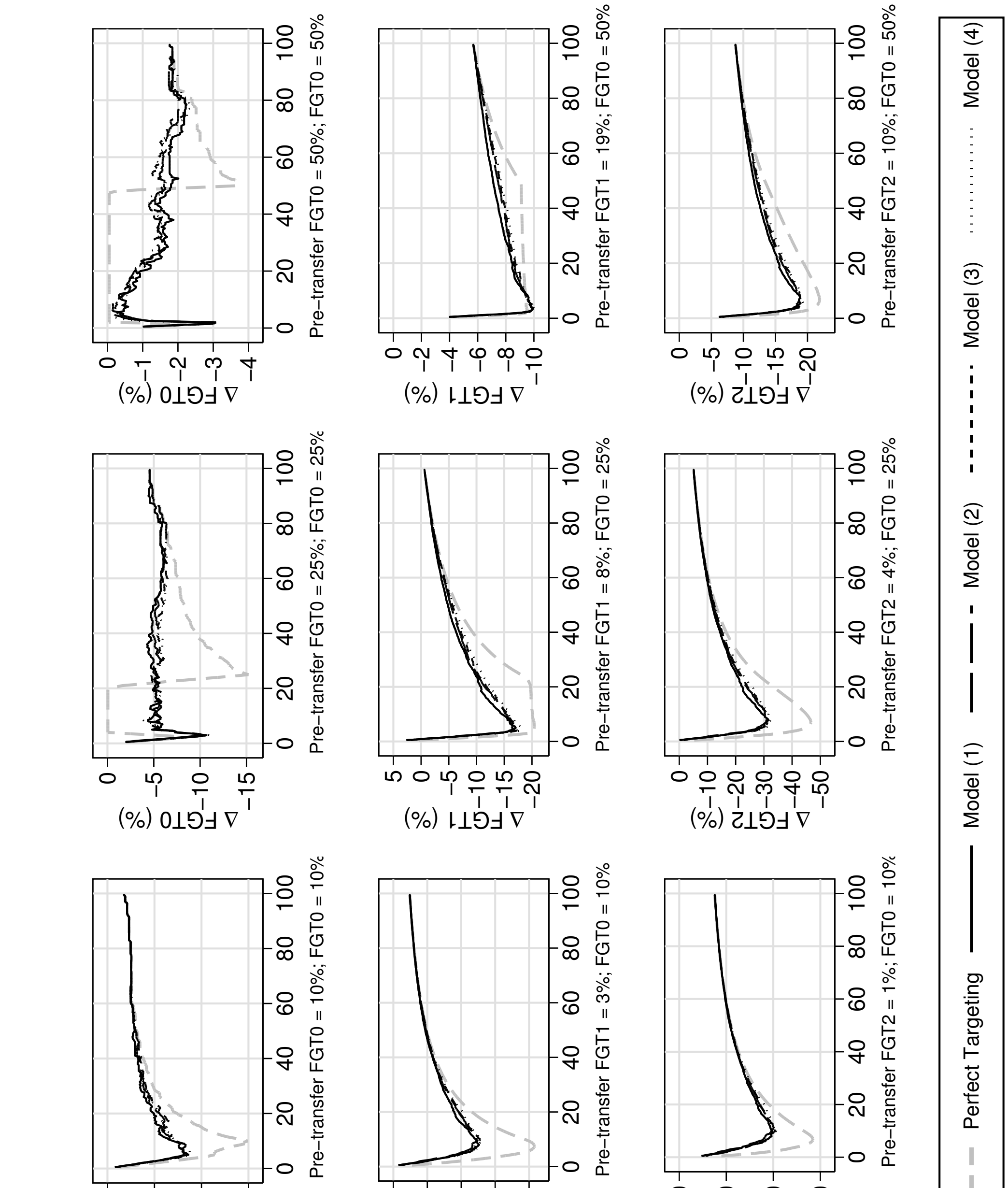

FIGURE 3.3: Simulated poverty effects based on the FGT-class of poverty measures against inclusion rate.

Table 3.4: Selected results from poverty simulations: ideal targeting, PMTs M1 and M4 based on 2011 calibration sample.

	FGT0 (headcount)			FGT1 (gap)			FGT2 (squared gap)		
	Ideal	M1	M4	Ideal	M1	M4	Ideal	M1	M4
Panel A. Poverty headcount of ten percent.									
Share targeted (%)	10.0	5.0	5.5	7.5	8.0	9.0	7.0	9.5	10.0
Poverty impact (%)	-59.7	-34.4	-35.6	-83.1	-50.5	-51.4	-92.1	-49.3	-53.3
Panel B. Poverty headcount of 25 percent.									
Share targeted (%)	25.0	3.0	3.0	4.0	5.0	4.0	7.0	8.0	8.0
Poverty impact (%)	-15.2	-9.9	-11.1	-20.4	-17.1	-18.0	-46.7	-31.6	-32.4
Panel C. Poverty headcount of 50 percent.									
Share targeted (%)	50.0	1.5	1.5	3.0	3.5	3.0	7.0	5.5	5.5
Poverty impact (%)	-3.7	-2.9	-3.0	-9.5	-9.9	-10.0	-21.9	-18.9	-19.3

Authors' own calcuations based on data from the *Encuesta de Hogares* 2008 and 2011.

terms of targeting performance between these two PMTs and that they differ considerably in the number of proxies they take into account. Despite these differences, very similar poverty impacts can be had.

The second central result relates to the ideal inclusion rate: unsurprisingly, going from universal schemes to targeting initially increases the poverty effect in almost all cases. The largest poverty effects are achieved with very narrowly targeted schemes. With the exception of the headcount in the case of ideal targeting, the maximum effect occurs at inclusion rates below ten percent. Given our previous finding that only very broadly targeted schemes led to high TPRs, it seems there is a trade-off here between including a large proportion of the poor and attaining a significant poverty effect.

The results for the poverty headcount are more complicated. It is worth noting first that the headcount ratio is less appropriate for poverty comparisons in this exercise as it violates Sen's (1976) *monotonicity axiom*: all else equal, a reduction in welfare of a person below the poverty line must decrease poverty. For pre-transfer headcounts of 25 and 50 percent, we find that the poverty headcount is initially reduced substantially in the case of perfect targeting but subsequently increases as targeting becomes less narrow. The initial transfer is large enough to lift the poorest above the poverty line. However, as the number of targeted individuals increases, the resources available *per individual* decrease rapidly and quickly become insufficient to lift the poorest households out of poverty. Despite the fact that the poorest households are better of, the headcount remains unaltered over a considerable range. It only drops sharply as the population proportion targeted approaches the headcount. This is where the maximum poverty effect occurs (see table 3.4). As a result, PMTs bring about a greater effect on poverty over a considerable range in the case of headcounts of 25 and 50 percent. This does not occur with higher-order FGT measures.[25]

These peculiar effects of transfers on the headcount are not as stark for PMT-based targeting. This, however, points to a problem: the effect on poverty headcounts is larger because the PMTs

[25] Another way to rationalize this is to note the difference between ideal targeting in terms of maximizing the effect on the headcount and ideal targeting as we define it, namely, targeting only the poorest. In order to maximize the effect on the poverty headcount, resources should be targeted to those close to the poverty line in order to lift them out of poverty.

erroneously identify some of the not-so-poor as very poor, a point we already made with the analysis of ROC-curves when we considered only a tenth of the population as poor. Similar inconsistencies are not observed for the poverty gap and squared poverty gap indices. Instead, we observe minima with perfect targeting when only a small fraction of the population is targeted. In most cases the poverty effect is maximized when about five percent of the population is targeted. We do not observe this peculiar pattern for a headcount of ten percent (top right graph). In this case and with perfect targeting, all the poor that receive the transfer are initially lifted above the poverty line. The transfer becomes insufficient to lift the lowest percentiles above the poverty line when all the poor (i.e. ten percent of the population) are targeted. From that point on, transfers are 'wasted' on non-poor beneficiaries.

Note also that in simulations assuming a headcount of 25 and 50 percent, there is a flatter portion of the curve under perfect targeting for the poverty gap index. The effect is nearly constant over a range of inclusion rates before it starts to decline shortly before all the poor are targeted. In fact, the effect starts to decline as the first poor is reached for which the shortfall is less than the transfer. Again, this reflects particularities of the poverty measure considered, in this case the fact that the poverty gap index is not sensitive to changes in the distribution of real expenditure among the poor.

Finally, we compare the performance of our PMTs to outcomes under ideal targeting. While poverty reductions under ideal targeting strictly dominate those under PMT-based targeting for the FGT1- and FGT2-measures, the differences are often only pronounced for narrowly targeted schemes and when the headcount is low. In that case, perfect targeting results in a poverty effect that is about 65 percent greater. However, the difference is much less pronounced when 25 or 50 percent are poor.

3.7 Conclusion

Based on Bolivian household survey-data, this study stipulates and evaluates a series of regression-based PMTs in terms of targeting accuracy and potential poverty effects. In line with similar studies, we find that PMT-based targeting results in fairly large margins of errors when the poverty headcount is small and only a small share of the population is targeted. On the other hand, targeting accuracy improves considerably when both the headcount and the targeting share increases. Naturally, leakage also increases but much of the non-poor receiving the transfer will still be close to the poverty line. Finally, a large share of the poor are covered even at high levels of poverty when only the rich are not targeted. Our findings further suggest that targeting using a lengthy list of proxies often does not increase accuracy over more parsimonious models, particularly when the entire population is considered.

Interestingly, we find that PMTs are fairly robust over the time period of three years, i.e. PMTs trained based on somewhat dated datasets perform very similar in terms of accuracy and poverty effects. While one would conclude from this that PMTs should rather be used for broad targeting, our poverty simulations suggest that poverty effects are maximized when a small fraction—about five to ten percent—of the population is targeted. PMT-based targeting achieves a poverty effect of about three-fourths of what a perfectly targeted program would achieve at this level of targeting. Again, differences across PMTs are small.

While our study shows that the effect of PMT-based schemes on poverty should be evaluated in addition to traditional targeting measures such as total error rates and true and false positive rates, results from our poverty simulations should be interpreted cautiously. It is important to keep in mind that we do not consider any behavioral responses and that the budget available for the poverty-alleviation program likely depends on the population share targeted. Political-economy considerations would suggest that schemes that target a larger fraction of the population command more political support and thus dispose of larger budgets. If these considerations were taken into account, it would seem reasonable to assume that the trade-off between targeting accuracy and poverty impact would be somewhat less severe and that the maximum poverty effect is reached at higher levels of targeting.

A more general problem with PMT-based targeting is that one has to decide on a well-being variable on which to train models. Measurement error (e.g. item-non response) in household survey data may be a problem but it seems there is little that can be done. It is also unclear whether our indicator of economic welfare captures all of what the population targeted itself sees as relevant for their well-being. While we consider household expenditure rather than income (which is even more problematic) and adjust for price differences and differences in the size and demographic makeup of households, certain dimensions of well-being cannot be captured within such an approach. Alternative methods that acknowledge such measurement problems (e.g. community-based targeting) potentially suffer from other severe problems. There is still much to be said about addressing it by implementing complementary measures that tackle particular dimensions of poverty that are found to be uncorrelated with real expenditure and PMT scores. It would certainly be worthwhile to assess the instruments developed here in terms of community satisfaction with the resulting allocation of benefits.

3.A Calculation of the welfare indicator

3.A.1 General considerations

We use data from the consumption aggregate in order to calculate an indicator of economic welfare. The reliability of this approach depends primarily on data quality and availability. Excessive item non-response and other measurement errors (such as non-random recall bias) may undermine the usefulness of the aggregate as a welfare measure. Depending on the application at hand, problems associated with item non-response may be addressed to some degree by reasonable imputation methods. In our application, however, it is important to stress that imputations should not rely too heavily on additional information that are also used to later predict economic welfare. Imputations based on, say, household size and location, would be self-defeating when the same variables are employed to predict welfare. In what follows we rely on agnostic imputations whenever viable.

The expenditure categories considered include expenditure on food (including consumption of food produced by the household and the value of food received without payment), meals outside the home, other non-food items, education, rents, services, and the service flow from durable goods. In the case of expenditure on food and own-consumption, we replace expenditure on items greater than 3.5 times the mean of log per capita expenditure in this category by median per capita expenditure in the respective primary sampling unit. We also remove in total nine households from the dataset that do not report any food expenditure, neither inside nor outside the home.

Rents are actual rents payed or, in case the household owns the dwelling, hypothetical rents reported by the household. This may be problematic as home-owners may have no knowledge of the rental market. While potentially self-defeating, we have no choice but to rely on inputing values based on hedonic price regressions when information on both actual and hypothetical rents are missing. Rents, however, only account for a small share of total expenditure (see table 3.5).

Including the purchase value of durable goods (e.g. refrigerators) would likely introduce measurement error. These items are typically purchased at a certain point in time while they are used over several years. Households that actually have the same level of economic welfare may thus appear different only because of differences in the timing of purchases. The preferred alternative if data on time of purchase, purchase price, and (hypothetical) current value is available is to calculate service flows as the depreciation rate. We do so following the procedure pointed out by Deaton [1997]. Data on deposit and inflation rates were obtained from ECLAC [2010, p. 138] and the web page of the *Instituto Nacional de Estadísticas* (INE), respectively. Our results concerning median depreciation rates (not reported) are in line with what Deaton reports for similar countries.

Economic shocks such as unexpected hospital fees are a potential source of measurement error: such unexpected outlays will increase total expenditure even though necessary one-off spending on hospitalization may have displaced essential expenditure on, say, food. We therefore exclude reported expenditure that is related to some adverse shock, in particular, hospitalization fees. Moreover, including repairs and enhancement of durables (e.g. repairs of a car or furniture) may lead to double counting when the increase in value resulting from the enhancement is already captured in the current value reported by households. We therefore exclude expenditure reported on car repairs. Furniture, another category that would be available, is excluded on the ground that it is likely included in the durables listings.

3.A.2 Equivalence scale

An indicator of economic welfare should account for the fact that (i) household members do not necessarily require the same amount of expenditure in order to arrive at the same level of welfare and (ii) not all goods are completely private within households. The first point is obvious enough: depending on physical stature, humans require different amounts of calories and thus different levels of expenditure on food. The second point refers to goods and services that are partly public in the sense that utility per person decreases less than proportionally in the number of household members. The dwelling itself is a good example: it seems reasonable that the same level of expenditure *per person* on the dwelling will lead to higher utility levels *per person* for a larger household. We thus require an *equivalence scale* that accounts for these two points.

We rely here on the simple yet flexible functional form suggested by Citro and Michael [1995, p. 159] [see also ECLAC, 2006, p. 38], which can be written as

$$s_i = (A_i + pK_i)^\theta,$$

where A_i is the number of adults in household i, K_i is the number of children, p is the proportion of a child's needs relative to an adult's, and $\theta \in [0, 1]$ is a measure of economies of scale. $\theta = 0$ would imply that all within-household consumption is completely public and $\theta = 1$ would imply that consumption is completely private. For our analysis, we define 'children' as household members below the age of 14 and assume that $p = 0.7$ and $\theta = 0.8$. We consider neither domestic servants nor their relatives, as we assume that the larger share of their consumption goes unrecorded. We exclude all expenditure that can be attributed to these individuals from our calculations.

Deaton [1997, p. 205] notes that "the state of knowledge and agreement in the area is not such as to allow incontrovertible conclusions or recommendations." Theory-based methods to estimate equivalence scales are not yet fully convincing. The recommendation is thus to rely on arbitrary but reasonable equivalence scales. On the other hand, Drèze and Srinivasan [1997] show that if household expenditure can be divided between purely private and purely public components so as to maximize average utility among identical members, θ is equal to the share of private goods in total household expenditure. Expenditure shares are reported in table 3.5 below. Food is usually considered a purely private good and its expenditure share is about 60 percent. This can be considered a lower bound for θ since other goods and services are mostly neither fully private nor fully public. Our choice of $\theta = 0.8$ is thus sensible.

3.A.3 Differences in prices

Using consumption expenditure properly adjusted for household size and composition as an indicator of economic welfare will be misleading if prices for relevant goods and services vary across households. Rather, the welfare indicator should reflect the command over commodities of a given individual. A way forward is to use a *true cost-of-living index* that reflects the price of a reference bundle of commodities as a deflator [Ravallion, 1998]. Here, we use the

Table 3.5: Distribution of expenditure and expenditure patterns.

	2011			2008		
	All-Bolivia	Urban	Rural	All-Bolivia	Urban	Rural
Panel A: Means						
Per capita expenditure	908	1,072	577	751	915	440
Per capita real exp.	897	1,059	571	852	1,038	498
Per adult real exp.	1,258	1,479	812	1,197	1,454	710
Food	61.8	56.2	73.1	61.2	56.0	71.1
Other non-food	11.5	12.0	10.6	14.6	14.7	14.3
Education	6.5	7.3	4.9	6.4	7.1	4.9
Rent	11.8	14.7	5.9	10.1	12.4	5.3
Services	3.6	4.3	2.0	3.9	4.8	2.0
Durables	4.9	5.6	3.6	4.2	5.0	2.8
Panel B: Medians						
Per capita expenditure	719	854	465	582	718	346
Per capita real exp.	710	843	460	661	816	391
Per adult real exp.	1,052	1,223	683	986	1,189	598
Food	63.3	57.6	75.2	62.7	57.0	72.4
Other non-food	9.9	10.8	8.8	12.3	12.7	11.9
Education	4.6	5.4	3.2	4.5	5.2	3.3
Rent	8.9	12.1	4.4	7.7	9.9	4.0
Services	3.2	3.8	1.6	3.4	4.2	1.5
Durables	3.2	3.9	1.8	2.7	3.5	1.5
Observations	8,842	3,937	5,954	2,328	2,888	1,609

Calculations weighted using household inflation factors. Authors' own calculations based on data from the *Encuesta de Hogares* 2008 and 2011.

official series of consumer price indices (CPIs) for each of the nine departments. The data were provided by Unidad de Análisis de Políticas Económicas y Sociales (UDAPE)-staff. We proceed by first normalizing CPIs to equal unity in the department of La Paz in 2011. We then deflate expenditure per adult equivalent by the normalized CPIs.

For household i in region r in year t we define *real consumption expenditure per adult equivalent* as

$$y_{rti} = \frac{x_i}{s_i p^*_{tr}}, \tag{3.1}$$

where p^*_{tr} is the price index for geographic region r at time t, x_i is total household expenditure recorded for household i, and s_i is the value of the equivalence scale as explained above.

3.A.4 Expenditure patterns

Table 3.5 reports means of per capita expenditure, real per capita expenditure, and real per adult expenditure per month in Bolivia by quintiles of the later and by location for 2008 and 2011. For 2011 (2008) we find that average real expenditure *per adult* is 1,258bs (1,197bs) while real *per capita* expenditure is 897bs (852bs). This is roughly comparable to results reported by Moratti [2010].

Expenditure shares are broadly in line with our expectations: on average, more than half of total expenditure is food expenditure in both 2008 and 2011. Service fees and service flows from durables play a minor role across all subsamples. Overall, these shares are well in line with what other studies find. For instance, they accord closely to what Deaton and Zaidi [2002, p. 24] report for Ecuador for data from the mid-1990s.

Chapter 4

Livestock as an Imperfect Buffer Stock in Poorly Integrated Markets

Abstract Livestock holdings in rural areas of the West African Semi-arid Tropics (WASAT) are often substantial yet there is little evidence for precautionary saving in the form of livestock. This paper re-visits farm households' ability to smooth consumption *ex post* via savings in the form of livestock. Based on data covering Burkina Faso's 2004 drought, we find that livestock sales increase significantly in response to drought with households citing the need to finance food consumption. Some consumption smoothing is achieved via adjustments to grain stocks, but households apparently fail to smooth consumption by adjusting livestock holdings. We argue that this seemingly contradictory finding is largely due to a decrease in relative livestock prices during droughts. This renders selling livestock a costly coping strategy and underlines the need for market integration.

4.1 Introduction

Understanding whether and how the poor are able to smooth consumption in the event of adverse income shocks is of great importance. The theory of optimal savings in the absence of formal insurance mechanisms and credit markets predicts that households facing covariate risks will use liquid assets for self-insurance [e.g. Deaton, 1990, 1991, 1992, Zeldes, 1989]. Since poor households are more likely to be credit-constrained, they are also more likely to hold extra savings. However, such behavior may give rise to poverty traps by preventing agents from undertaking profitable investments that are viable in principle [Fafchamps and Pender,

The chapter is based on joint work with Malte Reimers. We are grateful to Kimseyinga Savadogo, Denis Akouwerabou and Omer Combary of the University of Ouagadougou and to the Burkinabe Ministère de l'Agriculture, de l'Hydraulique et des Ressources Halieutique for sharing their data. We have benefited from comments by Joachim von Braun, Michael Grimm, Stephan Klasen, Matin Qaim, Hermann Waibel, Marijke Verpoorten, and Meike Wollni. Participants at the AEL Conference 2014 in Passau, the PEGNet Conference 2014 in Lusaka, the NEUDC Conference 2014 in Boston, the CSAE conference 2015 in Oxford, as well as seminar participants in Göttingen provided valuable comments.

1997, Zimmerman and Carter, 2003]. The existence of such traps would strengthen the case for governments to intervene in order to correct market failures.

The West African Semi-arid Tropics (WASAT), which have been the site of recurring droughts [Shanahan et al., 2009], are an ideal setting for researchers to study households' responses to shocks in terms of adjustments to buffer stocks. The set of strategies to cope with drought-induced shortfalls in income resident households can choose from is severely limited. Despite substantial covariate risks [Carter, 1997], households typically lack access to formal insurance mechanisms [Binswanger and McIntire, 1987, Binswanger and Rosenzweig, 1986]. Informal insurance arrangements that do not extend beyond villages are ineffective as adverse shocks are to a large extent covariate [Carter, 1997, Sakurai and Reardon, 1997].

Against this background, it has long been hypothesized that in this region and elsewhere in developing countries livestock sales are an important means to smooth consumption [Binswanger and McIntire, 1987, Reardon et al., 1988].[1] However, empirical studies investigating households' responses to adverse shocks are inconclusive. Based on data from India and war-time Rwanda, Rosenzweig and Wolpin [1993] and Verpoorten [2009] find that households increase sales of livestock during droughts and episodes of civil conflict, respectively. In contrast, Fafchamps et al. [1998], Kazianga and Udry [2006], and Carter and Lybbert [2012] find—using data collected in rural Burkina Faso during the early 1980s—that a large majority of households do not generate additional revenues via sales.[2]

In this paper we systematically re-assess the importance of livestock as a buffer stock in developing countries using two large and recent panel datasets from rural Burkina Faso. We offer a straight-forward explanation for existing ambiguities and thus reconcile previous empirical studies in the literature. As a first step, it is important to recognize that the above two strands of the literature actually pursue slightly different questions which are both, however, related to distress sales:

1. Do households increase (net) sales of livestock in times of economic hardship (e.g. droughts, conflict)?
2. Does net saving in the form of livestock vary positively with transitory income, i.e. do households generate additional revenues for consumption from sales of livestock in times of adverse income shocks?

The difference between the above questions lies in the role assumed by price movements. If prices were constant, the answer to both questions would always be the same. If, however, livestock prices varied positively with transitory income, which is to be expected when markets are poorly integrated and shocks are correlated across households, an increase in net sales would not necessarily translate into an increase in revenues. The above studies differ in which of the two questions they address: while Rosenzweig and Wolpin [1993] and Verpoorten [2009] find evidence for an increase in sales in response to adverse shocks, studies that focus on precautionary savings typically find no relationship between savings in the form of livestock and transitory income [Carter and Lybbert, 2012, Fafchamps et al., 1998, Kazianga and Udry, 2006].

[1] Binswanger and McIntire [1987] argue that animals are more resistant to droughts as there might still be vegetative growth that provides fodder and that animals can be shifted to neighboring areas in case of local droughts while grain storage is expensive as stocks exhibit limited durability and are often affected by pests.

[2] Carter and Lybbert [2012] find that a small fraction of households, those with large livestock holdings, that accounts only for ten to 15 percent of their sample do generate additional revenues through sales.

Instead of investigating only one of the above questions in isolation, as previous studies have done, this paper addresses both questions at the same time in order to explain the apparent contradiction. We employ two large panel datasets from Burkina Faso that cover the harvests of 2003 and 2004 as well as 2004–2007, respectively, which saw considerable variation in rainfall including a drought in 2004. Our findings with regard to the first question suggest that livestock sales increase in response to adverse rainfall shocks at the province-level with no off-setting increase in purchases. Based on count data models, we also find that sales are negatively related to rainfall at the household-level. Consistent with a need to compensate for a decrease in revenues from cropping, households cite food consumption as the main motive for extra sales.

We then turn to the second question and ask whether households save in the form of livestock out of transitory crop income based on an Instrumental Variables (IV)-estimator that identifies the transitory component of crop income from unanticipated variation in rainfall. Our framework shares key components with specifications typically employed in the literature and results in similar findings. Importantly, we find no evidence for a significant role of savings in the form of livestock. In contrast, there is evidence for an effect of transitory income on savings in the form of grain stocks and a sizeable yet somewhat less robust effect on consumption expenditure.

Viewed in isolation, our results are largely in line with previous findings in the literature and beg the question of why there are additional sales yet no additional revenues. We argue that prices account for this puzzle: in a province-level panel dataset, we show that cattle prices decline in the event of an adverse weather shock. This is consistent with drought-induced sales in markets that are not fully integrated. Price-effects potentially explain the lack of correlation between transitory income and net purchases of livestock in monetary terms: an increase in the net number of animals sold is off-set by a decrease in livestock prices. This renders adjustments to livestock holdings a costly strategy to smooth consumption. It also explains why households bear consumption cuts despite livestock holdings that would allow them to completely offset transitory income losses. At least some households will find that post-shock prices are too low for their livestock and hence abstain from selling. While a potential role for price effects has been acknowledged in the theoretical literature [Park, 2006, Zimmerman and Carter, 2003] and was also discussed by Fafchamps et al. [1998], these insights are often ignored or addressed only insufficiently in empirical work. This paper aims to fill this gap.

The paper proceeds as follows: section 4.2 describes the setting of our empirical investigation. Section 4.3 introduces and compares datasets used in this study and provides summary statistics. Section 4.4 provides some descriptive evidence on behavioral responses to the drought in 2004. Based on province-level panel regressions and household-level count data models, we show in section 4.5 that sales increase in response to adverse rainfall and that there is no off-setting change in purchases. Section 4.6 details our empirical strategy to identify the effects of transitory crop income on consumption and savings in the form of grain storage and net purchases of livestock and presents our results. Section 4.7 investigates livestock price responses to adverse rainfall. Section 4.8 concludes and discusses policy implications.

4.2 Agricultural production in Burkina Faso

Burkina Faso is a landlocked country located in West Africa. It is among the poorest countries in the world with a purchasing power parity (PPP)-adjusted Gross Domestic Product (GDP) per capita in 2009 of about $450 (in 2005 dollars) [World Bank, 2013]. More than two-thirds of the population live on less than $2-a-day. As in other Sahelian countries, most of the poor still live in rural areas where livelihoods depend crucially on rain-fed agriculture. While agriculture accounts for about one-third of total output, the sector employs four in five persons in Burkina Faso. The rate of technological change in agricultural production in the WASAT is low [Eicher and Baker, 1982].

According to definitions by Sivakumar and Gnoumou [1987], there are three distinct climatic zones in Burkina Faso: the *Southern Sudanian Zone* to the south characterized by rainfall above 1,000mm annually on average, a rainfall season that lasts for more than six months, and comparatively low temperatures; the *Central North Sudanian Zone* with rainfall between 650 and 1,000mm that does not exceed six months; and the *Northern* or *Sahelian Zone* with a short rainy season, considerable variability in rainfall and high temperatures. These three climatic zones roughly coincide with the shaded areas in figure 4.1 which clearly shows the north-south-gradient in average annual precipitation in millimeters between 2001 and 2012. While onset dates differ somewhat across regions, most of the rain is typically received between June and September.

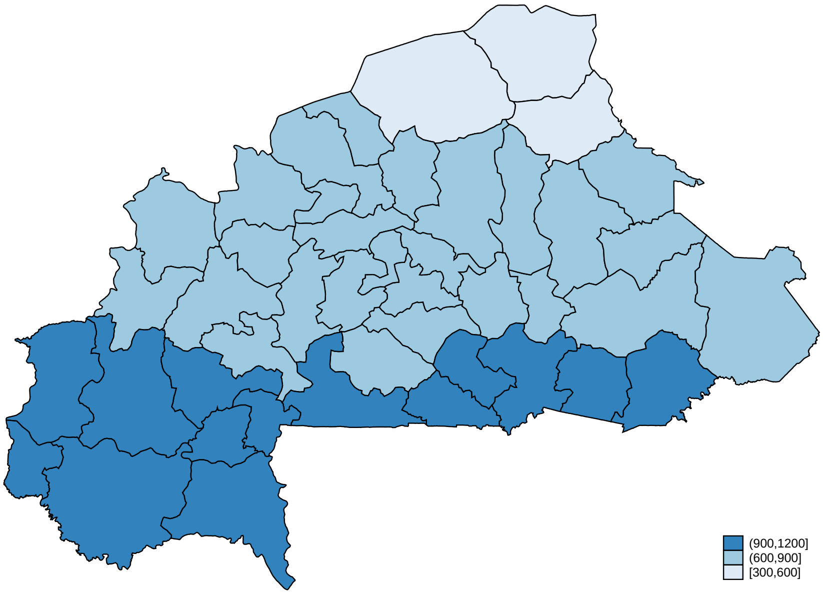

FIGURE 4.1: Average annual precipitation in millimeters, 2001–2012.

Crops grown for domestic consumption are mostly sorghum, millet, maize, and rice [see Kassam, 1979, for an early exposition]. Millet is the dominant staple crop in the arid northern provinces, whereas sorghum is the principal subsistence crop elsewhere [Sivakumar and Gnoumou, 1987]. The most important cash crop is cotton and its production has increased rapidly over the last years. It accounts for more than half of the countries' export earnings [Amo-Yartey, 2008]. Production is concentrated in the southwestern regions. In the Sahel, in contrast, cash crops include groundnuts and sesame but at a smaller scale [Traore and Owiyo, 2013].

Animal husbandry, dominated by cattle, goats, and sheep, is traditionally an important income source in the northern and eastern parts of the country [Sivakumar and Gnoumou, 1987]. Recent reports, however, point to a decline in the economic importance of livestock in the Sahel—particularly in response to droughts in 2004 and 2010. For instance, Traore and Owiyo [2013] cite an accelerated degradation of pastures in recent years. Consistent with an intensification of livestock management practices, respondents in their study also report a shift away from purely cattle-based livelihoods towards a combination of crop production and livestock keeping.

4.3 Datasets and descriptives

For the empirical part of our paper, we use two different panel datasets which both cover the 2004 drought, the *Enquête Permanente Agricole* (EPA) and the *Deuxième Programme National de Gestion des Terroirs* (PNGT) datasets. These datasets are separately described below where we also provide a comparison of our data with the data collected by the International Crop Research Institute for the Semi-Arid Tropics (ICRISAT) that previous studies of consumption smoothing in Burkina Faso relied on. We combine the panel datasets with precipitation data from the Famine Early Warning Systems Network (FEWS NET) [USAID, 2013], also described below.

4.3.1 EPA surveys

For our analysis, we use four waves of nationally representative EPA data collected between 2004 and 2007. The survey is designed to provide *inter alia* reliable information on agricultural production, livestock holdings and transactions, as well as grain stock holdings. The balanced panel we retain for our analysis contains 2,364 households residing in all 45 provinces of Burkina Faso. Data collection was conducted by enumerators coming from the same area who are typically farmers themselves and are hence familiar with local conditions for agriculture. This proximity enables them to visit surveyed households repeatedly, usually at least once during the growing and once during the harvest season (see also figure 4.2 for the timing of surveys).

Given the survey's major importance for the Burkinabe Ministry of Agriculture as a tool to collect information on recent and expected harvests, extraordinary efforts are made to capture each household's agricultural production as precisely as possible. For example, each local enumerator is equipped with an isosceles triangle which is used at the first visit during the growing season to randomly mark on each plot an area of exactly 25m^2 with wood pegs. Randomness is assured by following an exhaustive predefined process. Shortly before the plot is ready to be harvested, the household head contacts the enumerator in order to schedule a second visit. This allows the enumerator to be present when the marked area is harvested, threshed, and weighed. The measured output is ultimately extrapolated to the entire plot area.[3] The above-described procedure together with very strict protocols also for other questionnaire modules lets us be confident that the data are of high quality.

[3] This procedure apparently worked quite well since, according to the data, an enumerator was present at the time of harvest in almost 60 percent of the cases.

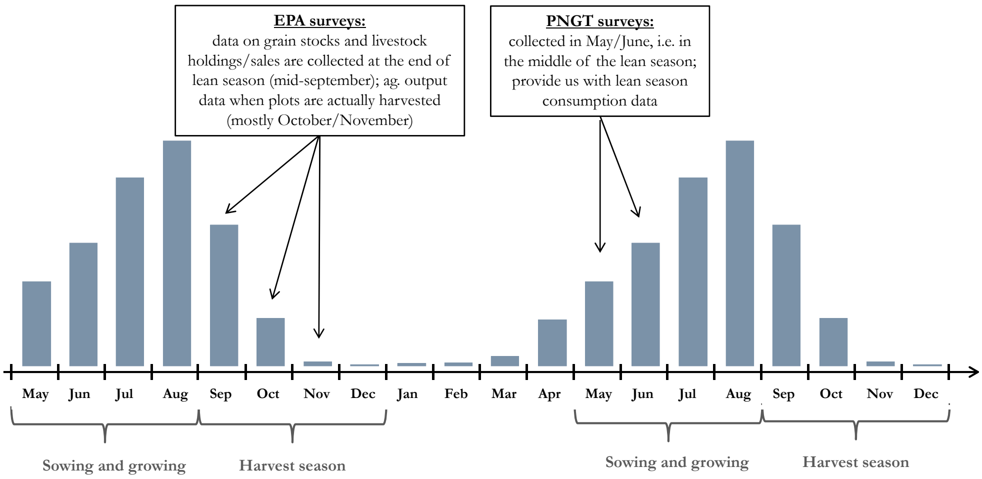

FIGURE 4.2: Timing of surveys relative to average levels of rainfall and the agricultural cycle. Authors' calculation based on data from USAID [2013].

4.3.2 PNGT surveys

We further use two waves of the PNGT panel surveys for our analysis which were collected in May and June of 2004 and 2005, respectively.[4] These surveys are administered by the University of Ouagadougou in collaboration with the Burkinabe Ministry of Agriculture and aim to quantify improvements in the livelihoods of households in rural Burkina Faso. The data we retain cover a total of 1,492 households in 60 villages in all of Burkina's 45 provinces [Wouterse, 2011].

Our main motivation for using the PNGT data lies in the fact that they include a consumption and expenditure module as well as detailed data on self-reported consumption cuts. They come complete with a village questionnaire which we use to extract price data. The data are collected during each year's lean season allowing us to grasp the extent to which harvest shortfalls translate into actual reductions in food consumption without relying on flow accounting methods (as in Kazianga and Udry [2006]).

Figure 4.2 depicts the timing of the two surveys relative to average levels of rainfall and the agricultural cycle.[5] Since the EPA surveys are fielded before and after harvesting and PNGT surveys during the lean season, crop output in a given survey year is associated with different harvests—current year or previous year—depending on the provenance of the data.

4.3.3 Comparison to ICRISAT data

It may be helpful to compare our data to the ICRISAT data used in other prominent studies of consumption smoothing in Burkina Faso [Carter and Lybbert, 2012, Fafchamps et al., 1998, Kazianga and Udry, 2006]. The ICRISAT data were collected between 1981 and 1985 and cover households in six villages in three different agro-climatic zones. Similar to our data, the survey provides data on crop production, livestock holdings and transactions, and grain stocks[Matlon, 1988]. There is also data on consumption albeit only for the last two years, i.e. 1984 and 1985.

[4] We also have access to a third PNGT wave collected in November 2006. However, after careful consideration, we decided not to use this dataset since the change in survey timing would complicate comparisons over time.

[5] See also Bonjean et al. [2012] for a more detailed depiction of the agricultural cycle in Burkina Faso.

Table 4.1: Descriptive statistics: means and standard deviations by year.

	EPA				PNGT		
	2004	2005	2006	2007	2003	2004	2005
Household size	10.95	10.97	10.92	10.95	-	9.21	10.10
	(6.74)	(6.85)	(6.94)	(6.83)		(5.79)	(6.30)
Age of HH head	50.99	51.27	51.53	51.77	-	48.26	49.21
	(14.90)	(14.85)	(14.47)	(14.60)		(15.50)	(15.32)
Mean age of HH members	22.37	22.70	23.02	23.14	-	22.31	22.31
	(7.96)	(8.52)	(8.61)	(8.56)		(8.93)	(8.71)
Cultivated area (ha)	4.03	4.15	4.04	4.09	-	5.31	5.09
	(3.37)	(3.58)	(3.47)	(3.48)		(5.08)	(5.41)
Agg. grain output (kg)	2,233.43	2,955.98	2,861.26	2,566.60	1,716.10	1,206.49	-
	(2,388.95)	(2,968.61)	(2,870.08)	(2,914.91)	(1,825.60)	(1,345.96)	
Crop profit (1,000 CFA)	486.63	572.81	567.36	463.56	322.43	273.43	-
	(547.04)	(607.03)	(575.17)	(551.02)	(430.60)	(340.09)	
Agg. grain stock (kg)	324.61	116.04	283.23	301.96	-	-	-
	(753.25)	(408.43)	(625.40)	(690.89)			
Herd size (Cattle equiv.)	8.13	7.67	7.90	7.53	-	4.67	4.79
	(20.35)	(18.35)	(22.55)	(18.78)		(11.73)	(12.36)
Observations	2,364	2,364	2,364	2,364	1,492	1,492	1,492

Standard deviations reported in parentheses. To calculate *tropical livestock units* (TLUs) we follow Jahnke [1982]: cattle enters with a weight of unity while sheep and goats enter with a weight of one-seventh. TLUs are thus 'cattle equivalents.'

While the strict protocols for data collection that come with the EPA surveys ensure that the data are of comparably high quality, there are three main advantages in comparison to the ICRISAT data: first, both the EPA and PNGT surveys aim to be representative of rural Burkina Faso. While our data inlude information on households residing in all of Burkina's 45 provinces, the ICRISAT data cover only six villages. Second, our datasets both provide much larger sample sizes than the ICRISAT data. They cover 2,364 and 1,492 households in each year, respectively, while the ICRISAT data only provide information on 126 households. Finally, our data were collected during the mid-2000s while the ICRISAT surveys date back to the early 1980s.

4.3.4 Descriptives

Table 4.1 reports descriptive statistics for some of the key variables in our analysis. As further detailed in section 4.4, our data cover a major drought in the northern part of Burkina Faso in the year 2004 which caused the grain output to be considerably lower. Households in the EPA data comprise on average around eleven members while households are slightly smaller in the case of the PNGT data. Also, we observe that the latter have younger household heads and fewer livestock holdings. The figures also seem to indicate that they have lower grain output despite cultivating more land.[6]

The drought year of 2004 is apparent in the data via a reduction in aggregate grain output and crop income[7] that is observed in both datasets. Grain stocks, which are only available from the EPA datasets, are much lower in 2005 following the 2004 drought. However, averages reported in table 4.1 disguise considerable spatial variation. Therefore, section 4.4 scrutinizes the effects of the 2004 events in more detail with a particular focus on livestock related variables.

[6] A possible explanation for this is recall bias: output is directly measured in the case of the EPA data, often in the presence of enumerators, whereas the PNGT data rely on recalls elicited during the following lean season, i.e., several months after harvesting.

[7] What we refer to as *crop income* is actually *net crop income*, the value of crop output less the value of agricultural inputs excluding labor.

4.3.5 Prices

A challenge we encounter with the otherwise excellent EPA data is the elicitation of crop prices. The data neither come with a village level-survey in which local market prices were collected separately by enumerators as in the case of the PNGT data, nor do they include a consumption aggregate. Hence, we would be left with prices inferred from unit values calculated from households' reports on sales of agricultural output. Such sales, however, are rare in an environment characterized by subsistence farming and low incomes. In general, households sell their grain output only when prices are particularly good.[8] Households in our data are rarely net sellers of crops. Asked about the proceeds from the 2003 harvest in mid-2004, in as many as eight out of 45 provinces less than ten households report sales of millet. This share increases to 21 out of 45 provinces in 2004/2005.[9]

Another potential problem is intra-seasonal price variation. Respondents were asked about sales of output between the last harvest and the time of the interview, a period of almost ten months in case of the EPA surveys. Intra-seasonal prices in African agriculture are known to fluctuate substantially within localities, a phenomenon that has received some attention recently [Burke, 2014, Stephens and Barrett, 2011]. We would thus expect average unit prices based on reported sales to exhibit high variability within province-years.

We therefore rely on an alternative that takes advantage of the fact that the PNGT data provide us with village-level prices from all provinces at three different points in time between spring 2004 and fall 2006 (see also footnote 4). These data are used to calculate province-level prices and then supplemented with monthly crop price data for Ouagadougou from the *Statistical Yearbooks* of the *Institut National de La Statistique et de la Démographie* [INSD, 2012]. We impute province-level prices based on regression models. Details are reported in appendix 4.A.

These prices together with average expenditure shares estimated based on data from the PNGT surveys' consumption modules are then used to compute a province-level consumer price index (CPI) as follows: first, we calculate the average annual quantities consumed of the major food items for which we have prices in the PNGT village-level surveys over all households and years. Second, these quantities are valued using the current province-level market prices giving us the monetary value of the food basket. On average, food expenditure accounts for roughly two-thirds of total consumption expenditure of households in the PNGT sample. For non-food expenditures, accounting for the other third of the basket of goods, we assume a moderate inflation rate of three percent annually. The resulting CPI is normalized to be unity on average across all households and time periods.

4.3.6 Precipitation data

The precipitation data used in this paper come from USAID [2013] and are estimated based on a combination of actual station-level rainfall data and satellite-measured cloud top temperatures. For our analysis, we use province-level precipitation data for ten-day intervals for the years 2001

[8] For instance, Barrett and Dorosh [1996] report that of their sample of rice-producing farm households in Madagascar only five percent of households accounted for about half of rice sales while about 60 percent purchased rice. Similarly, Budd [1999] shows that few of the farm households in Côte d'Ivoire that he studies were fully self-sufficient and very few were net sellers. This finding is usually attributed to high transaction costs in environments with poor infrastructure [e.g. Park, 2006, Renkow et al., 2004].

[9] The problem is even more pronounced in the case of non-grain crops such as wandzou for which we hardly ever observe more than ten transactions from which unit prices could be calculated.

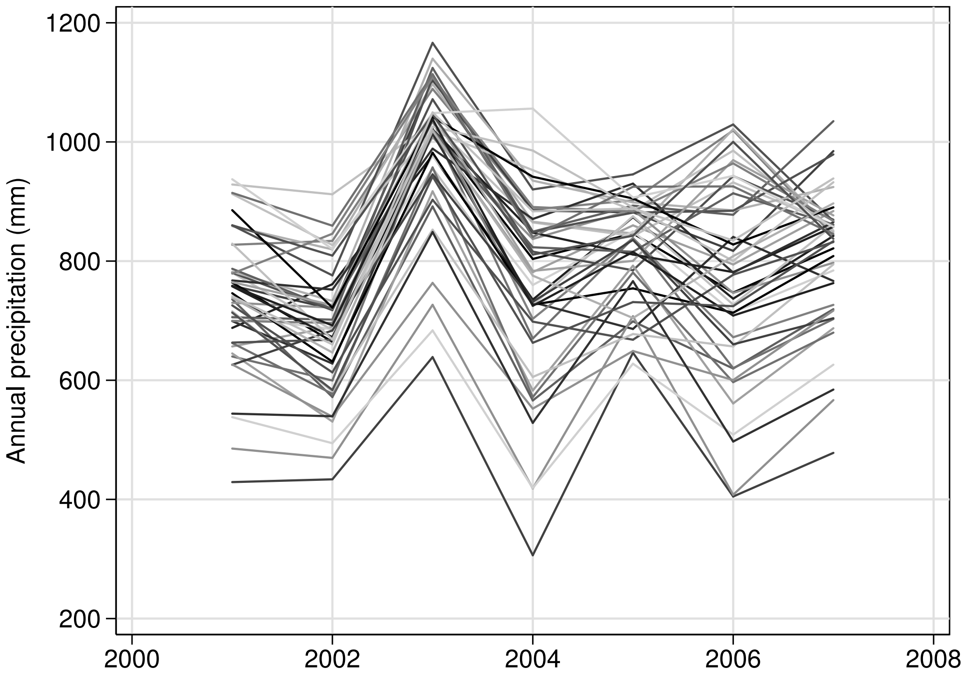

Figure 4.3: Annual precipitation in 45 provinces, 2001–2007. Authors' calculation based on data from USAID [2013].

until 2012 (i.e. 36 data points per year and province) which we aggregate to total rainfall in millimeters for each province-year.

While this may seem a simplistic indicator of rainfall at first sight, it is appropriate in the context of Burkina Faso as rainfall follows a unimodal distribution and is highly concentrated during the summer months (see also figure 4.2). Nevertheless, we experimented with more complex measures of rainfall such as squared terms, quarterly aggregates, and estimates of the duration of the rainy season. This did not significantly improve the predictive power with regard to net crop income, the relationship of interest in our IV framework below.

Another issue is the level of aggregation. Ideally, one would want to use rainfall data at the village-level which, unfortunately, is not available to us. However, province-level rainfall still results in a valid first stage in our IV models below. In our reduced-form regressions, a higher level of aggregation renders our estimates lower bounds as this may introduce measurement error in the main explanatory variable.

The resulting panel is depicted as a time series plot for all of Burkina's 45 provinces in figure 4.3. It is clear from this figure that 2003 was a particularly good year in almost all provinces whereas 2004 saw less rainfall.

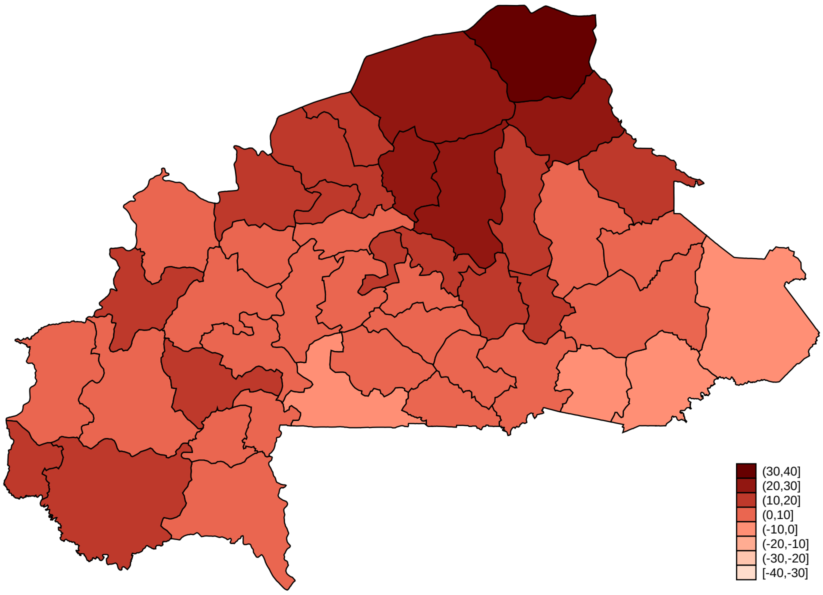

Figure 4.4: Shortfall in precipitation relative to long-term mean, 2004. Authors' calculation based on data from USAID [2013], 2001–2012.

4.4 Impact of the 2004 drought

In terms of agricultural production, 2004 was a particularly bad year for farmers in the northern provinces of Burkina Faso. The rainy season started later than usual, precipitation was irregular and overall rainfall levels were considerably below the long-term mean [FAO, 2005].[10] This shortfall in rain is depicted in figure 4.4, where we report the proportional shortfall in 2004 relative to the long-term mean calculated for 2001–2012. On average, provinces experienced about ten percent less rain in 2004. Provinces most severely hit during that year were Sanmatenga, Namentenga, Soum, Séno and Oudalan, all situated in the north of the country. In the northernmost province, Oudalan, one of the driest provinces within Burkina Faso, rainfall levels were about 30 percent below the long-term mean.

Crop Output The consequences of these events are reflected in figure 4.5 which depicts the shortfall in crop output per hectare relative to the 2004–2007 average. In line with the above explanations, the output shortfall was largest in the three provinces Oudalan, Séno and Soum which suffered from an average shortfall in excess of three-fourths. Also beyond these three provinces, the rainfall map in figure 4.4 and the output per hectare map in figure 4.5 match quite well suggesting a relationship between these two indicators. At the national-level, we observe an average shortfall in output per hectare of slightly more than 25 percent[11]

Food Consumption Lean season data from the PNGT surveys allow us to examine how Burkinabe households reacted to the events of 2004 in terms of consumption. Figure 4.6 depicts the share of households reporting reduced food in-take during the last seven days (separately for men, women and children) as well as the share that left out entire meals or did not consume

[10] In addition, there were reports of desert locust swarms from North Africa invading many West African Sahel countries including Burkina Faso [IFRC, 2005]. This phenomenon also affected primarily northern provinces and is likely to have further aggravated the loss of output and grain stocks in 2004.

[11] All figures are unweighted averages across provinces.

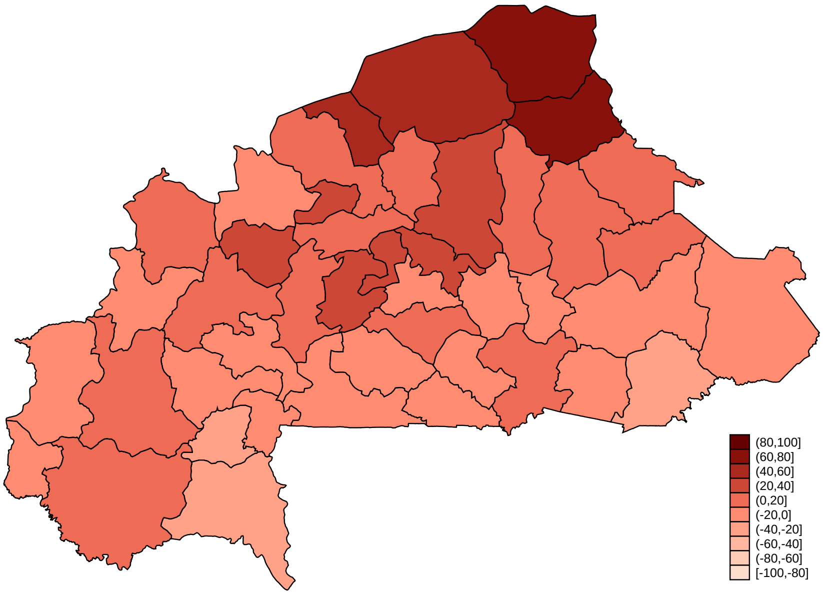

FIGURE 4.5: Shortfall in crop output relative to 2004–2007 average, 2004. Authors' calculations based on EPA data, 2004–2007.

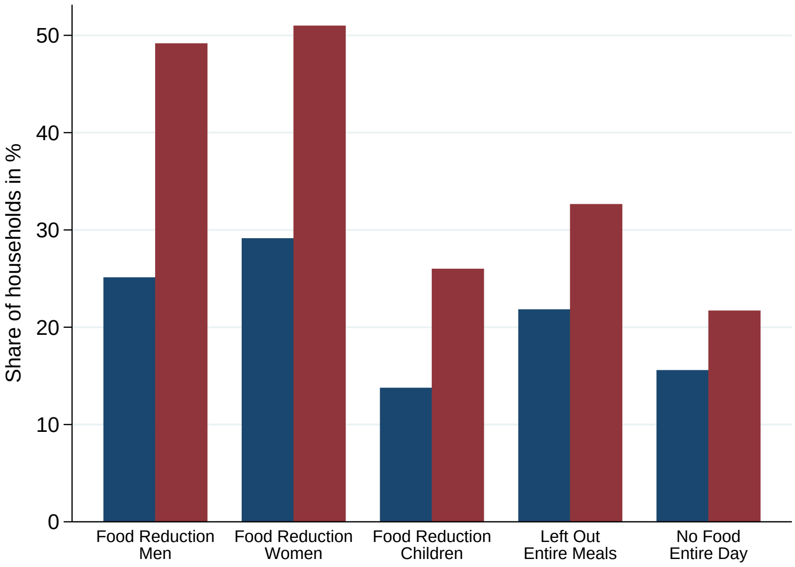

FIGURE 4.6: Reported cuts in food consumption during the 2004 lean season (blue bars) and the 2005 lean season (red bars). Authors' calculations based on data from PNGT surveys.

any food for an entire day in this time period. Three issues are particularly noteworthy. First, a considerable share of households in rural Burkina Faso is structurally poor given that, even following a good year in terms of rainfall and output such as 2003, between 25 and 30 percent of households report reduced food in-take for at least some household members. More than 15 percent of households even report going without food for at least one day during the last week. Second, it seems that households try to protect their children from food cuts to the extent possible given that the share of men/women experiencing reduced food consumption is in both years considerably higher than for children. Third, for all five indicators we see a clear upward shift between 2004 and 2005. Most notably, the share of households reporting reduced

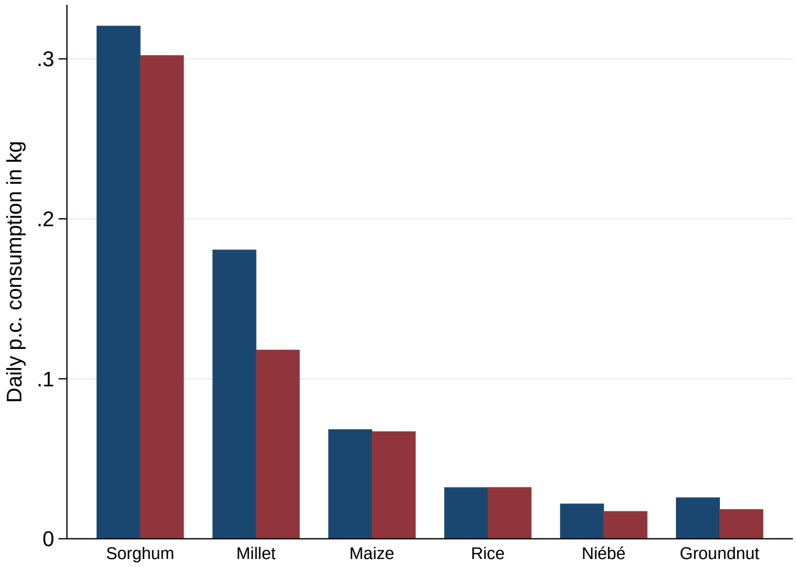

FIGURE 4.7: Daily per capita consumption (kg) of staple food during the 2004 lean season (blue bars) and the 2005 lean season (red bars). Authors' calculations based on data from PNGT surveys.

food in-take for men/women increases considerably to around 50 percent. This trend does not spare children since in the year 2005 approximately 26 percent of households report food cuts for children (compared to 14 percent in 2004). Analogously, the share of households abstaining from food consumption for an entire day also increases to approximately 22 percent.

While these time trends are indicative, it could be argued that they rely on rather subjective, categorical questions. Therefore, we analyze as a next step the daily food quantities consumed per capita for a total of nine crops.[12] As can be seen in figure 4.7, there has been a considerable drop in millet consumption between the years 2004 and 2005 (approximately 70g per person and day). There is also a reduction in in-take of sorghum, groundnut, and niébé, but to a smaller extent. While these reductions appear small at first sight, they nevertheless correspond to a reduction in food intake of approximately 330kcal from these crops (using calorie conversion factors from FAO [2010]). In this context, it should be noted that all nine crops together account for the median household in our dataset for approximately 70 percent of the value of total food consumption.

Livestock ownership and trading Are the 2004 events also reflected in livestock budgets? Table 4.2 reports means of variables related to production and trading of livestock for all available years. Livestock holdings are substantial in this setting: about three-fifths of all households report owning cattle and more than four-fifths own small livestock, i.e. sheep or goats. On average, families own more than five heads of cattle and more than 13 heads of small livestock in any year.

The share of households selling is lower: less than one-fourth of all households sold cattle, while about half report having sold small livestock. While there is virtually no increase in the

[12] Namely, the contemplated crops are: fonio, groundnut, maize, millet, niébé, rice, sesame, sorghum and wandzou. In figure 4.7, only six crops are shown since the average amounts of the other three crops (fonio, sesame, wandzou) are negligible.

Table 4.2: Livestock balance for cattle, sheep, and goats, 2004–2007.

	2004	2005	2006	2007
A. Cattle				
Share of households owning livestock	0.60	0.59	0.60	0.58
Share of households reporting sales	0.22	0.23	0.23	0.22
# of animals owned	6.04	5.76	5.97	5.64
# of animals sold	0.54	0.68	0.54	0.58
# of animals deceased	0.48	0.49	0.40	0.37
# of animals slaughtered	0.03	0.21	0.03	0.02
# of animals purchased	0.34	0.38	0.30	0.31
# of animals born	1.31	1.29	1.17	1.07
B. Sheep and Goat				
Share of households owning livestock	0.85	0.82	0.83	0.83
Share of households reporting sales	0.55	0.50	0.52	0.46
# of animals owned	14.59	13.40	13.52	13.22
# of animals sold	2.59	2.60	2.22	2.10
# of animals deceased	2.65	2.60	1.72	1.92
# of animals slaughtered	0.84	1.70	0.76	0.69
# of animals purchased	1.26	1.09	0.85	0.84
# of animals born	5.98	5.60	5.13	5.08

Based on EPA data.

number of households selling cattle between 2003/2004 and 2004/2005, the average number of cattle sold increases in 2004/2005 from 0.54 to 0.68 animals. No such increase is observed for small livestock although sales were higher in 2004/2005 than in subsequent periods.

While the figures do not indicate that there was an increase in the number of animals that died as a result of drought, the number of animals slaughtered increases substantially in 2004/2005, albeit from a very low level. This is surprising as several previous studies find that households rarely kill animals for own-consumption [see Fafchamps et al., 1998, and studies cited therein].

These averages disguise important spatial variation in sales. Figure 4.8 depicts net livestock sales relative to initial holdings (in percent) between harvests in 2004 and 2005. We combine different categories of livestock by considering livestock holdings and net sales in cattle equivalents (see note to table 4.1). In line with our expectations, the proportion of animals sold net of purchases is highest for the most drought-affected provinces in the North of Burkina Faso where households on average sold more than 30 percent of their livestock. However, this could also be a static effect if households in northern provinces have a higher tendency to engage in animal husbandry because of underlying differences in the rural economy. It is thus plausible that the pattern observed in figure 4.8 is unrelated to *changes* in rainfall and crop output. We will investigate this issue in more detail in section 4.5.

The EPA surveys also collect information on households' motives for livestock sales. The absolute number of sales of cattle as well as sheep and goats by motive is depicted in figure 4.9. We see that food purchases are the most prominent motive in all years. The pattern is fairly stable across years with the exception of sales to pay for food which increase substantially following the 2004 harvest (i.e. recorded in the 2005 data); the number of cattle and sheep/goats sold almost doubles between 2004 and 2005.

Taken together, the figures presented in this section show that livestock sales increased substantially between 2004 and 2005 and that the dominant motive behind sales was households'

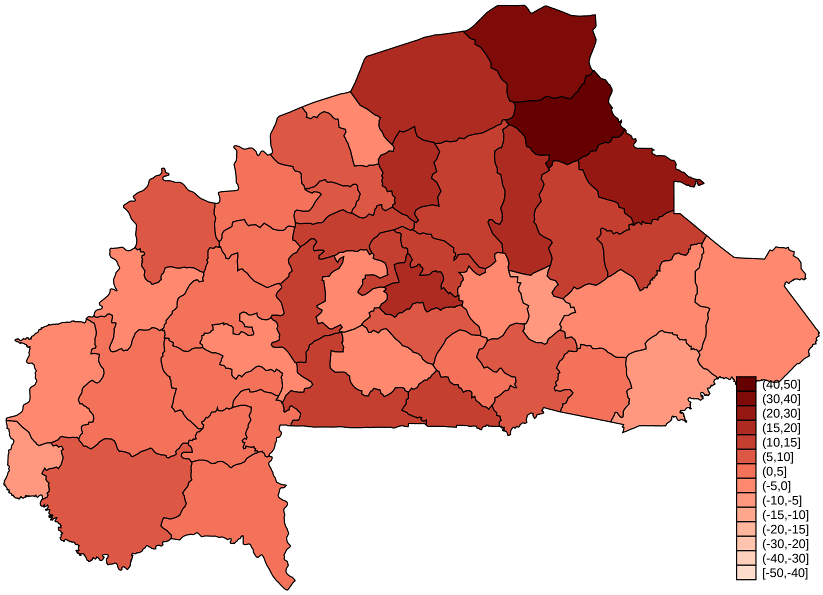

FIGURE 4.8: Net sales of livestock relative to holdings, 2004–2005. Authors' calculations based on EPA data, 2004–2007.

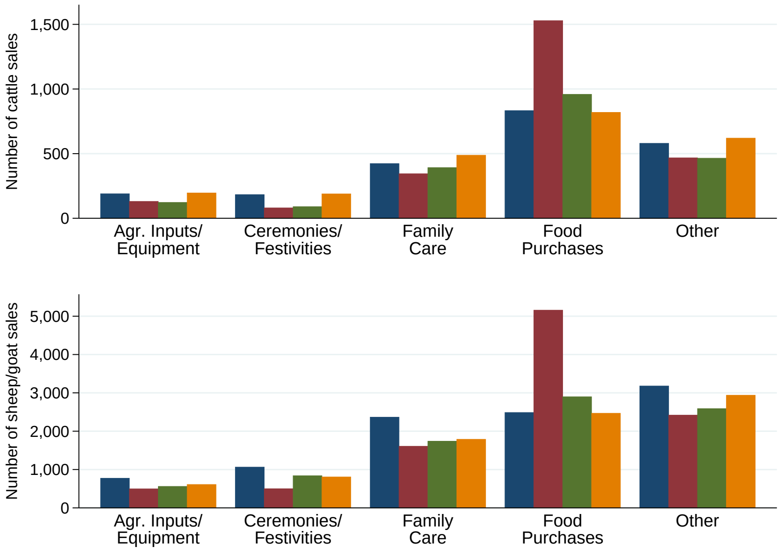

FIGURE 4.9: Motive for sales of cattle (top panel) and sheep and goats (bottom panel) following harvests of 2003 (blue bars), 2004 (red), 2005 (green), and 2006 (yellow). Authors' calculations based on EPA data, 2004–2007.

need to purchase food. It seems plausible that these extra sales were triggered by adverse rainfall conditions and the resulting shortfall in crop output. The next two sections investigate the relationship between rainfall and livestock sales more formally.

4.5 Rainfall and livestock trading

We now turn to the first question framed in the introduction: do households increase net sales of livestock in response to adverse rainfall? While the above descriptives are consistent with this

Table 4.3: Rainfall elasticities of sales and purchases of cattle and sheep/goat, 2004–2007.

	Log quantity sold of...		Log quantity purchased of...	
	...cattle.	...sheep and goats.	...cattle.	...sheep and goats.
	(1)	(2)	(3)	(4)
Log rainfall	-0.72**	-0.81*	0.08	0.09
	(0.34)	(0.44)	(0.30)	(0.36)
Obs.	178	180	179	180
R-squared	0.80	0.58	0.81	0.71

Robust standard errors clustered at the province-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. All regressions include province- and year-fixed effects. Based on EPA data.

notion, we can investigate this conjecture more formally, using regression models at both the province- and household-level.

4.5.1 Rainfall elasticities of livestock sales and purchases

We first run regressions of the form

$$\ln(x_{pt}) = \delta \ln(rainfall_{pt}) + \rho_p + \tau_t + \epsilon_{pt}, \tag{4.1}$$

where x_{pt} denote either aggregated sales or purchases of cattle, sheep, or goats in province p during year t and ϵ_{pt} is the usual error term. ρ_p and τ_t denote province- and year-fixed effects, respectively. The former capture time-invariant differences in livestock production across provinces. For instance, it is plausible that some geographical regions provide a relative advantage in producing livestock such that rural households are more likely to engage in animal husbandry. In that case, we would expect higher sales and purchases in every year. Year-fixed effects, on the other hand, capture trends in the supply and demand conditions that affect all provinces to the same degree such as world market prices for meat.

Since both sales (purchases) and rainfall enter the regression in logs, the coefficient of interest, δ, should be interpreted as the elasticity of sales (purchases) with respect to rainfall. A negative coefficient in a regression of sales on rainfall is consistent with consumption smoothing, i.e. it is consistent with households selling livestock in order to stabilize consumption.[13]

Results for cattle as well as sheep and goats combined are reported in table 4.3. Elasticities reported in columns (1) and (2) suggest that sales of both categories of animals decrease with rainfall. The implied elasticities are large and significant at the five and ten percent-levels for cattle and sheep/goats, respectively. This finding is consistent with livestock serving as a buffer stock and differs from results reported by Fafchamps et al. [1998]. In particular, at the village-level, they find no statistically significant relationship between rainfall and the number of cattle sold and only a weak relationship for sheep and goats.

If local economies were completely isolated, we would observe a concomitant increase in purchases. In that case, we would see animals being traded between farmers forced to sell in

[13] Note that accumulation of livestock is mainly accomplished through new births and under-selling and only to a minor extent through purchases. Here, we aim to investigate the reduced-form relationship between sales and purchases on the one hand and rainfall on the other. However, we control for livestock management variables in our micro-level analysis in section 4.5.2.

the wake of a bad harvest and others taking advantage of an increase in supply. This could potentially explain the puzzle found in the literature that, on average, there is no relationship between revenues from net sales and transitory income shocks. However, this explanation seems unlikely: first, in any given year, we find that the number of animals sold exceeds on average the number of animals purchased by a factor of two (see table 4.2). Second, we also estimate the absolute rainfall elasticity of purchases. Results are reported in columns (3) and (4) of table 4.3 and suggest that purchases do not vary significantly with rainfall. This suggests that increased sales are not absorbed within provinces through concomitant increases in purchases through rural households covered in our sample. A plausible explanation for this is that livestock is sold to butchers in urban localities.

4.5.2 Count data models

Having examined the relationship between rainfall and sales at the province-level, we now turn to the relationship at the household-level. Since sales are nonnegative integers, count data models are appropriate.[14] We opt for the fixed effects Poisson (FEP)-estimator [Hausman et al., 1984] which has several advantages[15] over commonly used alternatives.[16]

The mean function is specified as

$$m(\mathbf{x}_{it}, \beta) = c_i e^{\mathbf{x}'_{it}\beta}, \tag{4.2}$$

where c_i is a multiplicative fixed effect, $\mathbf{x}$, the matrix of covariates, includes a constant, and β is the vector of parameters of interest. Note that (4.2) has the advantage that parameters are easily interpreted as elastiticies if regressors are included in logs [Wooldridge, 2002, pp.647–648]. If they are included in levels, multiplying the coefficient by one hundred yields the semi-elasticity.

One drawback of the FEP estimator is that households for which the number of sales in all time periods is zero are not used in the estimation procedure.[17] The subsample to which the analysis applies is thus the set of households for which positive sales are observed at least once. This reduces the number of household-year observations available for estimation, particularly in the case of cattle as only about 44 percent of households actually sold cattle at least once. The share of households selling small livestock at least once, in contrast, is more than four-fifths. There are important differences between the two groups. Based on t-tests (not reported), we find that cattle-selling households are older, more likely to be headed by males, have more members, and are more likely to be residents of Burkina's Sahel region. The differences in average herd sizes is substantial: cattle-selling households own on average ten heads of cattle more than non-selling households. The picture that emerges for small livestock is very similar except that the

[14] Results presented below remain unaltered if we use two-way fixed effect-ordinary least squares (OLS) estimators with the log of sales as the dependent variable instead of count data models.

[15] Inference in standard Poisson models relies on the *Poisson variance assumption* that states that the conditional mean must equal the conditional variance [Wooldridge, 2002, pp. 646–647]. While there is evidence for overdispersion in our data—the standard deviation of sales of cattle, sheep, and goats is typically about three times the mean—Wooldridge [1999] shows that the only assumption required for consistency and asymptotic normality of the FEP estimator is that the conditional mean be correctly specified. In particular, the distribution of the dependent variable conditional on covariates and the fixed effects is entirely unrestricted; there can be overdispersion (or underdispersion) in the latent variable model.

[16] Allison [2000] and Greene [2005] show that the commonly used fixed effects-variant of the negative binomial model is not a "true" fixed effects-model as it builds the fixed effect into the variance of the random variable, not the mean.

[17] The FEP estimator is based on quasi-conditional maximum likelihood methods. The sum of counts across time is conditioned on in order to remove the unobserved c_is.

Table 4.4: Results from conditional (fixed effects) Poisson models for the number of sales of cattle, sheep, and goat, 2005–2007.

	Cattle			Goats and sheep		
	(1)	(2)	(3)	(4)	(5)	(6)
Log rainfall	-0.73***	-0.69**	-0.73***	-0.21	-0.45**	-0.25
	(0.26)	(0.27)	(0.25)	(0.27)	(0.23)	(0.26)
Log area cultivated in $t-1$			0.07			-0.04
			(0.08)			(0.06)
# of animals owned in $t-1$			0.00			0.00
			(0.00)			(0.00)
# of animals purchased		0.02*	0.02		0.01***	0.02**
		(0.01)	(0.01)		(0.00)	(0.01)
# of animals born		0.00	0.00		0.03***	0.03***
		(0.00)	(0.00)		(0.00)	(0.00)
# of animals deceased		0.01	0.00		0.00	-0.01*
		(0.01)	(0.01)		(0.00)	(0.00)
# of obs.	4,172	4,172	2,808	7,578	7,578	5,251
# of groups	1,044	1,044	937	1,902	1,902	1,759

Robust standard errors clustered at the province-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. All regressions include year-fixed effects and further controls: the age of the household head and age squared, the gender of the head, and the number of households members in a total of eight gender-age cells (the number of males and females below the age of seven, between seven and 14, between 15 and 64, and 65 and above). Based on EPA data.

proportion of households that has never sold small livestock is only about 20 percent. This is an important issue to keep in mind when comparing results from this section to those in later sections. In a robustness check in section 4.6, we deal with this issue by adjusting samples so that they match samples available for estimation in the present section.

Results from our main regressions are reported in columns (1) and (4) of table 4.4 for cattle and goats and sheep, respectively. In addition to log rainfall, the main variable of interest, we include year-fixed effects in order to control for aggregate shocks to demand and supply conditions. In order to also control for the availability of family labor, we include a set of household demographic variables: the age of the household head, her age squared, her gender, and the number of family members in a total of eight gender-age cells.

Other motives besides consumption smoothing might play a role in the decision to sell livestock [Moll, 2005]. In particular, households may make adjustments by selling livestock in order to maintain the optimal herd size. Regressions reported in columns (2) and (5) therefore include the number of animals purchased, born, and deceased over the last year. As a robustness check, models reported in columns (3) and (6) also include the number of animals owned and the log of the area cultivated in the previous period. Note that including these variables further reduces the number of observations available for estimation as observations on sales in the initial year can no longer be included.

Results reported in table 4.4 indicate that cattle sales are responsive to rainfall with an elasticity of about -0.7 (column (1)). The coefficient is significant at the one percent-level. The estimate is robust to the inclusion of herd management variables (column (2)) and remains unaltered if we include lagged stocks and lagged area cultivated (column (3)). Recall that, on average, households in our data sell about 0.5 heads of cattle each year (see table 4.2) and that in affected provinces the shortfall in precipitation in 2004 relative to the long-run mean was about 30 percent. An elasticity of -0.7 would thus suggest that households facing such a shortfall would step up sales by about one-tenth of a cow.

The coefficient on log rainfall in the regression of sales of goats and sheep is negative and significant at the five percent-level only when livestock management variables are included. The elasticity is also lower in absolute terms throughout: from column (5), a ten percent-increase in rainfall is associated with a decrease in sales by about 4.5 percent. Since an average household sells about 2.5 animals each year, a 30 percent-decrease in rainfall would be associated with an increase in sales by one-third of a goat or sheep. The estimated coefficient turns insignificant and is somewhat closer to zero if we include lagged stocks of sheep and goats, where our estimation sample now includes only 5,251 household-year observations rather than 7,578 as before.[18]

The number of animals purchased, born, and deceased, as well as the number of animals owned in the previous period are included in order to control for herd management considerations. Our results indicate that the number of animals purchased is positively associated with the number of animals sold for both categories of livestock. While all other coefficients are insignificant for cattle, we find that the number of animals born is an important determinant the number of sales for small livestock.

Lastly, recent contributions argue that consumption smoothing is only pursued by a subset of households with high levels of liquid wealth [Carter and Lybbert, 2012, Zimmerman and Carter, 2003]. To account for the possiblity that our findings are driven by farmers with large livestock holdings, we re-estimate all models and include interaction terms between log rainfall and binary variables that indicate holdings of more than 15, 20, and 25 cattle equivalents on average over all time periods.[19] We find no evidence for heterogenous effects across households differentiated by livestock holdings (results not reported); all coefficients on the interaction terms are insignificantly different from zero.

Overall, results in this section suggest that the first question can be answered affirmatively: households increase cattle sales in times of economic hardship. The evidence for the use of goats and sheep as a buffer stock is less robust but goes into the same direction.

4.6 Saving out of transitory income

We now investigate by which means farm households absorb adverse transitory income shocks, the second question outlined in the introduction. We start by motivating the empirical model. The PNGT data allow us to investigate the relationships between transitory income and consumption expenditure directly. Using the EPA data, we then consider saving in the form of grain stocks and livestock.

[18] A regression without these two variables but using only the the smaller sample excluding observations in 2004 reveals that this is *not* due to the inclusion of lagged stocks and area cultivated.

[19] Depending on the specification of their threshold models, Carter and Lybbert [2012] find that for the subgroup of households that own more than 15 and 25 cattle equivalents, livestock sales compensate for a large portion of shocks to transitory income.

4.6.1 Empirical framework

The empirical model is

$$s_{it} = \alpha + \beta y_{it}^P + \gamma y_{it}^T + \delta \sigma_i^y + \nu_{it}, \tag{4.3}$$

where s_{it} denotes savings of household i in period t in the form of some stock (i.e. net purchases of livestock or the accumulation of grain stock), y_{it}^P and y_{it}^T are the permanent and transitory components of total income y_{it}, respectively, and σ_i^y is the variance of the household's income.

As noted by Paxson [1992], a savings equation that is linear in permanent income, transitory income, and the variance of income such as (4.3) can be obtained by maximizing a utility function that is strongly inter-temporally separable and has either quadratic or constant absolute-risk-aversion-form. A linear specification also has the advantage that the coefficients have an easy interpretation: β and γ denote the propensity to save out of permanent and transitory income, respectively: an increase in transitory crop income by one Franc de la Communauté Financière d'Afrique (CFA) is associated with an increase in savings by γ CFA. While we remain agnostic about the degree of saving out of permanent income, we are interested in obtaining an estimate of γ, the propensity to save in different forms out of transitory income.

The challenge is, of course, that both y_{it}^P and y_{it}^T are unobserved in practice. However, there are several ways in which γ might still be identified. As is common in the literature [Carter and Lybbert, 2012, Fafchamps et al., 1998, Kazianga and Udry, 2006, Paxson, 1992], we rely here on unanticipated variation in the level of rainfall in order to isolate the component of rainfall that is orthogonal to permanent income.

First, write $y_{it}^T = y_{it} - y_{it}^P$ such that

$$s_{it} = \alpha + \gamma y_{it} + (\beta - \gamma) y_{it}^P + \delta \sigma_i^y + \nu_{it}. \tag{4.4}$$

De-meaning this equation allows us to purge $\delta \sigma_i^y$. Write

$$\tilde{s}_{it} = \gamma \tilde{y}_{it} + (\beta - \gamma) \tilde{y}_{it}^P + \tilde{\nu}_{it}, \tag{4.5}$$

where the tilde denotes de-meaned variables. This is of course equivalent to introducing a set of household-fixed effects. Equation (4.5) relies solely on variation across time for identification.

Note that if permanent income were (close to) constant over time, an assumption that seems defendable in a setting where there is little technological progress [Deaton, 1992], we would also have purged permanent income from the equation just by the virtue of allowing for household-fixed effects.[20] If, however, permanent income is changing, IV techniques can be applied in order to estimate γ consistently. In practice, instrumenting also serves to safeguard estimates from attenuation bias due to measurement error. We will return to this issue below.

Allowing $(\beta - \gamma)\tilde{y}_{it}^P$ to be absorbed into the error term, γ can be estimated provided a suitable instrument is available that is correlated with changes in transitory income yet uncorrelated to changes in permanent income. Rainfall levels, conditional on household-fixed effects, are both relevant in the first stage and exogenous in the second stage.

[20] In his work on consumption smoothing and saving in Côte d'Ivoire, Deaton [1992] assumes that incomes follow a stationary process. He cites very little real economic growth in rural areas in decades prior to his study in justification of that assumption, an argument that arguably also applies to Burkina Faso in the mid-2000s.

First, rainfall has been shown to be an excellent predictor of farm profits in the WASAT region. For instance, Carter [1997] shows that about half of the variation in crop income in the ICRISAT data is accounted for by rainfall variability. The key assumption is that rainfall conditional on controls and household-fixed effects has no effect on savings other than through its effect on crop income. There are two particular circumstances in which this assumption is violated that are tested routinely in the literature [e.g. Fafchamps et al., 1998, Paxson, 1992]. First, if there was a common trend in rainfall over time, it would seem likely that permanent income would also be trending. Rainfall would thus be correlated with the error term which includes permanent income—see equation (4.5). While this could easily be remedied by considering only rainfall conditional on households-fixed effects *and* year-fixed effects—something that we will do below—we can also test for trends in rainfall data collected at eight rainfall stations across Burkina Faso that stretch back to the early 1970s. Results are reported in Appendix 4.B. Since we find no evidence for linear trends in these data, we conclude that including a common time trend is not necessarily warranted. This result is in line with Fafchamps et al. [1998] who find no evidence for a trend in their rainfall data.

Second, if rainfall were serially correlated, current deviations from long-term means would contain information on deviations in the future. If the AR(1)-parameter was positive and households were aware of this, they would reason that the likelihood of a bad rainfall-year increases following a bad year. This could lead them to hold on to buffer stocks.[21] Also in appendix 4.B, we show that there is no evidence for serial correlation in rainfall.

Our empirical strategy shares key ideas with approaches found in other studies in the literature but there are also some important differences. Fafchamps et al. [1998], Kazianga and Udry [2006], and Carter and Lybbert [2012] rely on an empirical strategy originally advanced by Paxson [1992] that proceeds in two steps. First, a regression model for crop income is specified. This regression typically includes household and farm characteristics, as well as rainfall and interactions of rainfall with farm characteristics on the right hand-side. In addition, this regression typically includes household-fixed effects and, in some cases, village-year-fixed effects [e.g. Carter and Lybbert, 2012]. Second, crop income is decomposed into its permanent, transitory, and unexplained component based on the resulting estimates: household-fixed effects and household- and farm-characteristics multiplied with the respective estimates account for permanent income, while transitory income is determined by rainfall and its interactions and, if included, village-year-fixed effects. Finally, the residuals are taken to be unexplained income. Predicted income components are then used on the right hand-side of a regression of saving together with household-fixed effects and a set of controls. The functional form is similar to the one we start with in equation (4.3) in that it relates savings to permanent and transitory income in levels. The difference is that income variability does not appear on the right hand-side and that, instead, unexplained income is also included.

The first step in this strategy amounts to estimating a first stage-equation in an IV framework manually. Our approach is very similar in terms of the main idea, the reliance on rainfall as an instrument for income in order to identify the effect of transitory income changes. In particular, the assumption that rainfall conditional on household-fixed effects is both unrelated to permanent income and the error term in the second stage is crucial in both frameworks. There are, however,

[21] Deaton [1990] shows that serial correlation in the income-generating process will decrease the viability and desirability of precautionary saving.

Table 4.5: Estimates of the effect of transitory crop income on consumption expenditure (both 1,000 CFA).

	OLS		IV		Reduced form	
	(1)	(2)	(3)	(4)	(5)	(6)
Crop income (1,000 CFA)	0.09**	0.09**	0.78**	0.84		
	(0.04)	(0.04)	(0.34)	(0.64)		
Precipitation (mm)					0.13**	-0.18
					(0.05)	(0.13)
Year 2005		-37.43***		2.86		-83.95**
		(13.06)		(39.86)		(34.51)
Cragg-Donald F statistic (weak identification test).						
F-statistic			11.91	1.93		
# of obs.	2,946	2,946	2,922	2,922	2,972	2,972
# of groups	1,485	1,485	1,461	1,461	1,486	1,486

Robust standard errors clustered at the village-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. Consumption expenditure and crop income measured in 1,000 CFA. All regression include a complete set of household-fixed effects and additional regressors: age of the household head and age squared, the gender of the head, and the number of households members in a total of eight gender-age cells. Based on PNGT data.

three advantages of our framework: first, there is no need for us to adjust standard errors in the second stage. Carter and Lybbert [2012], for instance, bootstrap the two steps outlined above in order to account for the fact that the regressors in the second stage depend on estimated quantities. Second, it is unclear how to interpret coefficients on unexplained income. Finally, we can directly conduct tests of over-identifying restrictions provided that more than one instrument is available for transitory income. We return to this this in section 4.7.2.

4.6.2 Consumption

We first investigate whether households adjust consumption in response to shocks to transitory income. Table 4.5 reports results from regressing consumption expenditure on crop income, where both variables are in real terms and the latter is instrumented using rainfall levels. All models reported in this section include a full set of household-fixed effects and the same set of household demographic variables as in section 4.5. Standard errors are clustered at the level of villages and reported in parentheses.

Columns (1) and (2) of table 4.5 report results from simple OLS-estimation without and with a year-2005-dummy, respectively. Since only two years are available from the PNGT dataset, this is equivalent to running the regression in first-differences. The estimates of the effect of transitory crop income in columns (1) and (2) both suggest that for each increase in transitory per capita crop income by 1,000 CFA, consumption per capita increases by about 90 CFA. While these coefficients are significantly different from zero, they are much lower than comparable estimates in the literature. Kazianga and Udry [2006], for instance, report estimates in the range of 0.50–0.75.

IV estimates based on *two stage-least squares* (2SLS) are presented in columns (3) and (4). We also report Cragg-Donald-F-statistics [Cragg and Donald, 1993] which can be compared to critical values provided by Stock and Yogo [2005]. Our instrument passes the weak identification-test only in the specification that does not include a year-2005-effects.[22]

[22] The critical value for an IV bias relative to the bias in OLS of at most ten percent is 7.03 in this case.

These estimates are greater than OLS estimates by an order of magnitude. The point estimate in column (3) is at the upper end of the range reported by Kazianga and Udry [2006]. It suggests that more than three-fourths of transitory income is transmitted to consumption. However, since we only have two consecutive years of data, standard errors on these coefficients are comparatively large. Based on OLS estimates, one could get the (false) impression that households achieve a high degree of consumption smoothing. As mentioned above, we believe that the difference is due to measurement error in the main explanatory variable, crop income, a problem that is often compounded when identification relies solely on within unit-variation. Similar discrepancies between OLS and 2SLS estimates with income as the main explanatory variable in a fixed effects-specification have been encountered recently by Bengtson [2010]. The problem has also been discussed in the literature on demand for calories [see Deaton, 1997].

At the same time, the standard errors on these coefficients are also substantially larger. In fact, while the estimate reported in column (4) is of similar magnitude, we cannot reject that the coefficient is zero. The finding is not surprising considering the pattern of rainfall during harvests prior to the PNGT surveys (i.e. figure 4.3). Rainfall varies only at the province-level and over time. Considering only rainfall in 2003 and 2004, i.e. rainfall that drives crop income reported by PNGT households during the lean seasons of 2004 and 2005, slightly more than half of the variation in rainfall is accounted for by province-fixed effects. However, if we also include year-fixed effects, roughly 95 percent of the variation in rainfall is captured. Thus, our instrument lacks predictive power when both sets of fixed effects are included. It is important to note that this is not so much of a problem when we analyze EPA data as year-on-year changes in rainfall are much less uniform during later years. Province- and year-fixed effects explain only about 80 percent of the variation in rainfall if we consider the years 2004, 2005, and 2006. We also report results from estimating the reduced forms without and with the year-2005-effect in columns (5) and (6), respectively. Consistent with collinearity between rainfall and the year-fixed effect, the positive and statistically significant effect of our instrument on consumption expenditure vanishes if we include a year-2005-effect.

Despite these shortcomings of the PNGT data, both the descriptive evidence presented in section 4.4 and our regression results here suggest that households reacted to a drop in rainfall levels by cutting consumption. In particular, the coefficient in column (3) implies a very sizeable effect of transitory crop income on consumption expenditure. Albeit insignificant for the reason stated above, the coefficient in column (4) is of similar magnitude.

4.6.3 Grain stocks

Next, we investigate the importance of savings in the form of grain stocks in *ex-post* consumption smoothing. This is done by regressing subsequent changes in grain stocks (i.e. forward first-differences), valued in real CFA, on crop income and household-fixed effects and instrumenting crop income again with rainfall levels. Hence, of the four years of data from 2004 to 2007 in the EPA surveys, the last contributes only one observation on grain stock levels required to construct the first-differenced dependent variable associated with crop income in 2006. Results are reported in table 4.6. Again, all regressions include additional control variables (not reported) that capture households' demographic make-up. In this case, we also include year-fixed effects in all regressions.

Table 4.6: Estimates of the effect of transitory crop income on subsequent changes in grain stocks (both 1,000 CFA).

	OLS (1)	IV (2)	Red. (3)
Precipitation (mm)			0.08^{***}
			(0.03)
Crop income (1,000 CFA)	0.02^{**}	0.26^{**}	
	(0.01)	(0.12)	
Year 2005	38.66^{***}	18.10	36.22^{***}
	(3.38)	(11.57)	(4.11)
Year 2006	28.02^{***}	8.70	28.36^{***}
	(3.47)	(10.18)	(3.27)
Cragg-Donald F statistic (weak identification test).			
F-statistic		20.62	
# of obs.	7,071	7,071	7,092
# of households	2,357	2,357	2,364

Robust standard errors clustered at the village-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. Changes in grain stock and crop income measured in 1,000 real CFA. All regressions include a complete set of household-fixed effects and additional regressors: age of the household head and age squared, the gender of the head, and the number of households members in a total of eight gender-age cells. Based on EPA data.

Before considering results from OLS and IV estimations in columns (1) and (2), respectively, note that in contrast to our findings for consumption, the reduced form-estimate indicates that rainfall predicts changes in grain stocks conditional on year-fixed effects (column (3)). The coefficient is positive and significantly different from zero at the one percent-level despite our inclusion of year-fixed effects. Since these results are based on EPA data, more time periods are available and the number of households observed in each year is larger. As a result, the Cragg-Donald F-statistic reported in column (2) of table 4.6 indicates that the partial correlation between crop income and our instrument is sufficiently high.[23]

OLS and IV results are reported in columns (1) and (2), respectively. Again, the difference is large: while both coefficients are significant at least at the five percent-level, the IV-estimate is larger by an order of magnitude. As noted above, this is likely due to attenuation bias that is a result of measurement error in the independent variable. The IV-estimate suggests that grain storage plays an important role in *ex post*-consumption smoothing: households absorb approximately one-fourth of transitory crop income by adjusting grain stocks. This is in line with findings reported in Kazianga and Udry [2006] for Burkina Faso during the early 1980s and Udry [1995] for northern Nigeria.

4.6.4 Livestock

We now turn to savings in the form of livestock by regressing net purchases of livestock on crop income. The empirical set-up is the same as in the 4.6.2 and 4.6.3. The first three columns of table 4.7 report results from specifications that mirror those in table 4.6. The coefficient in the OLS regression reported in column (1) is statistically significant yet close to zero. Again,

[23] The test statistic exceeds the critical value, 16.38, reported by Stock and Yogo [2005] that corresponds to a bias in the IV estimate relative to the OLS estimate of ten percent.

TABLE 4.7: Estimates of the effect of transitory crop income on subsequent net purchases of livestock (both 1,000 CFA).

	OLS	IV	Red. form	Owners	Sellers	Sellers (cattle only)
	(1)	(2)	(3)	(4)	(5)	(6)
Precipitation (mm)			0.06*			
			(0.03)			
Crop income (1,000 CFA)	0.02***	0.19		0.18	0.21	0.42
	(0.01)	(0.12)		(0.12)	(0.13)	(0.30)
Year 2005	-7.17	-21.33**	-7.74*	-22.14**	-27.31**	-66.52
	(4.88)	(10.22)	(4.57)	(10.60)	(12.67)	(41.75)
Year 2006	-12.10**	-25.24**	-10.99**	-26.14**	-32.91**	-64.27**
	(5.20)	(10.94)	(5.09)	(11.19)	(13.36)	(29.33)
Cragg-Donald F statistic (weak identification test).						
F-statistic		20.00		19.85	19.37	6.26
# of obs.	7,027	7,025	7,048	6,650	5,561	2,783
# of households	2,357	2,355	2,364	2,230	1,865	935

Robust standard errors clustered at the village-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. Net purchases of livestock and crop income measured in 1,000 real CFA. All regressions include a complete set of household-fixed effects and additional regressors: age of the household head and age squared, the gender of the head, and the number of households members in a total of eight gender-age cells. Based on EPA data.

measurement error is suspected. The IV estimate in column (2) is larger by a factor of about ten yet insignificant at conventional levels of significance. It suggests that 20 percent of transitory crop income is saved in the form of livestock. Finally, the reduced form coefficient is significant only at the ten percent-level, suggesting a weak partial correlation between rainfall and net purchases of livestock. Taken together, there is little evidence of significant savings out of transitory crop income in the form of livestock.

These findings are despite the fact that most households' holdings of livestock would have allowed them to completely absorb the income shock caused by adverse weather conditions. If we define the shock as the negative deviation in crop income from its four-year-mean between 2004 and 2007 and compare this for 2004 to livestock holdings at the end of the lean season in 2005, we find that in each region more than half of the households disposed of enough livestock to compensate for the entire shortfall. In seven out of the 13 regions and including in the Sahel more than 80 percent of the households had sufficient livestock holdings.

Our findings here are in line with the literature. In particular, Fafchamps et al. [1998] find that at most 30 percent and probably closer to 15 percent of income shortfalls are compensated via livestock sales. The latter is close to the point estimate reported in column (2). While we cannot reject the hypothesis that the coefficient on crop income in column (2) is equal to 30 percent, it is also insignificantly different from zero. Note that this is despite the fact that we find evidence for the use of grain stocks as a buffer stock in the same data. Hence, a lack of statistical power is unlikely to account for our non-finding.

The findings from savings regressions differ substantially from those obtained considering only the number of sales in section 4.5.2. As noted above, however, one concern is comparability:

since households that are never observed selling livestock do not contribute to the conditional log likelihood in the case of the FEP estimator, results in that section were based on subsamples that are potentially selective. We therefore also investigate savings behavior for the subset of households that owned either cattle, sheep, or goats at any point after 2004 (column (4)); those that report positive sales in any of these categories between 2005 and 2007 (column (5)); and those that report postive sales of cattle over this period (column (6)). The last subsample corresponds closely to the subsample used in column (2) of table 4.4. While all three estimates of the propensity to save in livestock are positive and have the expected sign, they are insignificant at conventional levels.

Taken together, the above regressions show that cuts to consumption and adjustments to grain stock go a long way in explaining how households absorb transitory crop income. For instance, if we would combine our estimates in columns (2) and (4) of tables 4.5 and 4.6, respectively, we would already be able to account for all of the change in transitory income.

4.6.5 Alternative income sources

At least some households might have had the opportunity to resort to other sources of income in order to compensate for output loss due to adverse rainfall. Transfers (including in-kind transfers, remittances, and aid), revenue from non-agricultural businesses, wages from off-farm employment, and the use of credit might play a role in households' risk management.[24]

We investigate this issue further based on the PNGT data which provide information on (gross) revenues from households' non-cropping enterprises, net transfers, wages earned, and use of credit. Crop income accounts for more than 50 percent of the total in eleven out of 13 provinces. Only in the Centre-Nord region and in the Sahel is the share smaller. To investigate whether alternative sources of income become important in case of an adverse shock to crop output, we run regressions of these alternative income sources on crop income instrumenting with rainfall and controlling for households characteristics as before. Results (not reported) indicate that these alternative sources of income do not allow households to smooth consumption *ex post*. Coefficients on crop income are either positive or insignificant suggesting that income from these alternative sources will decline if crop income does.[25]

Overall, the results in this section suggest that we have to negate the second question spelled out in the introduction, i.e. revenues from livestock sales do not make-up for losses in crop income during droughts. Instead, there is some evidence that transitory income shocks are passed on to consumption expenditure and stronger evidence for the use of grain stocks as a buffer.

[24] Reardon et al. [1988], for instance, show that the share of food aid accounted for 60 percent of transfers received by the poorest households in the Sahelian region of Burkina during the 1984 drought. Reardon et al. [1992] argue that non-farm activities of households in the same data were an important means of *ex ante* income diversification accounting for 30–40 percent of total income. A more recent study by Lay et al. [2009] that investigates patterns of income diversification in Burkina Faso between 1994 and 2003 concludes that the extent of income diversification stagnated.

[25] As noted by Fafchamps et al. [1998], the finding is also consistent with anecdotal evidence reported in Sen [1981] who argues that droughts often lead to a collapse in the demand for local services and crafts.

4.7 Rainfall, prices, and quantities

4.7.1 Evidence from province-level price regressions

In essence, our results above are in line with findings reported in the two different strands of the literature and make explicit the contradictory conlusions: we have first shown that households increase livestock sales following a negative rainfall shock. Directly asked about the reason for sales, households cite the need to finance food purchases. At the same time, we find no evidence for consumption smoothing via asset sales. The logical next step is to explain this apparent contradiction. In this section we probe into one plausible explanation: price adjustments in the wake of adverse weather shocks.

If prices for livestock decline in response to a rainfall-induced increase in market supply, the effect of rainfall on net purchases in monetary terms as investigated in 4.6.4, will be attenuated. This would require that the demand schedule for livestock that households face in a given locality is downward sloping and it seems likely that markets in Burkina Faso match this description. For instance, there is some evidence that in Niger, a country bordering Burkina Faso to the Northeast, markets for livestock are poorly integrated [Fafchamps and Gavian, 1997].[26]

To examine whether such an explanation for the puzzle is plausible, we investigate how prices for livestock react to changes in rainfall. We do so by regressing log prices for cattle, sheep, and goat on log rainfall. The resulting coefficient can thus be interpreted as the rainfall elasticity of livestock prices.

The set-up is the same as in (4.1) only that prices are now on the left hand-side of the equation. Again, we include both province- and year-fixed effects in our regressions. The former account for province-specific differences in market structures that affect prices and are potentially correlated with levels of precipitation. The latter account for common shifts in demand and supply of livestock. Prices are unit values calculated from the EPA data and then averaged within each province.[27] The precision of these averages will depend on the number of sales reported. Hence, there is an econometric argument for weighting each province-year observation in the resulting panel dataset in proportion to the number of observations for which unit values could be calculated. However, this would give a higher weight to provinces in which many sales are reported, i.e. in which markets are well-functioning, potentially biasing our results towards a lower price response. Running both weighted and unweighted regressions, we find that the differences between the estimated elasticities are only minor. Therefore, we only report the former.

Results are reported in table 4.8, where we consider both nominal (columns (1)–(3)) and real prices, i.e. prices divided by our CPI discussed in section 4.3 (columns (4)–(6)). The estimates reported are positive and statistically significantly different from zero for cattle but not for other types of livestock. The elasticity of the nominal cattle price is 0.3 and is significantly different from zero at the five percent-level. Estimated elasticities are also positive but lower for sheep and goat at 0.15 and 0.10, respectively. In both cases, however, we cannot reject that they are zero at conventional significance levels. This is consistent with small livestock being easier to transport. Estimates are very similar when real prices are considered (columns (4)–(6)).

[26] Fafchamps and Gavian [1997] nevertheless find that relative prices respond to changes in urban meat demand, signalling at least some degree of integration.

[27] Households sampled in the EPA surveys were asked to report on quantities and values of livestock sold within the last twelve months.

Table 4.8: Results from province-level fixed effects-regressions of log nominal and log real prices for livestock on log rainfall, 2004–2007.

	Nominal price			Real price		
Log price of...	...cattle. (1)	...sheep. (2)	...goat. (3)	...cattle. (4)	...sheep. (5)	...goat. (6)
Log rainfall	0.30**	0.15	0.10	0.28**	0.13	0.08
	(0.13)	(0.13)	(0.11)	(0.13)	(0.12)	(0.11)
Obs.	177	177	177	177	177	177
R-Squared	0.75	0.88	0.77	0.77	0.89	0.83

Robust standard errors in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. All regressions include year- and province-fixed effects. Based on EPA data.

Our results are in line with Fafchamps and Gavian [1997] who find that livestock prices respond to droughts in Niger. Several authors have also commented on the potential importance of general equilibrium effects in the context of consumption smoothing more broadly [Fafchamps et al., 1998, Zimmerman and Carter, 2003]. For instance, Fafchamps et al. [1998] point out that in the extreme case in which villages constitute closed markets, net sales of livestock will necessarily total zero and that prices will adjust downward accordingly. However, in section 4.5 we found no evidence for a positive elasticity between rainfall and purchases, suggesting that livestock was sold to economic agents outside our sample of rural farmers.

Results presented in this section explain the puzzling finding in the literature of no consumption smoothing via sales of livestock. In particular, two effects seem to be at work during drought episodes that to some extent have a tendency to cancel each other out. First, households try to make-up for a drop in crop income by selling extra livestock. Second, this leads to a downward adjustment of livestock prices and, hence, to a decrease in what can be earned per unit of livestock sold. The result is that revenues remain unaltered despite increased sales.

4.7.2 Rainfall, prices, and exclusion restrictions

While the above results explain the apparent lack of association between rainfall and net purchases, it also potentially threatens the appropriateness of rainfall as an instrument for crop income in a regression of savings on income as in section 4.6. Such specifications derive from partial-equilibrium models in which prices are exogenous. If, however, rainfall affects prices and, at the same time, local prices are important for households' decision in which form to make provisions for the future, rainfall is potentially correlated to the error term in a specification such as (4.5).

In appendix 4.C, we therefore test underlying exclusion restrictions in two ways: first, we insert prices for cattle and the CPI directly into the estimation equation and test whether they are individually and/or jointly significant. Second, we generate additional instruments and test exclusion restrictions based on standard Hansen/Sargan-type tests. In both cases we cannot reject that our instruments are rightly excluded from the main equation of interest. Thus, we are confident that coefficient estimates are consistently estimated.

4.8 Conclusion

The present paper re-visits a puzzle stated in the empirical literature on optimal saving in developing countries in the absence of formal insurance mechanisms. While livestock holdings were traditionally hypothesized to constitute the main means of households to smooth consumption in the wake of shocks, empirical work in this area usually finds no evidence for a significant relationship between the *monetary value of net livestock sales* and transitory income. On the other hand, studies with a focus on the *number of sales* often find evidence for a sizeable increase in sales in response to adverse shocks.

The event we study is a severe drought in the northern provinces of Burkina Faso that occurred in 2004 and a subsequent return to normal levels of rainfall. Our empirical investigation is based on two household-level datasets that provide ample information on consumption, grain stocks, and transactions of livestock.

Our results can be summarized as follows: rainfall positively affects sales of livestock with no off-setting effect on purchases at the level of provinces. A similar increase in sales in response to adverse rainfall is observed at the household-level. Reportedly, extra sales were a reaction to an increased need to finance food purchases. However, we find no evidence for precautionary savings in the form of livestock. On the other hand, grain storage plays a significant role in *ex-post* coping. There is some evidence that a substantial portion of transitory income is transmitted to consumption expenditure, although our data are in this case insufficient to robustly establish this result.

We then show that cattle prices at the province-level vary positively with rainfall and our estimates suggest that the elasticity is high. This is consistent with a general equilibrium-effect that adversely affects revenues from livestock sales in times of harvest failure, rendering precautionary saving in the form of livestock a costly strategy to smooth consumption. Households thus seem to manage a difficult trade-off between selling more livestock at low prices and destabilizing consumption and safeguarding assets that may fetch higher prices in the future. Consequently, asset-smoothing may be considered the outcome of poor prices to be had in times of crises, an issue that has only insufficiently been accounted for in previous empirical investigations.

Our findings underline the lack of market integration in rural Burkina Faso witnessed by massive price changes and inter-regional discrepancies over the course of the 2004 drought. These imply that savings in forms other than grain stocks are subject to major price risks. An increased focus on integrating livestock markets (e.g. by investing in road infrastructure) would potentially mitigate welfare losses incurred by farm households during episodes of economic distress. Ultimately, of course, appropriate insurance mechanisms should be put in place that would allow households to stabilize incomes *ex ante*.

4.A Imputation of prices

We rely on predicted crop prices obtained from regressions in which prices from the PNGT are fitted based on price data from Ouagadougou. Denote the price for crop c in province p in year t and month j (May or November) p_{pctj} and the contemporaneous price in Ouagadougou $p_{ctj}^{Ouag.}$. The model can then be written

$$\ln(p_{pctj}) = \phi_p \ln(p_{ctj}^{Ouag.}) + \rho_p + \gamma_c + \epsilon_{pctj}, \tag{4.6}$$

where ϕ_p is the province-specific price elasticity with respect to capital city-price, ρ_p is a province-fixed effect, and γ_c is a crop-fixed effect. t indicates two months in each year, May and November. We allow the price-price-elasticity ϕ to vary across provinces as we expect different degrees of integration of local markets. The resulting model has an R^2-statistic of 71.7 percent.

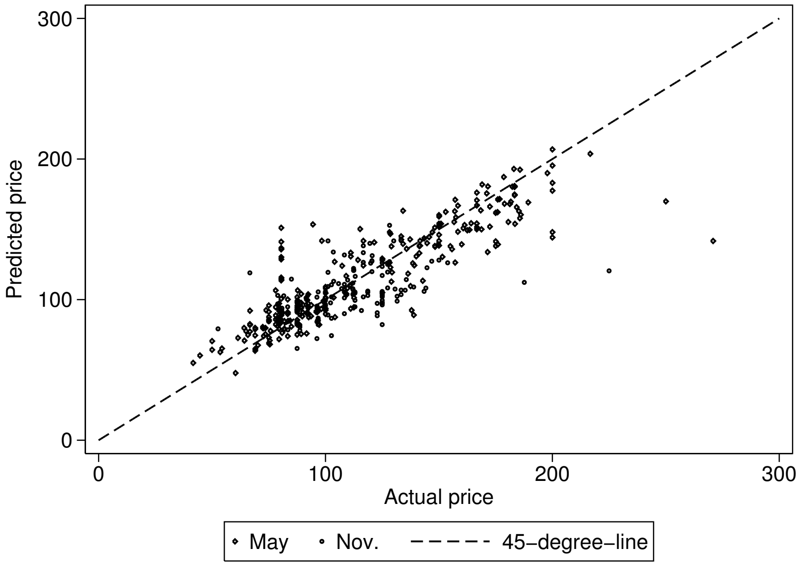

FIGURE 4.10: Predicted vs. actual prices for sorghum, millet, and maize; May 2004, May 2005, and November 2006.

Figure 4.10 plots predicted prices against actual observed prices. While there are some outliers in the sense that the actual price was much higher than the predicted price, the overall fit seems reasonable. Figures 4.11 and 4.12 plot time series of predicted prices and prices in Ouagadougou for each province separately for sorghum and millet, the main staples in Burkina Faso, respectively. Also displayed are the actual province-level price observations from the PNGT data. Regional market prices are added for comparison. As one would expect for locally produced goods, movements of predicted prices closely track price movements in Ouagadougou yet prices are lower and less volatile in the provinces.

In a second step, we regress log prices from the three PNGT datasets on capital city-log prices for sorghum and maize and a set of province-fixed effects for all remaining crops separately. These crops and the respective R^2-statistics are rice (40.3 percent), groundnut (41.3), niébé

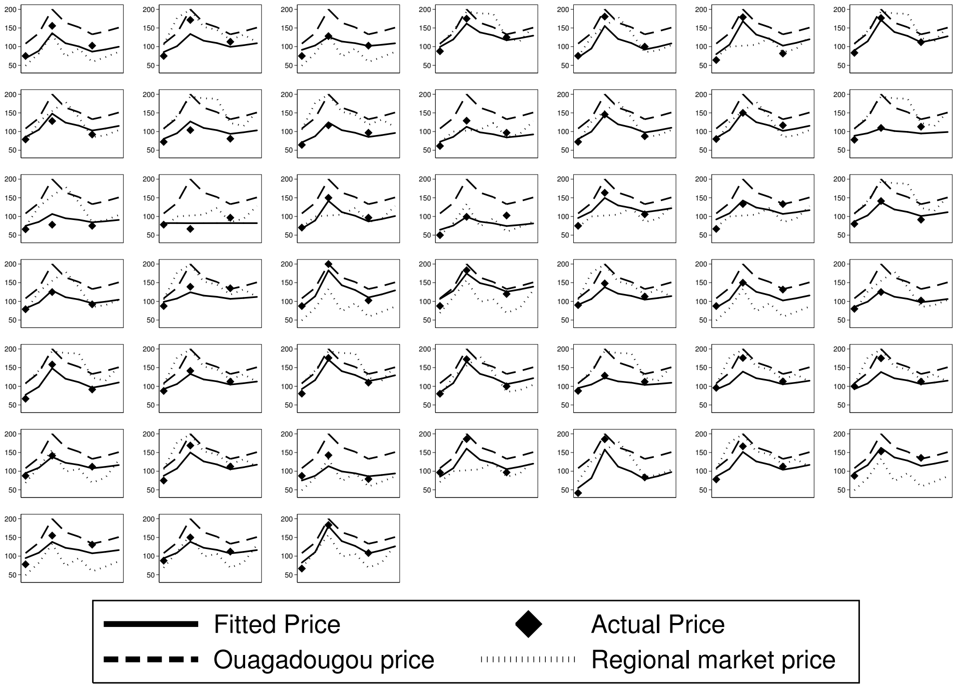

FIGURE 4.11: Actual, imputed, and nearest large city-, and capital city-prices for sorghum, May and November 2004, 2005, 2006, and 2007.

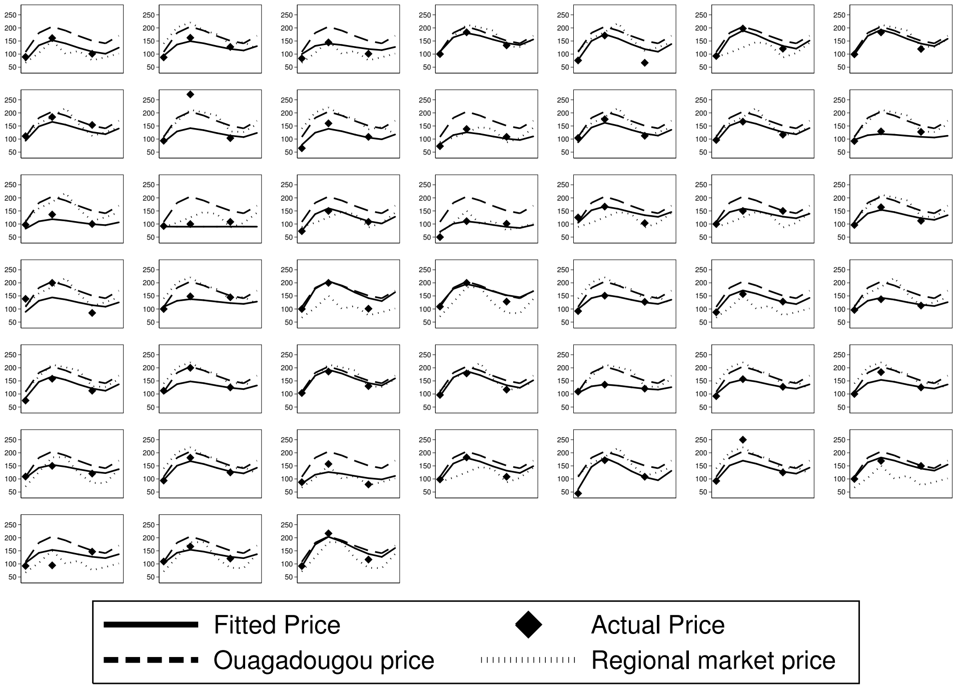

FIGURE 4.12: Actual, imputed, and nearest large city-, and capital city-prices for millet, May and November 2004, 2005, 2006, and 2007.

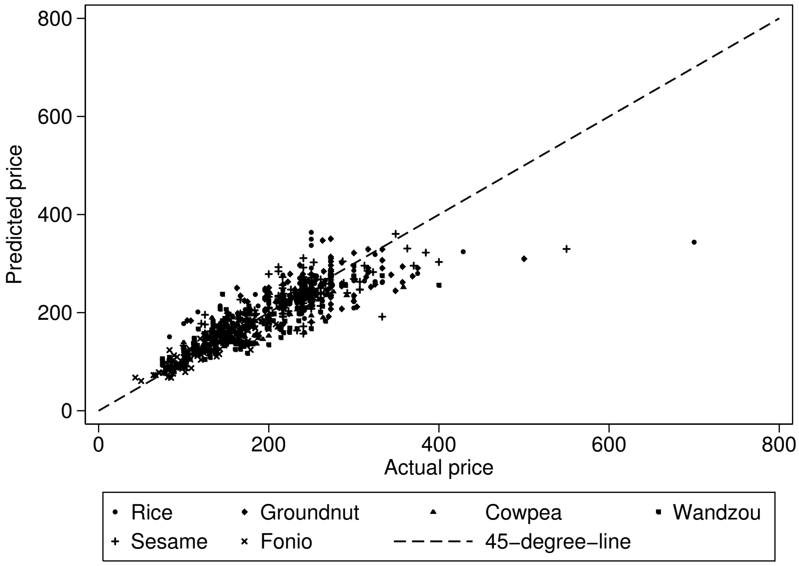

FIGURE 4.13: Predicted vs. actual prices for rice, groundnut, niébé, wandzou, sesame and fonio; May 2004, May 2005, and November 2006.

(57.2), wandzou (55.9), sesame (52.6), and fonio (83.9). Figure 4.13 plots predicted against actual prices.[28]

4.B Levels of rainfall across Burkina Faso, 1970–2009

In this appendix we report results from analyzing time series data from eight rainfall stations across Burkina Faso for years prior to our study period. For the validity of our instrument in the empirical application of this paper, it is crucial that levels of rainfall neither exhibit significant trends over time nor that conditional on the long-term mean past observations provide any information about future rainfall. In that case, deviations of rainfall from its long-term mean will be orthogonal to permanent income; income associated with good rainfall will be transitory [see also Deaton, 1997, p. 290].[29]

The data analyzed here come from FAO's 2014a *Climate Impact on Agriculture*-website and contain information on monthly rainfall. To prepare the series for analysis we first aggregate rainfall at the level of years, retaining only station-year-observations for which observations in each month were available. In a second step, we discard all stations for which we have less than 25 years of observations. The final series are depicted in figure 4.14.

The location and elevation of weather stations is reported in panel A of table 4.9. Given the geographical locations of weather stations which capture much of the agro-climatic differences across Burkina Faso, the data allow us to make statements about rainfall patterns in very different parts of the country. We subject the series to simple tests for linear and exponential

[28] Prices for cotton are fixed as the state is the monopoly buyer of cotton.

[29] There is a long-standing tradition in economics of using rainfall variability in order to identify effects of income components. Wolpin [1982] uses information on historical regional rainfall for rural Indian households assuming that households residing in regions with favorable weather conditions have higher permanent income. Paxson [1992] shows that the deviation of rainfall from its local mean in Thailand is serially uncorrelated and thus unpredictable. It is therefore uncorrelated with permanent income yet a strong predictor of transitory income. More recent examples include Fafchamps et al. [1998], Kazianga and Udry [2006], and Carter and Lybbert [2012].

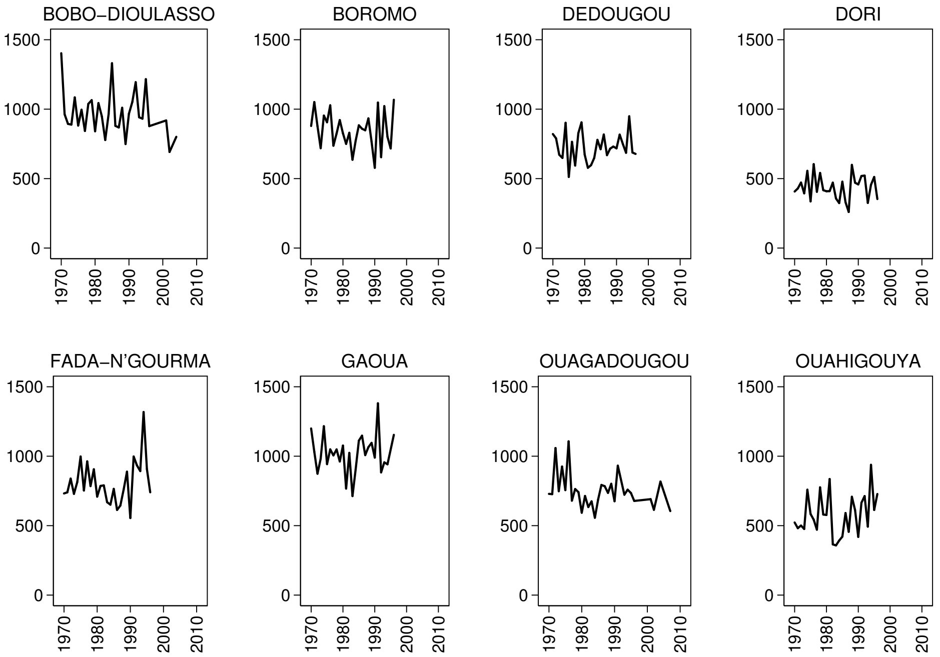

Figure 4.14: Rainfall levels recorded at eight stations across Burkina Faso, 1970–2009. Data from FAO [2014b].

time trends. We regress rainfall and log rainfall on years for each series separately. Results are reported in panel B of table 4.9. There is only one coefficient that is statistically significant at the ten percent-level, namely for the series from Ouagadougou. Results from Breusch-Godfrey-tests [see Breusch, 1979, Godfrey, 1978] in panel C indicate that the null hypothesis of no serial correlation cannot be rejected for any of the eight series. We conclude that there is no evidence that deviations of rainfall from long-term means are predictable. Rainfall levels *conditional on household-fixed effects* thus seem an appropriate instrument.

4.C Testing exclusion restrictions

In section 4.7.1, we show that cattle prices are responsive to rainfall. This finding potentially threatens the identification strategy we pursue in savings regressions presented in section 4.6 where we regress savings in the form of livestock on crop income and instrument the latter with rainfall levels since prices might also affect households' decision to purchase and sell livestock directly. In this appendix we therefore aim to test directly whether our instrument can in fact be excluded from (4.5).

One way to test this is to include the livestock and the price level of other goods, captured by the CPI, on the right hand-side of the equation of interest. Note that we are not suggesting that the resulting point estimates are in some way more valuable in judging whether households rely on livestock to smooth consumption. The hypothetical question how households would react to transitory changes in crop income *conditional on real livestock prices* is arguably not of interest. Here, we are solely interested in whether our identification strategy is valid. A significant coefficient on the price of cattle should be interpreted as a sign that the exclusion

TABLE 4.9: Analysis of station-level rainfall data, 1970–2009.

	Gaoua	Bobo-Dioulasso	Boromo	Fada-N'Gourma	Ouagadougou	Dedougou	Ouahigouya	Dori
Panel A. Location of weather station.								
Latitude	10.33	11.17	11.75	12.03	12.35	12.47	13.57	14.03
Longitude	−3.18	−4.32	−2.93	0.37	−1.52	−3.48	−2.42	−0.03
Elevation (m)	335	460	271	309	306	300	336	277
Panel B. Testing for time trends.								
Coef.: rainfall on year	0.97	−4.21	−2.04	4.06	−3.78*	0.78	4.78	−0.38
	(3.70)	(3.07)	(3.29)	(3.80)	(2.15)	(2.64)	(3.63)	(2.22)
Coef.: log rainfall on year	0.00	−0.00	−0.00	0.00	−0.00*	0.00	0.01	−0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)
Panel C. Breusch-Godfrey Lagrange multiplier test for serial correlation.								
F-value	0.308	0.454	2.539	0.678	0.005	0.894	0.000	2.474
p-value	0.584	0.506	0.124	0.418	0.944	0.354	0.997	0.128

Standard errors in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. Sorted by latitude (rather than alphabetically). Based on data from FAO [2014a].

TABLE 4.10: Tests of over-identifying restrictions: net purchases of livestock in 1,000 CFA.

	2SLS		GMM	
	(1)	(2)	(3)	(4)
Crop income (1,000 CFA)	0.22*	0.27	0.21**	0.04
	(0.13)	(0.20)	(0.10)	(0.03)
Log price of cattle		-34.59		
		(43.42)		
Log CPI		93.94		
		(314.95)		
Year 2005	-18.94*	-28.75	-18.14**	-5.61
	(10.92)	(41.28)	(8.26)	(3.70)
Year 2006	-22.64*	-18.91*	-21.90**	-5.37
	(11.74)	(11.07)	(9.01)	(4.29)
Cragg-Donald F statistic (weak identification test).				
F-statistic	18.62	9.75	14.96	9.86
Hansen/Sargan-test (over-identifying test of all instruments):				
Degrees of overidentification	0	0	2	12
χ^2-statistic	0.00	0.00	1.43	8.73
p-value			0.49	0.73
# of obs.	6,857	6,857	6,857	6,857
# of households	2,353	2,353	2,353	2,353

Robust standard errors clustered at the village-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. Net purchases of livestock and crop income measured in 1,000 real CFA. All regressions include a complete set of household-fixed effects and additional regressors: age of the household head and age squared, the gender of the head, and the number of households members in a total of eight gender-age cells. Based on EPA data.

restriction does not hold. Results are reported in column (2) of table 4.10. For comparison, we also report results from a regression without these prices in column (1). This model corresponds to the one in column (4) of table 4.7 and is reported solely for comparison.[30] We cannot reject that the coefficients on log price of cattle and log CPI in column (2) are both individually and jointly zero.

One might also test exclusion restrictions directly based on Hansen's J-test. However, this requires the model to be over-identified. One way of generating additional over-identifying restrictions is to specify a set of instruments as in Holtz-Eakin et al. [1988], where each time period is instrumented separately. This results in three instruments that convey information about rainfall in 2004, 2005, and 2006, respectively. A second option is to allow rainfall to affect crop income differently across Burkina Faso's 13 regions. Both approaches convey slightly more information. However, it is the second approach that we believe is more appropriate as rainfall likely affects agricultural production differently across Burkina's regions. In both cases, the additional moment conditions are perfectly valid if rainfall itself meets the exclusion restriction.

Under the null, Hansen's J-statistic is χ^2-distributed with degrees of freedom equal to the degrees of over-identification—two and twelve with the instrument sets described above. It is consistent in the presence of heteroskedasticity and serial correlation. The test is a general

[30] The only difference is that we exclude households in three provinces, Boulkièmdé, Tapoa, and Loroum, in one year, 2005, for which there are no reports on livestock sales and hence no lean season prices that we could calculate for 2004/2005. However, the resulting estimate is broadly in line with the one reported in table 4.7.

specification test: if rejected, either the orthogonality conditions or other assumptions of the model or both are likely to be false [see Hayashi, 2000, pp. 198–201 and 217–218, for details]. In any case, a rejection will cast doubt on the appropriateness of the instruments employed. Results from two-step efficient *general method of moments* (GMM)-estimation are reported in columns (3) and (4). We find that we cannot reject that our instruments are jointly valid and thus conclude that rainfall is rightly excluded from the estimation equation.

Chapter 5

Tracking and the Intergenerational Transmission of Education: Evidence from a Natural Experiment

Abstract Proponents of tracking in education argue that the creation of more homogeneous classes increases efficiency while opponents fear that tracking aggravates initial differences between students. We estimate the effects on the intergenerational transmission of education of a reform that delayed tracking by two years in one of Germany's federal states. While the reform had no effect on educational outcomes *on average*, it increased educational attainment among individuals with uneducated parents and decreased attainment among individuals with educated parents. The reform thus lowered the gradient between parental education and own education. The effect is driven entirely by changes in the gradient for males.

5.1 Introduction

Can education policies help level the playing field between students from different social backgrounds? And if so, how? One major design feature of education systems that has frequently been related to equality of opportunity is tracking, the practice of grouping students by ability [Betts, 2010]. Countries differ widely in the way they track students[1] but almost all education systems in practice feature some kind of tracking. Proponents argue that the creation of more homogeneous classes increases efficiency by allowing educators to tailor lessons to students' specific needs. Opponents, on the other hand, fear that misclassification of students is

The chapter is based on joint work with Marten von Werder. It has benefited from comments by Rafael Aigner, Mikkel Barslund, Thushyanthan Baskaran, Ronny Freier, Stephan Klasen, Steffen Lohmann, Malte Reimers, Ramona Rischke, Robert Schwager, Viktor Steiner, Sebastian Vollmer, as well as seminar participants in Berlin and Göttingen and participants of the 2015 Spring Meeting of Young Economists in Ghent.

[1] For instance, countries differ in the age at which students are tracked and in whether tracking occurs within schools (i.e. sorting students into different classrooms as in the United States and Canada) or across schools (i.e. sending students to different types of schools as in some European countries).

often rife—especially when students are tracked at an early age—and that tracking aggravates initial differences.

In this paper we exploit a policy reform implemented during the 1970s in one of Germany's federal states, Lower Saxony. While most German states continued tracking students after fourth grade, the reform shifted the timing of tracking from grade four to grade six (roughly age ten and twelve, respectively). This was achieved through the introduction between 1972 and 1982 of a new intermediate school, the *orientation stage*[2] (henceforth, OS). We investigate the effects of this reform on the gradient between parental education and own education[3] based on a *difference-in-differences* (DD)-framework, comparing changes in this gradient across cohorts and states.

On average, the reform neither increased years of education nor the likelihood of being eligible to apply for university nor university graduation. We find, however, that the reform had a significant negative effect on the gradient in terms of years of education (which we measure as the time usually required to obtain the highest degree earned): years of education among individuals with uneducated parents increased by about one-third of a year and decreased by about the same time for individuals with educated parents. There is no evidence for a corresponding effect for females. Hence, the gap in years of education decreased by 1.4–1.6 years for males and by about 0.7–0.8 years overall. We observe a similar pattern when we investigate the impact on the likelihood of being eligible for university and obtaining a university degree. These results are robust to a wide range of changes in the empirical specification and in sample definitions.

In an extension to our analysis, we find no evidence for a positive effect of the reform on parental involvement, one of its stated aims. We argue that our results are consistent with the presence of peer effects and improvements in the allocation of students to tracks associated with de-tracking. We discuss recent evidence from the literature on gender differences in the development of non-cognitive skills and conclude that such differences potentially explain the gendered pattern in our findings.

The German reform we study offers a particularly well-suited setting to study the effects of tracking: education policies in Germany are to a large extent the responsibility of the states while the federal government is in charge of most other policy areas that may affect educational outcomes. Hence, our analysis is unlikely to be subject to confounding trends in unobserved socio-economic and institutional factors.[4] Because of Germany's system of equalization payments between states and sharing of tax revenues, there is limited potential for different trends in resources allocated to public education between states. Also, private education institutions that would potentially mitigate the effect of policy reforms aimed at improving intergenerational mobility play a negligible role in Germany's education system. Finally, Germany stands out among OECD countries for early tracking at which large effects on educational careers may be expected.

Only a limited number of studies investigate the long-term effects of changes in the tracking regime.[5] Identification strategies that rely on cross-sectional variation [e.g. Bauer and Riphahn, 2006] potentially suffer from bias due to selection-on-unobservables or omitted variables [Betts

[2] In German: *Orientierungsstufe.*

[3] We define this gradient as the gap in educational outcomes between individuals with educated parents and those with uneducated parents.

[4] Waldinger [2007] argues that this may be an issue in cross-country comparisons such as Hanushek and Woessmann [2006]. See, however, Woessmann [2010].

[5] See Betts [2010] for a review.

and Shkolnik, 2000, Pischke and Manning, 2006, Waldinger, 2007]. Hanushek and Woessmann [2006] exploit differences in the degree of tracking across countries based on an identification strategy that replaces changes within countries with changes across grades. They find a positive effect of tracking on performance inequality and no effect on performance levels. Waldinger [2007], however, presents evidence hinting at the presence of an omitted variable that affects both the intergenerational transmission of education and the likelihood of a country to implement early tracking.

Brunello and Checchi [2007] and Schütz et al. [2008] rely on cross-country variation in design features of education systems in order to estimate their impact on the importance of family background characteristics. The former authors focus on long-term outcomes such as earnings and employability and are thus closer to the present study. Both studies find that early tracking accentuates the importance of family background characteristics.

There are also several recent studies that investigate policy reforms within countries. Malamud and Pop-Eleches [2011] exploit a policy reform implemented in Romania in 1973 to study the effect of postponing tracking on educational outcomes. The reform increased significantly the proportion of students in general as opposed to vocational secondary schools. While they find a sharp increase in the proportion of students eligible to apply for university overall and among disadvantaged students, they do not find an increase in the likelihood of disadvantaged students to attend university. Their results are thus in stark contrast to ours.[6]

Studies similar to our own rely on variation across both cohorts and space to analyze de-tracking reforms in Scandinavia. Meghir and Palme [2005], Pekkarinen et al. [2009], and Kerr et al. [2013] all report findings that are broadly in line with ours. Hall [2012], however, finds no effect of such a reform on university enrollment in Sweden. An important difference is that country-wide reforms in Scandinavia involved other significant changes to educational institutions and it is not possible to disentangle the effects of de-tracking from other components of these reforms.[7] Lower Saxony's reform, in contrast, entailed only adjustments to curricula in addition to delayed tracking. Nevertheless, our results below suggest that the pro-equality character of Germany's reform is broadly comparable to that of Scandinavian reforms.

The remainder of this paper is organized as follows. The next section describes the German education system and the process that finally led to the introduction of OS schools in Lower Saxony. Section 5.2 discusses channels through which tracking may affect educational outcomes and educational inequality. Section 5.4 describes the data and our identification strategy. Section 5.5 presents our main results and section 5.6 a number of robustness checks. Section 5.7 investigates one potential transmission channel, parental involvement with their children's academic performance. Section 5.8 offers an interpretation of our results. Section 5.9 concludes.

[6] Note, however, that the setting also differs in many important ways.

[7] Sweden's reform entailed an increase in compulsory years of schooling and cash transfers to compensate families for foregone earnings. De-tracking in Finland coincided with the abolition of a vast network of private schools which were placed under municipal authority.

5.2 Background: Germany's education system

5.2.1 Tracking in Germany's education system

Most states in Germany track students into three different types of secondary schools at the end of fourth grade when students are about ten years old:[8] low-achieving primary students usually attend the lower vocational track. The leaving certificate, awarded after five additional years of schooling, qualifies graduates to enroll in upper secondary vocational training courses (i.e. apprenticeships in the dual-system). Students with about average marks from primary education usually attend the intermediate secondary track. After another six years of schooling students are eligible to choose from an extended set of apprenticeships within the dual-system of vocational training. Enrollment in the academic track is recommended to highly-achieving students. This school type is the only secondary school track that awards after eight to nine years of schooling the *Abitur*, the most prestigious school-leaving certificate that permits students to apply to a university.

The streaming procedure varies across states but usually involves teachers formally recommending a secondary school track to parents based on their child's performance. In ten out of 16 states, parents have the final say about the placement of their child, whereas in the remaining six states, recommendations are binding yet parents have the right to let their child take an entry exam or attend test lessons.

Tracks differ in several respects in terms of the quality and quantity of inputs [Dustmann et al., 2014]: first, teachers in the academic track usually receive higher wages. Second, their university degree differs in terms of requirements, is more subject-oriented, and also typically takes one additional year to complete. Third, students in the academic track cover more topics and more advanced topics each year and they are often required to attend more hours per week. Finally, Brunello and Checchi [2007] report that the ratio of students to teachers in German schools varied substantially across tracks in 2004 with 11.89 students per teacher in the general track and 21.25 students per teacher in the vocational track.

Tracking across schools may be inconsequential if the education system exhibits a high level of permeability. In principle, German students are allowed to switch between tracks at any time if their academic records justify such a step. However, research on the topic suggests that switching between tracks prior to completion is rare. Mühlenweg [2008], for instance, examines administrative data from Hesse and reports for the school years 2003/2004 and 2005/2006 that more than 96 percent of students in grades five through seven remain in their initial track. Similar results are reported by Dustmann et al. [2014] who find that only two percent of students in the states of Bavaria and Hesse change track. Avenarius et al. [2001] present similar numbers for Lower Saxony. Moreover, Schnepf [2002] points out that students in the academic or upper vocational track are more likely to shift to the lower track than vice versa.On the other hand, upgrading upon completion of one of the two vocational tracks is quite common [e.g. Dustmann et al., 2014] and sometimes cited as evidence for the permeability of Germany's education system. Students in the lower vocational track may switch to the upper vocational track or stay on for an additional year in order to obtain the school leaving-certificate awarded upon completion of the

[8] Exceptions include Berlin and Brandenburg, in which elementary school comprises six grades, and Mecklenburg-West Pomerania, which tracks students after sixth grade but was part of the German Democratic Republic.

upper vocational track. Students in the upper vocational track, in turn, have the opportunity to continue schooling in the academic track if their academic record meets requirements.[9]

5.2.2 The introduction of OS schools in Lower Saxony

Serious attempts to reform Germany's three-tiered education system were first evident during post-war decades when some states started experimenting with less stringent forms of tracking in a few selected schools. The goal at the time was to improve the selection of students into academic careers. However, efforts in this direction were terminated in Lower Saxony in 1964 [Schuchart, 2006]. At the end of the 1950s, a federal advisory board formally recommended a prolongation of comprehensive schooling until sixth grade [Deutscher Ausschuss, 1959], a recommendation that had no effect on policies at the time [von Friedeburg, 1992]. Only Hesse introduced schools that closely resembled these recommendations but would also retain schools in the traditional three-tiered system.

A follow-up body, the *German Education Council*, presented a blueprint for structural reforms that first mentioned the introduction of an *orientation stage* in 1970 [Deutscher Bildungsrat, 1970]. In comparison to earlier plans, the focus was on tracking within schools across subjects and the dissemination of information about possible future careers to students and their parents [Schuchart, 2006]. OS schools were to be completely independent of schools in the three-tiered system. In 1974, however, it became clear that this proposal would not be approved by a majority of states. While all states agreed in principle to changes to the school system, the compromise would leave the decision over whether or not to delay tracking to the states [Ziegenspeck, 2000, p. 81]. While initial trials with OS-type schools were evident in several states, ultimately, only Lower Saxony and the city state of Bremen opted for tracking of all students at the age of twelve.[10]

Our analysis below compares changes across cohort in educational outcomes in students that received schooling in Lower Saxony to changes across cohorts that received schooling in other states of West Germany. We exclude the city states of Bremen, (West-)Berlin, and Hamburg as they differ in many important ways. We also exclude the state of Hesse which introduced an OS-style school but retained schools in the old system. The new school type was thus an alternative, not a replacement. Figure 5.1 depicts the locations of the states we focus on within Germany.[11]

There are several aspects of the design of the OS in Lower Saxony over and above delayed tracking that are important for the present study: first, the administration of OS schools was independent of other secondary schools. This was not the case, for instance, in states such as

[9] Alternatively, they qualify to attend specialized academic track-schools that often have a special focus on a particular subject. Completion of such a specialized academic-track school allows them to apply to universities, although the range of subjects from which they may choose may be limited.

[10] Bremen established six years of primary education. However, from 1957 onward, there was an option to switch to the academic track already upon completing fourth grade [Schuchart, 2006, p. 70].

[11] Out of the six remaining states that will serve as controls in our analysis below, two, Northrhine-Westphalia and Saarland, did not introduce OS schools of any stripe by 1975. In both states, however, the introduction was planned for the second half of the 1970s [Haenisch and Ziegenspeck, 1977, pp. 40ff]. Ultimately, reform efforts would run out of steam or would be prevented through referenda [Rösner, 1981]. While there were experiments with OS-type schools in the two southern states of Bavaria and Baden-Württemberg by 1975, these states eventually still opted for tracking after fourth grade. For instance, Bavaria would delay tracking into the two lower tracks by two years until the early 2000s but tracking after fourth grade into academic and vocational tracks was continued Piopiunik [2013, e.g.]. Rhineland-Palatinate and Schleswig-Holstein re-labeled fifth and sixth grade (as did Northrine-Westphalia eventually) but this did not lead to changes in the tracking regime. In Lower Saxony, OS schools were the norm until they were abolished in 2004.

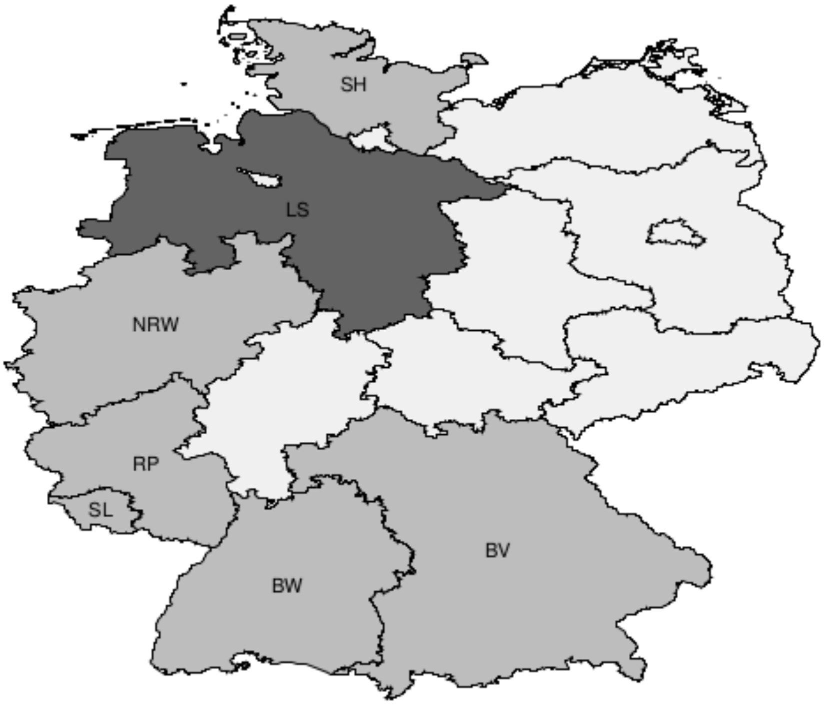

FIGURE 5.1: Treatment (dark gray) and control states (light gray). From North to South: SH (Schleswig-Holstein), LS (Lower Saxony), NRW (Northrhine-Westphalia), RP (Rhineland-Palatinate), SL (Saarland), BW (Baden-Württemberg), and BV (Bavaria).

Schleswig-Holstein and Rhineland-Palatinate in which OS schools were affiliated with secondary schools in the old three-tiered system in what amounts to an inconsequential re-labeling of fifth and sixth grade in the traditional system. Second, the reform entailed some degree of within-class tracking and ability grouping: teachers were expected to adjust tasks and demands within each class according to students' interests and abilities [Avenarius et al., 2001]. Students were also grouped into three different levels of academic achievement during English and math lectures, where the three levels were to roughly reflect the demand levels of secondary school tracks [Ziegenspeck, 2000, pp. 262ff.]. One could thus argue that the comparisons that we rely on in our analysis below are between tracking across schools and a system with a strong element of tracking within schools. Third, OS schools employed teachers from all secondary school tracks on a part-time basis. As every class was to be taught by teachers of all secondary school tracks, students were exposed to teachers with profound knowledge of the everyday practice at the secondary schools in order to ensure a suitable tracking. Fourth, teachers at OS schools were asked to regularly document students' behavior and academic progress according to fixed criteria. This was aimed at nudging teachers towards a more objective judgment of children's academic potential [Hornich, 1976, p.110].

Finally, there was an emphasis on frequent consultations with parents in order to inform them about their child's future perspectives and to explain and to convince them of the ensuing recommendation. This was important since from 1979 onward, parents had the final say about whether or not to heed the schools' recommendation.[12] To summarize, the reform intended to

[12] There is anecdotal evidence that parents at the time had a tendency to overrule recommendations in favor of choices typical of their social class. There was also an understanding among educators that working class parents had to be convinced that their child did not necessarily have to attend one of the two vocational tracks [Ziegenspeck, 2000, pp. 146ff.].

balance the advantages associated with creating homogeneous groups with an effort to dilute the influence of parental background.

5.3 Conceptual framework

5.3.1 Peer effects

Tracking may affect the efficiency of educational production as well as the variation in outcomes through altering the relationship between individual and peer quality.[13] While the early empirical literature on direct peer effects is beset with econometric problems [e.g. Manski, 1993, Sacerdote, 2001], recent studies suggest that students benefit from the presence of higher-achieving peers [see, *inter alia,* Ding and Lehrer, 2007, Lavy et al., 2011, Sacerdote, 2001]. Hence, if tracking succeeds in matching peers of similar quality, one would expect an increase in the overall variation in achievement. Moreover, if own achievement is linked to parental education, one would expect to find a positive relationship between tracking and the gradient between parental education and own education.

At the same time, it is often argued that tracking may allow for efficiency gains in educational production. Tracking will result in more homogeneous groups which, in turn, may allow teachers to tailor lessons more specifically to students' needs. Tracking may thus benefit *all* students. However, empirical studies on this channel are ambiguous [e.g. Epple and Romano, 2011]. While Ding and Lehrer [2007] and Duflo et al. [2011] find a positive effect of increasing peer homogeneity on achievement, Lyle [2009], for instance, finds that a *higher* variance in peer math scores benefits students.

5.3.2 Mis-allocation of students to tracks

A second point relates to the precision with which students' are tracked. While track choice should arguably be based on academic potential, it is often maintained that this is difficult to observe initially and that the signal becomes stronger over time [e.g. Brunello et al., 2007]. Early tracking may thus be associated with a mis-allocation of students to tracks. In particular, it may be the case that non-cognitive skills such as attentiveness become more important for track choice at an early age when cognitive skills are still difficult to observe. Non-cognitive skills, in turn, have been shown to be related to parental background variables such as parents' educational attainment [DiPrete and Jennings, 2012, Segal, 2008].

There is indeed evidence that early tracking fails in separating students effectively by academic potential, particularly in Germany. First, there is considerable overlap in test scores between different school tracks [Baumert et al., 2003]. Lehmann and Peek [1997] find that students of uneducated parents have to score significantly higher in standardized tests in order to receive an academic track recommendation with the same probability as students with educated parents.

Second, track choice is often found to be determined by variables that are arguably unrelated to academic potential. A large literature, for instance, documents relative age-effects in education: Puhani and Weber [2007], Mühlenweg and Puhani [2010], Jürges and Schneider [2011], and Dustmann et al. [2014] all find that students' exact birthday relative to an arbitrary enrollment

[13] Epple and Romano [2011] provide a comprehensive review of the literature on peer effects in education.

cut-off date predicts track choice in Germany.[14] While such findings are indicative of inefficiencies associated with tracking, both Jürges and Schneider [2011] and Dustmann et al. [2014] find no evidence for a persistent effect of relative age on educational outcomes. Moreover, Jürges and Schneider [2011] find no evidence that the age at which states track affects the strength of the relative age effect based on variation across states.

On the other hand, initial misallocation may be inconsequential in a system such as Germany's in which upgrading upon completion is common (see section 5.2.1). In a recent study, Dustmann et al. [2014] use relative age at birth as an instrument to show that track choice is largely inconsequential for marginal students. They argue that built-in flexibilities, the possibility of upgrading upon completion, eventually allow marginal students in the lower tracks to fully compensate for the exposure to low-ability peers.

There is also an increasing interest in gender differences in student achievement, the evolution of such differences over time, and their interaction with tracking. Bedard and Cho [2010] report that tracking is pro-female in Germany in that females are placed in classes with higher average ability. Both Lehmann and Peek [1997] and Jürges and Schneider [2011] report that boys are less likely to be recommended to the academic track in Germany *conditional on academic achievement*, suggesting that girls outperform boys in other relevant dimensions.[15] Interestingly, Jürges and Schneider find no evidence that the gender effect varies across states with different tracking procedures and while the relative age-effect seems to fade over time, the number of female students in the academic track is still greater in ninth grade. The authors conclude that delaying tracking by two years would not reduce gender bias in track attendance.

5.4 Estimating the impact of the reform on educational outcomes

5.4.1 Data and descriptives

In this section, we describe some key variables of our analysis and explain coding decisions. The dataset we use is the German Socio-Economic Panel (GSOEP) v29, a nationally representative survey carried out on an annual basis between 1984 and 2012. See Wagner et al. [2007] for a description of the dataset.

Outcome variables in our analysis are total years of education and two binary variables indicating whether an individual has attained the highest school-leaving certificate (i.e. is eligible to apply to a university without restrictions) and whether an individual has graduated from a university. Means are reported in panel A of table 5.1. 'Years of education' in this dataset refers to the number of years usually required to obtain certain degrees, not to the time spent in education, and includes all stages from primary to tertiary education.[16] Grade repetition will thus not be reflected in this measure of educational attainment.

[14] Schneeweis and Zweimüller [2014] present evidence for Austria in which tracking also occurs at the age of ten.

[15] Jürges and Schneider relate this finding to "differences in verbal and non-cognitive skills at age 10."

[16] The maximum this variable takes is 18 years, 13 years until the highest school-leaving certificate plus five years in order to obtain a university degree. Obtaining a BA degree or a degree from a university of applied sciences usually takes three years. Depending on the type of job, vocational training adds 1.5 or two years to the total. Completing the lower and upper secondary track takes nine and ten years, respectively.

Table 5.1: Means of key variables by state in the estimation sample.

	LS	BV	BW	NRW	RP	SH
Panel A. Outcomes						
Years of education	12.62	12.49	12.69	12.89	12.33	12.55
Eligible for university	0.39	0.37	0.41	0.45	0.36	0.39
University degree	0.23	0.24	0.25	0.26	0.22	0.21
Panel B. Individual background characteristics						
Panel B1. Basic characteristics						
Age	45.20	45.39	44.59	45.44	46.45	44.81
Male	0.48	0.48	0.49	0.47	0.47	0.48
Educated parent	0.37	0.35	0.35	0.36	0.30	0.45
Panel B2. Childhood place of residence was mostly...						
...city.	0.12	0.16	0.12	0.30	0.10	0.15
...large town.	0.16	0.12	0.19	0.23	0.15	0.21
...small town.	0.22	0.22	0.21	0.21	0.22	0.22
...rural area.	0.46	0.46	0.42	0.20	0.49	0.39
Panel B3. Migratory background is...						
...direct	0.02	0.03	0.06	0.04	0.03	0.01
...indirect	0.05	0.08	0.12	0.07	0.06	0.05
Panel B4. Sample information						
Last observed in 2012	0.56	0.57	0.53	0.56	0.60	0.49
Observations	1,593	2,430	1,977	3,484	981	488

LS: Lower Saxony; BV: Bavaria; BW: Baden-Württemberg; NRW: Northrhine-Westphalia; RP: Rhineland-Palatinate/Saarland; SH: Schleswig-Holstein. Based on GSOEP v29-data.

Variables that serve as controls in our analysis capture socio-demographic characteristics such as year of birth, gender, and migrant status. We also include in our analysis a complete set of indicators for the size of the respondent's locality during childhood.[17] Means of these variables are reported in panel B of table 5.1.

Information on parents' educational attainment is available for almost all individuals in the data. There are two variables for each mother and father that relate to the school-leaving certificate and the type of tertiary schooling or vocational training completed. We code a binary variable equal to unity if either the mother or father has (i) attained the highest school-leaving certificate, has (ii) completed the upper vocational track of secondary *and* a full course vocational education, or has (iii) completed training as a clerk, a public health worker, a civil servant, or an engineer, or holds a degree from a tertiary education institution.[18] This results in about 35 percent of our observations being classified as having educated parents. Note that our definition allows for either the father or the mother (or both) to have attained this level of education. Hence, the presence of only one educated parent is assumed sufficient to generate the relevant externality at the household-level [Basu and Foster, 1998].

Logit regressions reveal that our indicator of parents' educational attainment is highly predictive of parents' type of occupation[19] and the likelihood that respondents engaged in extracurricular activities at the age of 15 (such as having actively played a musical instrument or having

[17] Categories are 'no information available,' 'city,' 'large town,' 'small town,' and 'rural area.'

[18] This could be either a regular university (including foreign institutions) or a university of applied sciences.

[19] Conditional on fathers' (mothers') age and year of birth, we find that being educated increases their odds of having worked as a skilled, white-collar worker by a factor greater than four (three). These estimates are highly significant.

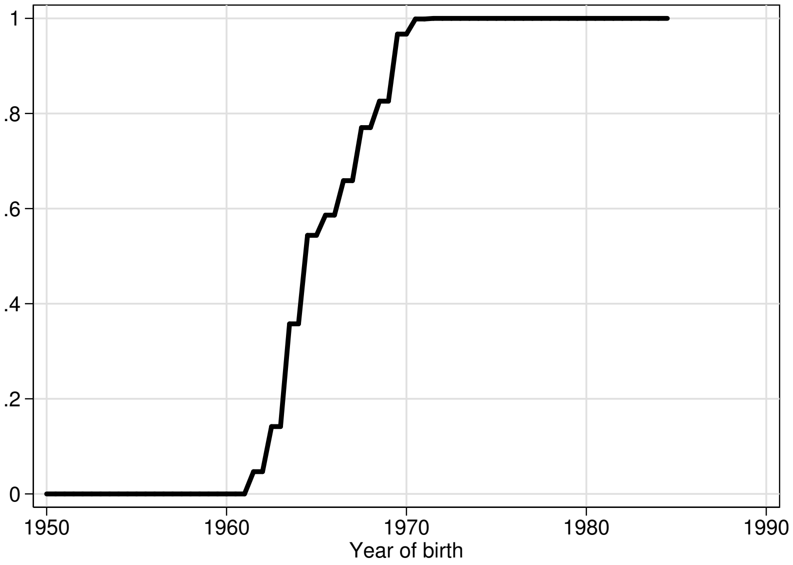

Figure 5.2: OS share against birth year in Lower Saxony.

done sports).[20] To further investigate differences between educated and uneducated parents, we also correlate this variable with an indicator of parents' preoccupation with respondents' academic achievement. This variable is coded on a five-point scale ranging from 'very much' to 'not at all.' We code the highest two levels as a binary variable. Results from logit regression in which we condition on own year of birth, parents' years of birth, parents' age, and the state of the last school visit indicate that having educated parents increases the odds that they were at least 'rather preoccupied' by a factor of two and 2.5 for males and females, respectively.

The GSOEP data do not provide direct information about the type of school individuals attended at the age of ten. We therefore supplement the data with information from the statistical reports on schooling in Lower Saxony (Landesamt für Statistik Niedersachen). These reports tabulate the number of students by grade and birth year in different school types. Combining these data allows us to calculate the percentage of students of one cohort that attended an OS school in a given school year.[21] Enrollment at the time occurred at the age of six and school years start after the end of the summer in July or August. For instance, an individual born in the first half of 1960 was supposed to start schooling in school year 1966/1967 and somebody born in the second half of that year was supposed to start schooling in school year 1967/1968. We therefore associate individuals of a given cohort in Lower Saxony with the share of students in OS schools with one of two subsequent school years depending on whether the individual was born during the first or second half of the year. Figure 5.2 plots the variable OS_c against birth years for individuals that attended schools in Lower Saxony.

There is usually no direct information in the GSOEP dataset about the state in which an individual resided at the age of ten. A second challenge is thus to infer the state in which individuals went to school at that age. Clearly, imputing the current state would confound treatment and control due to inner-German migration. There are several variables in the dataset

[20] Conditional on own year of birth, parents' years of birth, and parents' age, having educated parents increases the odds of having played an instrument and having done sports by a factor of 2.5 and 2.6, respectively.

[21] We ignore private schools which play a negligible role in terms of student in-take in Germany.

that allow us to close in on the required information such as whether the current place of residence was also the childhood place of residence and the state of the last school visit. We also drop observations on individuals that did not attend school in one of the West German states. The exact procedure is described in appendix 5.A. While these steps are reasonable, there will still be misclassification error. Note, however, that confounding treatment and control will usually bias coefficients towards zero. Our estimates may thus be interpreted as lower bounds of the true effect.

Since our interest is exclusively in variables that do not change over time, we discard all observations except the most recent for each individual. This is 2012 for more than half of the observations in our sample but sometimes an earlier year (see panel B4 of table 5.1).

As was discussed in section 5.2, we retain only observations on individuals that are likely to have received their schooling in one of seven states. Due to low case counts and associated privacy issues, two states, Rhineland-Palatinate and Saarland, were until recently treated as one entity in the dataset. We further retain only observations on individuals above the age of 28 as these are likely to have completed their education.[22] In consequence, all individuals in our sample were born before 1985. We also exclude individuals born before 1950.

5.4.2 Empirical specifications

We estimate the causal effect of the reform on educational outcomes and inequality in educational outcomes and conduct this analysis separately for males and females. Our analysis is retrospective: individuals are observed only after they have acquired their education. We employ a DD estimator with a first difference across cohorts and a second difference across individuals in different states. Since the school year starts after the summer, a cohort c consists of all individuals born during the second half of a given year or during the first half of the subsequent year.

Define the treatment variable as the product of the percentage of a cohort in OS schools in Lower Saxony and an indicator variable for Lower Saxony, i.e. $D_{sc} = OS_c \times LS_s$. Note that D_{sc} is not a binary variable but takes on a range of values between zero and unity for cohorts born during the 1960s. We are primarily interested in the effect of the reform on the gradient between parental education and own education. Therefore, we include an interaction between the treatment variable and an indicator variable that is unity whenever the individual has educated parents. Our main model can be written as

$$y_{isc} = \beta_1(D_{sc} \times I[b = high]) + \beta_2 D_{sc} + \mathbf{x}'_{isc}\gamma + \lambda_s^b + \tau_c + \epsilon_{isc}, \tag{5.1}$$

where y_{isc} is the educational outcome for individual i with parental background $b \in \{low, high\}$ who went to school in state s and is a member of cohort c. As outcomes we consider years of education (as defined above) and binary indicators of university eligibility and graduation. For the latter two, the model is a linear probability model (LPM).[23]

[22] Schooling usually starts at the age of six and completing secondary education with the *Abitur*, the highest school-leaving certificate in Germany, requires 13 years of schooling over the time period that we study. Mandatory military service never exceeded two years. If we add five years required to obtain a university degree, the age at which one would complete university is 26.

[23] Estimating probit and logit models does not alter our results.

$\mathbf{x}_{isc}$ is a matrix of time-invariant demographic and socio-economic variables. We include indicators for the respondent's gender, migrant status, and locality size during childhood. Preliminary data analysis suggests that there was considerable variation in the gradient across states prior to the reform. We therefore allow the gradient to differ across states by including state-background-fixed effect denoted λ_s^b. τ_c denote cohort-fixed effects. When the sample is pooled across males and females, λ_s^b, τ_c, and all variables in $\mathbf{x}_{isc}$ are also interacted with the respondent's gender. Finally, ϵ_{isc} is the usual white noise-error term.

The parameters of interest are β_1 and β_2. The former measures the effect of the reform on the gradient and the latter the effect on individuals with uneducated parents. The effect on individuals with educated parents is calculated as the sum over both coefficients.

The reform was implemented during a time of a nation-wide educational expansion, mostly a response to rapid population growth during the 1960s and 1970s. We therefore estimate an alternative model that allows for differences in trends in educational attainment across states by including state-cohort-fixed effects, denoted ϕ_{sc}:

$$y_{isc} = \beta_1(D_{sc} \times I[b = high]) + \mathbf{x}'_{isc}\gamma + \lambda_s^b + \phi_{sc} + \epsilon_{isc}. \tag{5.2}$$

Note that the main effect of the reform is no longer identified and thus excluded here. β_1, however, is still identified from within-cohort, within-state variation.

The identifying assumption for the causal effect of the reform on educational outcomes in the framework above is that trends between states would not have systematically differed in the absence of the reform. While this may be reasonable, the assumption required to obtain a causal effect of the reform on the gradient was so far more demanding: we required that the gradient remains constant over time except for the effect of the reform. We relax the latter assumption by (1) allowing for common changes in the gradient across cohorts and (2) allowing for linear changes in the gradient that vary across states.

A specification that allows for changes in the gradient across cohorts that are common to all states is

$$y_{isc} = \beta_1(D_{sc} \times I[b = high]) + \mathbf{x}'_{isc}\gamma + \lambda_s^b + \psi_c^b + \phi_{sc} + \epsilon_{isc}, \tag{5.3}$$

where ψ_c^b denotes a set of cohort-background-fixed effects. Here, the assumption required to obtain a causal estimate of the effect of the reform on the gradient is that there would not have been any systematic differences in trends *in the gradient* absent the reform.

Since we observe several cohorts that turned ten before the introduction of OS schools in our dataset, we may also relax the assumption of common trends in the gradient in the absence of the reform by replacing cohort-background-fixed effects with linear state- and background-specific trends. This allows us to pick up state-specific changes in the gradient that are discernible prior to the introduction of the reform. The specification is

$$y_{isc} = \beta_1(D_{sc} \times I[b = high]) + \mathbf{x}'_{isc}\gamma + \lambda_s^b + \eta_s^b c + \phi_{sc} + \epsilon_{isc}, \tag{5.4}$$

where η_s^b denotes background- and state-specific parameters that multiply the cohort variable. Note that we again include state-cohort-fixed effects. (5.4) mostly serves a robustness check as linear trends are likely more restrictive than cohort-background-fixed effects.

A general concern one may have in estimating the above specifications is that the reform may have caused selective migration: if families decided to move from or to Lower Saxony in response to the reform and if this decision was correlated with parental education, our estimates would be biased. There are, however, no reports of such responses that we know of. Note also that this would have required parents to move to another state which, in most cases, would imply leaving one's job. It seems implausible that a reform that only affected two grades would have induced this behavior.

If there had been an important migratory response, however, the reform would be associated with a change in the state-specific probability of having educated parents. To check that this is not the case, we regress the binary indicator of having educated parents on the treatment variable as well as a complete set of state- and cohort-fixed effects (results not reported). We also include all controls that we include when we estimate (5.1) and run this regression for the pooled sample and males and females separately. We find no evidence that the reform had an effect on the probability of having educated parents: all coefficients are close to zero and statistically insignificant. Selection due to strategic migration is not a confounding element in our analysis.

5.5 Results

We first consider years of education as the outcome variable. Results from estimating all the above equations are reported in table 5.2. Standard errors clustered at the state-cohort-level are reported in parentheses in all tables that follow.[24]

We exclude the interaction effect between parental background and the reform from (5.1) in column (1) in order to obtain an estimate of the average effect of the reform. Column (2) reports results from estimating (5.1) and column (3) from estimating (5.2), the model including state-cohort-fixed effects. Finally, columns (4) and (5) reports results from estimating (5.3) and (5.4), respectively.

Results reported in panel A are for the pooled sample. There is no indication here of an effect of the reform on years of education on average (column (1)). The coefficient is positive yet small and insignificantly different from zero. If we include the interaction term, however, we find that the reform had a significant negative effect on the gradient between parental and own education (column (2)). The point estimate on the interaction term is -0.61 years and significant at the ten-percent level. We may calculate from the two coefficients reported in column (2) the effect on individuals with uneducated and educated parents, respectively, as the estimate of the main effect, 0.31, and the sum of the two coefficients, -0.31. Both effects fall just short of being significant at the ten-percent level (p-values of 0.12 and 0.16, respectively).

Our conclusions regarding the reform's effect on the gradient does not change when we include state-cohort-fixed effects (column (3)). Our estimate of the effect on the gradient is now somewhat larger in absolute terms at -0.73 years and is significant at the five-percent level. Allowing for a common trend in the gradient over time results in a similar estimate (column

[24] We also experimented with conventional, robust ('sandwich'-type) standard errors, and standard errors clustered at the level of cohorts and states. Clustering at the state-cohort-level turned out to be the most conservative option and not very different from conventional standard errors, robust standard errors, and standard errors clustered at the level of cohort. Clustering at the state-level resulted in much smaller standard errors. We address potential problems with standard errors in a robustness check in section 5.6.1 below.

Table 5.2: Impact of the reform on years of education by gender: OLS estimates.

	(1)	(2)	(3)	(4)	(5)
Panel A. Pooled sample ($N = 10,953$)					
Reform × ed. parents		-0.613*	-0.731**	-0.798**	-1.542**
		(0.315)	(0.339)	(0.358)	(0.712)
Reform	0.076	0.305			
	(0.137)	(0.193)			
R-squared	0.164	0.164	0.190	0.195	0.192
Panel B. Females ($N = 5,698$)					
Reform × ed. parents		-0.066	-0.100	-0.048	-0.867
		(0.350)	(0.358)	(0.406)	(0.768)
Reform	-0.011	0.014			
	(0.180)	(0.226)			
R-squared	0.166	0.166	0.191	0.195	0.192
Panel C. Males ($N = 5,255$)					
Reform × ed. parents		-1.235**	-1.435***	-1.629***	-2.380*
		(0.500)	(0.507)	(0.543)	(1.217)
Reform	0.173	0.617**			
	(0.207)	(0.274)			
R-squared	0.158	0.159	0.186	0.192	0.187
Fixed effects:					
Cohort	✓	✓			
State-cohort			✓	✓	✓
State-background	✓	✓	✓	✓	✓
Cohort-background				✓	
Linear cohort trend:					
State-background-specific					✓

Standard errors clustered at the state-cohort-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. All regressions include dummies for the size of respondents' childhood place of residence and migrant status. When we use the pooled sample, we also include a dummy variable for the respondent's sex and interact it with all other controls (including fixed effects). Based on GSOEP v29-data.

(4)). Including linear, state- and background-specific cohort trends results in an increase in the estimate in absolute terms by a factor of close to two. However, the standard error is also considerably larger. The coefficient remains significant at the five-percent level. This suggests that state-specific trends in the gradient that were present prior to the introduction of the reform cannot explain our finding.

Turning to panels B and C of the same table it is apparent that the effect on the gradient is entirely due to an effect on males. Both the estimate for the overall effect of the reform (column (1)) and the estimate on the interaction term for females are close to zero and insignificant at conventional levels. The coefficient estimate reported in column (5) is large and negative but also insignificant. Estimates in panel C, on the other hand, suggest that the gap in years of education narrowed considerably for males. The coefficient estimate is -1.24 in column (2) and -1.44 when we include state-cohort-fixed effects. Both estimates are statistically significant. A similar estimate is obtained when we allow for cohort-background-fixed effects (column (4)). Again, we

Table 5.3: Impact of the reform on the probability of being eligible for university by gender: OLS estimates.

	(1)	(2)	(3)	(4)	(5)
Panel A. Pooled sample ($N = 10,953$)					
Reform × ed. parents		-0.050	-0.081	-0.101	-0.159
		(0.063)	(0.066)	(0.065)	(0.137)
Reform	-0.008	0.010			
	(0.024)	(0.037)			
R-squared	0.123	0.123	0.149	0.153	0.150
Panel B. Females ($N = 5,698$)					
Reform × ed. parents		0.021	0.002	0.013	0.028
		(0.075)	(0.076)	(0.078)	(0.148)
Reform	-0.023	-0.032			
	(0.032)	(0.045)			
R-squared	0.126	0.126	0.151	0.155	0.153
Panel C. Males ($N = 5,255$)					
Reform × ed. parents		-0.130	-0.173**	-0.227***	-0.391**
		(0.082)	(0.082)	(0.083)	(0.172)
Reform	0.009	0.056			
	(0.032)	(0.044)			
R-squared	0.100	0.101	0.127	0.133	0.129
Fixed effects:					
Cohort	✓	✓			
State-cohort			✓	✓	✓
State-background	✓	✓	✓	✓	✓
Cohort-background				✓	
Linear cohort trend:					
State-background-specific					✓

Standard errors clustered at the state-cohort-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. All regressions include dummies for the size of respondents' childhood place of residence and migrant status. When we use the pooled sample, we also include a dummy variable for the respondent's sex and interact it with all other controls (including fixed effects). Based on GSOEP v29-data.

observe that both the estimate and the associated standard error increase in absolute terms when we replace cohort-background effects with linear cohort trends. However, the estimate remains significant at the ten-percent level.

We conclude that the reform had no effect on average but narrowed the gap in years of education between sons of educated and sons of uneducated parents. Coefficient estimates suggest a decrease in years of education of about three-fifths of a year for the latter and this effects is significant at the five-percent level (column (2)). The sum over both coefficients in column (2), panel C, the effect of the reform on years of education for males with educated parents, is -0.62 and close to being significant at conventional levels (p-value of 0.11).

Next, table 5.3 presents estimates of the effect of the reform on the probability of obtaining the highest school-leaving certificate (and being thus eligible for university). The general pattern we observe is similar to that for years of education: there is no effect of the reform on this outcome on average. The coefficient on the interaction term is negative for the pooled sample.

Table 5.4: Impact of the reform on the probability of graduating from university by gender: OLS estimates.

	(1)	(2)	(3)	(4)	(5)
Panel A. Pooled sample ($N = 10,953$)					
Reform × ed. parents		-0.123**	-0.144**	-0.155***	-0.166
		(0.053)	(0.056)	(0.057)	(0.121)
Reform	-0.000	0.046			
	(0.023)	(0.029)			
R-squared	0.112	0.113	0.143	0.150	0.145
Panel B. Females ($N = 5,698$)					
Reform × ed. parents		-0.042	-0.060	-0.065	-0.132
		(0.064)	(0.065)	(0.072)	(0.159)
Reform	-0.031	-0.014			
	(0.027)	(0.034)			
R-squared	0.104	0.104	0.135	0.142	0.136
Panel C. Males ($N = 5,255$)					
Reform × ed. parents		-0.214**	-0.238***	-0.255***	-0.210
		(0.085)	(0.085)	(0.089)	(0.209)
Reform	0.034	0.111***			
	(0.035)	(0.042)			
R-squared	0.106	0.107	0.137	0.144	0.140
Fixed effects:					
Cohort	✓	✓			
State-cohort			✓	✓	✓
State-background	✓	✓	✓	✓	✓
Cohort-background				✓	
Linear cohort trend:					
State-background-specific					✓

Standard errors clustered at the state-cohort-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. All regressions include dummies for the size of respondents' childhood place of residence and migrant status. When we use the pooled sample, we also include a dummy variable for the respondent's sex and interact it with all other controls (including fixed effects). Based on GSOEP v29-data.

But, in contrast to our results for years of education, it is not statistically significant. The point estimates in this case suggest that the reform lowered the gap in the proportion of individuals that have attained a degree by five to 16 percentage points.

There is no evidence that the reform in any way altered the probability of university eligibility among females. All coefficients are close to zero and insignificant (panel B, table 5.3). On the other hand, the reform seems to have increased equality of opportunity in this respect among males: the coefficient estimates in columns (3) and (4), for instance, suggest that the gap in the probability of attaining eligibility decreased by 17 and 23 percentage points and these estimates are significant at the five- and one-percent level, respectively. As above, the effect seems more pronounced when we include linear trends but is less precisely estimated (column (5)).

As one would expect given the previous results, there is no indication that the reform increased the probability of obtaining a university degree on average (column (1) of table 5.4). The effect

Table 5.5: DD estimates of the reform's effect on educational gradients based on grouped data and by gender.

	Females			Males		
	Years of education (1)	Eligibility (2)	Graduation (3)	Years of education (4)	Eligibility (5)	Graduation (6)
Reform	-0.035 (0.496)	0.011 (0.086)	-0.061 (0.076)	-1.736*** (0.561)	-0.259*** (0.098)	-0.256*** (0.092)
R-squared	0.196	0.183	0.224	0.267	0.262	0.235
Observations	209	209	209	211	211	211

Asymptotic standard errors in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. Observations weighted by the number of individuals in each cohort-state. All regressions include sets of cohort- and state-fixed effects. We also include the share of individuals with direct and indirect migratory background and the share of individuals that grew up in rural areas, small towns, large towns, and cities in each cohort-state-cell as additional controls. Based on GSOEP v29-data.

on the gradient, however, is again seizable even for the pooled sample: once we control for state-cohort-fixed effects (columns (3)–(5)), the coefficient estimates on the interaction term indicate that the gap decreased by about 15 percentage points. These estimates are significant at the one-percent level except when we include state-background-specific cohort trends. However, this is likely to be a power issue as the point estimate hardly changes while the standard errors increase by a factor of about two. Again, the effect is driven by males for which the gap decreases by more than 20 percentage points according to most estimates (panel C).

5.6 Robustness

In this section we conduct different robustness checks that address potential problems with our analysis. We first investigate the robustness of our results to aggregating data at the level of cohort-states. Next, we investigate whether our results are robust to alternative sample restrictions.

5.6.1 Accounting for grouped data

The previous section shows that the result of a decrease in the gradient for males is robust to several alternative estimation methods. One concern may be that the treatment variable varies only at the state-cohort-level. Since the number of clusters is limited, standard errors may be unreliable. Following Angrist and Pischke [2009, p. 313], we collapse our dataset at the state-cohort-level. For each dependent variable of interest, we compute the mean difference between individuals with educated and those with uneducated parents. We regress this gap for males and females separately on the reform variable and include both cohort- and state-fixed effects. We also include variables that capture the share of individuals with direct and indirect migratory background and four variables indicating the level of urbanization of the childhood place of residence. We estimate these models by weighted least squares using the group size as weights.

There are two advantages of this approach: first, it makes our identifying assumption with respect to the reform's effect on the gradient explicit: the gradient is the outcome variable in a standard difference-in-differences equation. Second, it is also more evident now that the asymptotics of our analysis are based on the number of cohort-state-groups. Because the group means are close to being normally distributed, however, the finite-sample properties should closely resemble those of a regression with normally distributed errors. Usual asymptotic standard errors are therefore more likely to be reliable than clustered standard errors.

Results reported in table 5.5 confirm our previous findings: for males, the introduction of OS schools decreased the gap in years of education by about 1.7 years, and the gaps in university eligibility and graduation by 26 and 25 percentage points, respectively. Standard errors are very similar to those reported in columns (4) of tables 5.2–5.4. There is no evidence for an effect of the reform on the gradient for females.

5.6.2 Alternative samples

We now analyze the robustness of our result concerning the effect of the reform on the gradient to changes in the underlying sample. We consider (1) only fully treated or untreated cohorts, (2) only non-migrants, (3) only individuals that grew up in non-rural areas, (4) only individuals that grew up in states with levels and trends in enrollment in pre-primary similar to those of Lower Saxony, and (5) only individuals that grew up with both parents present. In all cases, we focus on the effect on the gradient. We therefore estimate (5.3), the specification that allows for state-cohort-fixed effects and common trends in the gradient. We only consider males here as we did not find any effects for females. Estimating models on the respective samples for females does not allow any other conclusion (results not reported).

A first concern is self-selection into OS schools. The introduction of OS schools in Lower Saxony took almost one decade and during this time period selection may have played a role. Given our findings above, it seems plausible that educated parents would have decided to send their children to schools in the old, three-tiered system as long as this was possible. In any case, the co-existence of two tracking regimes potentially produces outcomes that differ from those under complete de-tracking.[25] We therefore also estimate (5.3) including only cohorts that were either fully treated or fully untreated, i.e. cohorts born before 1962 or after 1971.

A second subsample considers only non-migrants. The effect of such a reform on migrants may differ significantly and there may also be problems in classifying the educational attainment of migrants' parents. Also, Lower Saxony has one of the smallest migrant populations in our sample.

A third subsample excludes individuals that grew up in rural areas. This is motivated by the fact that Lower Saxony is thinly populated in comparison to other states. It is the second-largest state of Germany in terms of area yet only the fourth-largest in terms of inhabitants. Kramer [2002] documents an increase in the access to upper vocational and academic track schools especially in rural areas of Lower Saxony following the construction of new schools during the 1960s and early 1970s. It may thus be that a differential change in access to different types of schools during the 1970s confounds our results.

[25] Recall that this was the reason for us to exclude Hesse, a state in which the two systems exist next to each other.

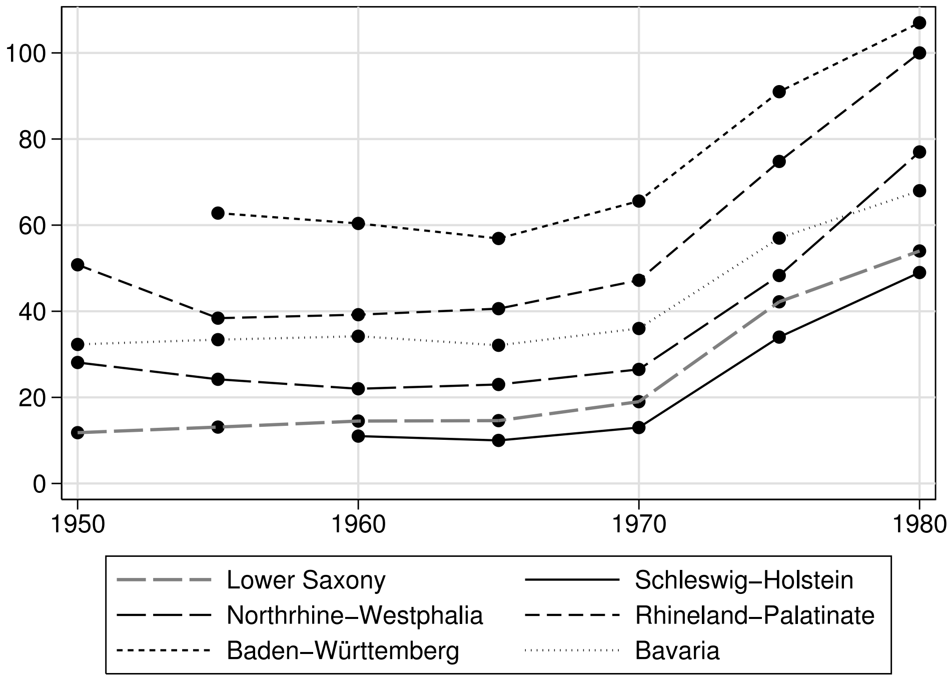

Figure 5.3: Pre-school capacity per 100 children between the age of three and six by state. *Source:* Erning et al. [1987, p. 37].

An additional concern relates to the percentage of children enrolled in pre-primary, an educational institution that has been linked to inequality in educational outcomes [Schütz et al., 2008]. Pre-school enrollment has repeatedly been shown to affect both cognitive and behavioral outcomes, particularly among disadvantaged children. See Barnett [1992], Currie [2001], Cunha et al. [2010], and Blau and Currie [2006] for reviews and Garces et al. [2002], Belfield et al. [2006], Magnuson et al. [2007], Berlinski et al. [2009], Schlotter and Wößmann [2010], and Schlotter [2011] for recent within-country evidence. Therefore, systematic differences in trends and levels in pre-school enrollment between states may affect outcomes and thus confound our results.

Differences in pre-primary enrollment are unlikely to explain our results for two reasons. First, much of the evidence on the effect of pre-school education comes from the US yet pre-school institutions in Germany are very different in that they offer mainly supervision and are not directly targeted towards disadvantaged segments of the population [Schlotter and Wößmann, 2010]. While Schlotter [2011] finds some evidence that pre-school education positively affects assertiveness and the ability to make friends, Schlotter and Wößmann [2010] find no evidence for an effect on reading test scores.

Second, while there are differences in the levels of pre-school education, there are no major differences in trends. Our data do not contain any information about whether respondents attended any pre-school institution and official records by state are not readily available. The number of places available per one hundred children between the ages of three and six are tabulated, however, in Erning et al. [1987, p. 37] and we reproduce this information in figure 5.3. We observe that pre-school capacity was lower in Lower Saxony than in all other states except Schleswig-Holstein and that capacity increased substantially in all states from 1970 onwards. If

Table 5.6: Impact of the reform on inequality in educational outcomes for males: robustness to alternative sample definitions.

	excl. partially treated (1)	excl. migrants (2)	excl. rural (3)	excl. BW and RP (4)	excl. incomplete families (5)
Panel A. Years of education.					
Reform × ed. parents	-1.872***	-1.744***	-1.822**	-1.485***	-1.691**
	(0.634)	(0.544)	(0.756)	(0.520)	(0.665)
R-squared	0.197	0.197	0.212	0.192	0.204
Panel B. University eligibility.					
Reform × ed. parents	-0.237**	-0.250***	-0.235**	-0.205**	-0.223**
	(0.101)	(0.085)	(0.108)	(0.079)	(0.095)
R-squared	0.142	0.139	0.158	0.138	0.141
Panel C. University degree.					
Reform × ed. parents	-0.333***	-0.265***	-0.277**	-0.232***	-0.285***
	(0.092)	(0.091)	(0.124)	(0.085)	(0.104)
R-squared	0.161	0.151	0.166	0.143	0.151
Observations	3,327	4,651	3,295	3,815	4,636

Standard errors clustered at the state-cohort-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. All regressions include sets of dummies capturing the size of respondents' childhood place of residence and migrant status. Based on GSOEP v29-data.

the effect of enrollment would be linear, at least, we would not expect trends to be systematically correlated with the roll-out of OS schools in Lower Saxony.

It may be that the effect of pre-school education depends on the level of enrollment [Schütz et al., 2008]. This will be the case, for instance, if there are positive effects of enrollment on educational attainment and children of educated parents are more likely to enroll initially. In this case, the effect of enrollment will follow an inverted u-pattern with equality of opportunity decreasing at low levels of enrollment and increasing at high levels. Schütz et al. find some evidence for such a relationship and report a turning point at an enrollment rate at about 60 percent. Note that this would result in *decreasing* inequality of opportunity in states with levels of enrollment near the turning point around 1970 (such as Baden-Württemberg) and *increasing* inequality in states with low levels of enrollment such as Lower Saxony.

Nevertheless, as an additional robustness check, we investigate the sensitivity of our results to excluding both Baden-Württemberg and Rhineland-Palatinate, two states that differed the most from Lower Saxony in that they had the highest levels of pre-school enrollment (see figure 5.3).

A final concern may be that systematic differences in family's demographic make-up confound our results. In particular, the absence of a father or a mother may both affect our indicator of parental education and may also be related to educational outcomes. While it is not clear why there should be systematic differences between states, we nevertheless investigate the robustness of our results to the exclusion of all individuals that did not grow up in complete families, i.e. with a mother and father present. The information is obtained from two questions in the dataset:

Table 5.7: Impact of the reform on parental involvement: OLS estimates.

	Females			Males		
Parents' education	All (1)	High (2)	Low (3)	All (4)	High (5)	Low (6)
Reform	-0.047 (0.04)	-0.080 (0.07)	-0.042 (0.05)	-0.029 (0.03)	0.015 (0.07)	-0.040 (0.04)
Observations	5,698	2,045	3,653	5,255	1,875	3,380
R-squared	0.06	0.10	0.07	0.09	0.11	0.10

Standard errors clustered at the state-cohort-level in parentheses. *, **, and *** denote significance at the ten-, five-, and one-percent level, respectively. All regressions include state-of-school-visit-, cohort-, and age-fixed effects. Further controls are dummies capturing the size of respondents' childhood place of residence and migrant status. Based on GSOEP v29-data.

parents' death years and a question about whether the respondent had conflicts with the parent at the age of 15.

Results for years of education, university eligibility, and university graduation are reported in panels A, B, and C of table 5.6, respectively. Note that the sample size changes considerably in comparison to the main sample. The samples used in columns (1) through (5) retain 58, 82, 58, 67, and 81 percent of the original sample, respectively. Nevertheless, we find that our estimates for effect of the reform on the gradient do not change significantly and remain highly significant throughout. They range from -1.48 to -1.87 and are qualitatively similar when we use university eligibility or graduation as the outcome variable. Overall, none of the above concerns seems to affect our main result of an attenuating effect of the reform on the gradient for males.

5.7 Effect of the reform on parental involvement

Previous sections show that the reform had no effect on educational outcomes on average. The effects on females that we find were small and insignificant. For males, on the other hand, we find that the effect varied by parental education with positive effects on individuals with uneducated parents and negative effects on individuals with educated parents. What accounts for this increase in equality of opportunity for males? In this section, we investigate one plausible channel through which the effects of the reform may have operated: parental preoccupation with their children's academic achievement.

The OS placed special emphasis on engaging parents with low educational attainment with the aim to increase the likelihood that they consider sending their children to the academic track granted teachers judged them suitable. Increased involvement of parents may thus be one channel through which educational attainment increased in some groups. Our data contain one item that captures the self-reported involvement of parents' with their children's academic performance. While we would prefer objective measures of parental involvement to self-reports and the question does not reference any particular grade, we are cautiously optimistic that this item serves us as a proxy. As discussed in section 5.4, the respective question asked respondents to rank how much parents cared about academic achievement on a five-point scale from 'very much' to 'not at all' and we code the top two responses as a binary variable. We regress this variable on the reform variable, our usual control variables, as well as fixed effects for schooling state and cohort. We also include age-fixed effects as the perception and judgment of past parental involvement may change with age. We run this regression separately for males and females and further disaggregate the analysis by parental education. Results are reported in table 5.7. While these coefficients are mostly negative and similar across subgroups, they are all insignificantly different from zero. There is thus no indication that the introduction of the OS increased parents' preoccupation with their children's academic achievement.

5.8 Discussion

Results presented in previous sections suggest that the reform achieved for male students what it was intended to achieve: it attenuated the importance of parental background for students' educational prospects. There is no evidence for a trade-off between efficiency and equity; the reform did not seem to have lowered overall educational attainment.

The effect on total years of education is comparatively large given that the reform delayed tracking by only two years: the effect we estimate on the gradient for males ranges from a reduction by 1.4 to 2.4 years of education. The preferred estimate is obtained from specifications that allow for common changes in the gradient across cohorts and state-cohort-fixed effects, i.e. the one reported in column (4), panel C of table 5.2, which indicates a decrease in the gap by 1.6 years. The unadjusted gradient in Lower Saxony prior to the reform was 3.2 years; hence, the reform lowered the gap by about 50 percent. The raw gap was 1.9 years on average for cohorts fully exposed to the reform. The reform's negative effect on the gradient has thus been to some extent abrogated by a modest increase in the gradient evident for other states.

Returns to one additional year of education for males in Germany are usually estimated to be around seven percent and often somewhat higher for tertiary education [Ammermüller and Weber, 2005, Lauer and Steiner, 2000]. The implied changes in hourly wages for males with educated and uneducated parents that we find, $(-1.235 + 0.617) \times 0.07 \approx -4.3$ percent and $0.617 \times 0.07 \approx 4.3$ percent, respectively, would thus be neither very large nor negligible. They are well in line with findings in Meghir and Palme [2005] who report estimates of -7.7 percent and 3.1 percent, respectively. The main difference is that we find no effects for females while Meghir and Palme find a positive effect for females with uneducated fathers (about 3.8 percent) and a negative effect for females with educated fathers (about -4.2 percent).

As one would expect in the presence of selection effects for partially treated cohorts, our estimates show a tendency to increase in absolute terms when we exclude cohorts that may have been in a position to choose between school systems. This suggests that a partial reform that allows parental discretion in choosing among systems may have very different effects from a complete reform.

A plausible explanation of our main findings is that tracking age interacts with gender differences in the development of non-cognitive skills. This would require that educated parents have an advantage in the production of these skills, that the resulting advantage dissipates over time (for instance, through in-class peer effects), and that at least one of the above is more relevant for boys than for girls. If these skills enhance learning or educators reward these skills, we would expect to find that delayed tracking is associated with a less prominent role of parental education for boys.

Differences in the levels and trajectories of cognitive and non-cognitive development between boys and girls are well-established in the literature. Matthews et al. [2009] find evidence for gender differences in self-regulation, the ability to control behavior, cognitions, and emotions, for kindergarten children with females outperforming boys in self-control but not in academic achievement. Lenroot et al. [2007] find important differences in the trajectories of brain development between boys and girls with boys generally trailing behind.

Recently, gender differences in non-cognitive skills have directly been linked to the gender gap in academic achievement. Based on data from the US, DiPrete and Jennings [2012] show that while there are no gender gaps in the returns to social and behavioral skills for children from kindergarten though fifth grade, girls lead boys by nearly 0.4 standard deviations at the start of kindergarten. They demonstrate that this gap grows over time and explains a considerable fraction of the gender gap in academic outcomes at that age. Moreover, these skills are significantly related to parents' socio-economic status and the presence of a father. While they reckon that this may reflect teachers rewarding social and behavioral skills, they also find evidence that these skills enhance learning. This is in line with Kerr et al. [2013] who find benefits from comprehensive schooling in terms of higher cognitive skills for students from disadvantaged backgrounds.

5.9 Conclusion

Recent research suggests that design features of education systems are an important determinant of the strength of the relationship between parental and own education. While tracking has often been found to increase variation in outcomes and the salience of parental background, it is often credited with increasing the efficiency of teaching through an increase in class homogeneity.

In this study, we investigate whether and how the introduction of delayed tracking in one of Germany's federal states in the 1970s affected educational outcomes of individuals differentiated by gender and parental education. Based on a difference-in-differences estimation strategy, we find no evidence that the reform led to a decrease in educational attainment for cohorts affected, that is, there is no evidence that tracking increases efficiency. We present strong evidence, however, that the reform is associated with increased equality of opportunity. The effect is entirely driven by males: the reform benefited males with uneducated parents at the expense of males with educated parents. These findings are robust to to a number of alternative specifications and restrictions imposed on the underlying sample. There is no indication that the reform had any effect on females.

We find no evidence that the reform affected parental involvement, one of its main aims. It seems plausible that the gender differences we observe are related to changes in the development of non-cognitive skills for males. However, more work on the precise transmission channel is clearly warranted. The interaction between gender differences in skill development and tracking seems an interesting alley for future research.

5.A Inferring the state of schooling at age ten

We proceed as follows in order to narrow down the state in which individuals received schooling at the age of ten:

1. We impute the state of the last school visit whenever the exact state cannot be inferred. This information was only gathered in one year, so there are many individuals in the dataset for which it is not available.

2. If neither the exact state nor the state of the last school visit is available, we resort to information from a question about the childhood state of residence. All individuals were asked whether they still live in the same place in which they lived during their childhood. While 'childhood' does not seem particularly well defined, we impute the first state in which an individual was encountered whenever the answer is 'yes.'

3. If all of the above fails, we impute the current state.

4. Finally, we remove individuals from our sample that match one of the following criteria: first, all individuals were asked about the place in which they lived in 1989, shortly before Germany's re-unification. If somebody states that she lived in East Germany and was born before 1978, it is nearly impossible that she received her schooling in one of the Western German states. We thus drop all these observations from our sample. Second, the dataset contains information about the year in which an individual migrated to Germany. Together with information on birth years, we can thus calculate the age at which an individual migrated. We drop all migrants that moved to Germany only after they turned eleven.

Table 5.8 reports the frequencies with which individuals were classified according to the above steps by states. It shows that for the majority of individuals, we make the assumption that the state of last school visit is also the state they lived in at the age of ten. Only ten to 15 percent of individuals are classified based on the first state in which they are observed.

Table 5.8: Provenance of information on state of school visit at the age of ten.

Schooling state is...	LS	BV	BW	NRW	RP	SH
...state at age ten.	0.03	0.04	0.05	0.04	0.02	0.02
...state of last school visit.	0.71	0.70	0.67	0.68	0.68	0.73
...childhood state.	0.15	0.14	0.14	0.16	0.19	0.14
...first state in which observed.	0.10	0.12	0.13	0.12	0.10	0.11
Observations	1,593	2,430	1,977	3,484	981	488

LS: Lower Saxony; BV: Bavaria; BW: Baden-Württemberg; NRW: Northrhine-Westphalia; RP: Rhineland-Palatinate/Saarland; SH: Schleswig-Holstein. Based on GSOEP v29-data.

Bibliography

AbouZahr, C. (2011). New Estimates of Maternal Health and How to Interpret Them: Choice or Confusion? *Reproductive Health Matters 19*(37), 117–128.

Akerlof, G. A. (1978). The Economics of "Tagging" as Applied to the Optimal Income Tax, Welfare Programs and Manpower Planning. *American Economic Review 68*(1), 8–19.

Alatas, V., A. Banerjee, R. Hanna, B. A. Olken, R. Purnamasari, and M. Wai-Poi (2013). Ordeal Mechanisms and Targeting: Theory and Evidence from a Field Experiment in Indonesia. *NBER Working Paper 19127*.

Alatas, V., A. Banerjee, R. Hanna, B. A. Olken, and J. Tobias (2012). Targeting the Poor: Evidence from a Field Experiment in Indonesia. *American Economic Review 102*(4), 1206–1240.

Allison, P. (2000). Problems with Fixed-effects Negative Binomial Models. *Unpublished: University of Pennsylvania*.

Ammermüller, A. and A. M. Weber (2005). Educational Attainment and Returns to Education in Germany—An Analysis by Subject of Degree, Gender and Region. *Centre for European Economic Research (ZEW) Discussion Paper 05-17*.

Amo-Yartey, C. (2008). Tackling Burkina Faso's Cotton Crisis.

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, New Jersey: Princeton University Press.

Arellano, M. and S. Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies 58*, 277–297.

Arellano, M. and O. Bover (1995). Another Look at the Instrumental Variable Estimation of Error-component Models. *Journal of Econometrics 68*, 29–51.

Atkinson, A. B. (1995). On Targeting Social Security: Theory and Western Experience with Family Benefits. In D. van de Walle and K. Nead (Eds.), *Public Spending and the Poor: Theory and Evidence*, pp. 25–68. John Hopkins University Press.

Attaran, A. (2005). An Immeasurable Crisis? A Criticism of the MDGs and Why They Cannot be Measured. *PlOS Medicine 2*(10).

Avenarius, H., H. Döbert, G. Knauss, H. Weishaupt, and M. Weiß (2001). Stand und Perspektiven der Orientierungsstufe in Niedersachsen. Technical report, Deutsches Institut für Internationale Pädagogische Forschung.

Bah, A. (2013). Finding the Best Indicators to Identify the Poor. *TNP2K Working Paper 01—2013*.

Bardhan, P. and D. Mookherjee (2005). Decentralizing Antipoverty Program Delivery in Developing Countries. *Journal of Public Economics 89*, 675–704.

Barnett, A. G., J. C. van der Pols, and A. J. Dobson (2005). Regression to the Mean: What It is and How to Deal with It. *International Journal of Epidemiology 34*(1), 215–220.

Barnett, B. J., C. Barrett, and J. R. Skees (2008). Poverty Traps and Index-Based Risk Transfer Products. *World Development 36*(10), 1766–1785.

Barnett, S. (1992). Benefits of Compesatory Preschool Education. *Journal of Human Resources 27*(2), 279–312.

Barrett, C. B. and P. A. Dorosh (1996). Farmers' Welfare and Changing Food Prices: Nonparametric Evidence from Rice in Madagascar. *American Journal of Agricultural Economics 78*(3), 656–669.

Basu, K. and J. E. Foster (1998). On Effective Literacy. *Economic Journal 108*(451), 1733–1749.

Bauer, P. and R. T. Riphahn (2006). Timing of School Tracking as a Determinant of Intergenerational Transmission of Education. *Economics Letters 91*, 90–97.

Baumert, J. and G. Schümer (2001). Schulformen als Selektionsbedingte Lernmilieus. In J. Baumert, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, P. Stanat, K.-J. Tillmann, and M. Weiß (Eds.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im Internationalen Vergleich*, pp. 454–467. Opladen, Germany: Leske + Budrich.

Baumert, J., U. Trautwein, and C. Artelt (2003). Schulumwelten—Institutionelle Bedingungen des Lehrens und Lernens. In J. Baumert, C. Artelt, E. Klieme, M. Neubrand, M. Prenzel, U. Schiefele, W. Schneider, K.-J. Tillmann, and M. Weiß (Eds.), *PISA 2000: Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland*, pp. 261–331. Opladen, Germany: Leske + Budrich.

Baumert, J., R. Watermann, and G. Schümer (2003). Disparitäten der Bildungsbeteiligung und des Kompetenberwerbs: Ein Institutionelles und Individuelles Modetationsmodell. *Zeitschrift für Erziehungswissenschaften 6*(1), 46–72.

Bedard, K. and I. Cho (2010). Early Gender Test Score Gaps across OECD Countries. *Economics of Education Review 29*, 348–363.

Belfield, C. R., M. Nores, S. Barnett, and L. Schweinhart (2006). The High/Scope Perry Preschool Program. *Journal of Human Resources 41*(1), 162–190.

Bengtson, N. (2010). How Responsive is Body Weight to Transitory Income Changes? Evidence from Rural Tanzania. *Journal of Development Economics 92*, 53–61.

Berlinski, S., S. Galiani, and P. Gertler (2009). The Effect of Pre-primary Education on Primary School Performance. *Journal of Public Economics 93*, 219–234.

Besley, T. and S. Coate (1992a). Understanding Welfare Stigma: Taxpayer Resentment and Statistical Discrimination. *Journal of Public Economics 48*, 165–183.

Besley, T. and S. Coate (1992b). Workfare versus Welfare: Incentive Argument for Work Requirements in Poverty-Alleviation Programs. *The American Economic Review 82*(1), 249–261.

Besley, T. and R. Kanbur (1988). Food-Subsidies and Poverty Alleviation. *The Economic Journal 98*(392), 701–719.

Besley, T. and R. Kanbur (1990). The Principles of Targeting. *World Bank: Policy, Research, and External Affairs (PRE) Working Paper No. 385*.

Betts, J. R. (2010). The Economics of Tracking in Education. In E. A. Hanushek, S. Machin, and L. Woessmann (Eds.), *Handbook of the Economics of Education*, Volume 3, pp. 341–381. Amsterdam: North Holland.

Betts, J. R. and J. R. Shkolnik (2000). Key Difficulties in in Identifying the Effects of Ability Grouping on Student Achievement. *Economics of Education Review 19*(1), 21–26.

Bigman, D. and H. Fofack (2000). Geographic Taregting for Poverty Alleviation: An Introduction to the Special Issue. *World Bank Economic Review 14*(1), 129–146.

Binswanger, H. P. and J. McIntire (1987). Behavioral and Material Determinants of Production Relations in Land-Abundant Tropical Agriculture. *Economic Development and Cultural Change 36*(1), 73–99.

Binswanger, H. P. and M. R. Rosenzweig (1986). Behavioral and Material Determinants of Production Relations in Agriculture. *Journal of Development Studies 22*(3), 503–539.

Blau, D. and J. Currie (2006). Pre-School, Day Care, and After-School Care: Who's Minding the Kids? In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, Volume 2, pp. 1163–1278. Amsterdam: Elsevier.

Blundell, R. and S. Bond (1998). Inititial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics 87*, 115–143.

Bond, S. (2002). Dynamic Panel Data Models: A Guide to Micro Data Methods and Practice. *Centre for Microdata Methods and Practice (CEMMAP) Working Paper CWP09/02*.

Bonjean, C. A., S. Brunelin, and C. Simonet (2012). Impact of Climate Related Shocks on Child's Health in Burkina Faso. *CERDI Etudes et Documents E 2012.32*.

Boyle, C. F., C. Levin, A. Hatefi, S. Madriz, and N. Santos (2014). Achieving a "Grand Convergence" in Global Health: Modelling the Technical Inputs, Costs, and Impacts from 2016 to 2030. `http://globalhealth2035.org/sites/default/files/working-papers/grand-convergence-2030.pdf`. Accessed: 2015-03-03.

Breusch, T. (1979). Testing for Autocorrelation in Dynamic Linear Models. *Australian Economic Papers 17*, 334–355.

Brunello, G. and D. Checchi (2007). Does School Tracking Affect Equality of Opportunity? New International Evidence. *Economic Policy 22*(52), 781–861.

Brunello, G., M. Giannini, and K. Ariga (2007). The Optimal Timing of School Tracking: A General Model with Calibration for Germany. In L. Woessmann and P. Peterson (Eds.), *Schools and the Equal Opportunity Problem*, pp. 129–156. Cambridge, Mass.: The MIT Press.

Budd, J. W. (1999). Changing Food Prices and Rural Welfare: A Nonparametric Examination of the Côte d'Ivoire. *Economic Development and Cultural Change 41*, 587–603.

Burke, M. (2014). Selling Low and Buying High: An Arbitrage Puzzle in Kenyan Villages.

Camacho, A. and E. Conover (2011). Manipulation of Social Program Eligibility. *American Economic Journal: Economic Policy 3*, 41–65.

Canstañeda, T. (2005). Targeting Social Spending to the Poor with Proxy-Means Testing: Colombia's SISBEN System. *World Bank Human Development Network Social Portection Unit Discussion paper 0529*.

Carter, M. R. (1997). Technology, and the Social Articulation of Risk in West African Agriculture. *Economic Development and Cultural Change 45*(3), 557–590.

Carter, M. R. and T. J. Lybbert (2012). Consumption Versus Asset Smoothing: Testing the Implications of Poverty Trap Theory in Burkina Faso. *Journal of Development Economics 99*, 255–264.

Citro, C. and R. Michael (1995). *Measuring Poverty: A New Approach.* Washington, D.C.: National Academy Press.

Clasen, T., W.-P. Schmidt, T. Rabie, I. Roberts, and S. Carncross (2007). Interventions to Improve Water Quality for Preventing Diarrhoea: Systematic Review and Meta-analysis. *British Medical Journal 39*(1), i193–i205.

Clemens, M. (2004). The Long Walk to School. International Education Goals in Historical Perspective. *Center for Global Development Working Paper 37*.

Clemens, M., C. Kenny, and T. Moss (2007). The Trouble with the MDGs: Confronting Expectations of Aid and Development Success. *World Development 35*(5), 735–751.

Coady, D., M. Grosh, and J. Hoddinott (2004). Targeting Outcomes Redux. *The World Bank Research Observer 19*(1), 61–85.

Conning, J. and M. Kevane (2002). Community-Based Targeting Mechanisms for Social Safety Nets: A Critical Review. *World Development 30*(3), 375–394.

Cragg, J. and S. Donald (1993). Testing Identifiability and Specification in Instrumental Variables Models. *Econometric Theory 9*, 222–240.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2010). Interpreting the Evidence on Life Cycle Skill Formation. In E. A. Hanushek and F. Welch (Eds.), *Handbook of the Economics of Education*, Volume 1, pp. 697–812. Amsterdam: North Holland.

Currie, J. (2001). Early Childhood Education Programs. *Journal of Economic Perspectives 15*(2), 213–238.

Das, J., J. Hammer, and K. Leonard (2008). The Quality of Medical Advice in Low-Income Countries. *Journal of Economic Perspectives 22*(2), 93–114.

de Haen, H., S. Klasen, and M. Qaim (2011). What Do We Really Know? Metrics for Food Insecurity and Undernutrition. *Food Policy 36*(6), 760–769.

de Vries, G., M. Timmer, and K. de Vries (2013). Structural Transformation in Africa: Static Gains, Dynamic Losses. *GGDC Research Memorandum 136*.

Deaton, A. (1990). Saving in Developing Countries: Theory and Review. *Proceedings of the World Bank Annual Conference on Developmet Economics 1989*.

Deaton, A. (1991). Saving and Liquidity Constraints. *Econometrica 59*(5), 1221–1248.

Deaton, A. (1992). Saving and Income Smoothing in Côte d'Ivoire. *Journal of African Economies 1*(1), 1–24.

Deaton, A. (1997). *The Analysis of Household Surveys: Microeconometric Analysis for Development Policy.* Baltimore: John Hopkins University Press.

Deaton, A. and S. Zaidi (2002). Guidelines for Constructing Consumption Aggregates for Welfare Analysis. *Living Statndards Measurement Study Working Paper No. 135*.

Demombynes, G. and S. K. Trommlerová (2012). What Has Driven the Decline of Infant Mortality in Kenya? *World Bank Policy Research Working Paper No. 6057*.

Deutscher Ausschuss (1959). Empfehlungen zur Umgestaltung und Vereinheitlichung des Allgemeinbildenden Öffentlichen Schulwesens. In H. Bohnekamp, W. Dirks, and D. Knab (Eds.), *Empfehlungen und Gutachten des Deutschen Ausschusses für das Erziehungs- und Bildungswesen 1953–1965*, pp. 59–115. Stuttgart, Germany.

Deutscher Bildungsrat (1970). Empfehlungen der Bildungskommission: Strukturplan für das Bildungswesen. Technical report, Deutscher Bildungsrat, Stuttgart.

Devarajan, S., M. Miller, and E. Swanson (2002). Goals for Development: History, Prospects, and Costs. *World Bank Policy Research Working Paper No. 2819*.

Ding, W. and S. Lehrer (2007). Do Peers Affect Student Achievement in China's Secondary Schools. *Review of Economics and Statistics 89*(2), 300–312.

DiPrete, T. A. and J. Jennings (2012). Social/Behavioral Skills and the Gender Gap in Educational Achievement. *Social Science Review 41*, 1–15.

Drèze, J. and A. Sen (1989). *Hunger and Public Action.* Oxford: Clarendon Press.

Drèze, J. and P. Srinivasan (1997). Widowhood and Poverty in Rural India: Some Inferences from Households Survey Data. *Journal of Development Economics 54*, 217–234.

Duflo, E., P. Dupas, and M. Kremer (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking. *American Economic Review 101*, 1739–1774.

Dustmann, C. (2004). Parental Background, Secondary School Track Choice, and Wages. *Oxford Economic Papers 56*, 209–230.

Dustmann, C., P. A. Puhani, and U. Schönberg (2014). The Long-Term Effects of Early Track Choice. *IZA Discussion Paper No. 7897*.

Dutta, P., R. Murgai, M. Ravallion, and D. van de Walle (2012). Does India's Employment Guarantee Scheme Guarantee Employment? *World Bank Policy Research Working Paper 6003*.

Easterly, W. (2001). The Lost Decades: Developing Countries' Stagnation in Spite of Policy Reform 1980–1998. *Journal of Economic Growth 6*(2), 135–57.

Easterly, W. (2009). How the Millennium Development Goals are Unfair to Africa. *World Development 37*(1), 26–35.

ECLAC (2006). *Compendium on Best Practices in Poverty Measurement.* Expert Group on Poverty Statistics (Rio Group), United Nations Economic Commission for Latin America and the Caribbean (ECLAC).

ECLAC (2010, September). Economic Survey of Latin America and the Caribbean 2009-2010. Technical report, Economic Commission for Latin America and the Caribbean.

Eicher, C. K. and D. C. Baker (1982). Research on Agricultural Development in Sub-Saharan Africa: A Critical Survey. *MSU International Development Papers No. 1*.

Epple, D. and R. Romano (2011). Peer Effects in Education: A Survey of the Theory and Evidence. In J. Benhabib, M. O. Jackson, and A. Bisin (Eds.), *Handbook of Social Economics*, Volume 1B, pp. 1053–1163. Amsterdam: North Holland.

Erning, G., K. Neumann, and J. Reyer (1987). *Geschichte des Kindergartens: Institutionelle Aspekte, Systematische Perspektiven, Entwicklungsverläufe*, Volume 2. Freiburg, Germany: Lambertus.

Fafchamps, M. and S. Gavian (1997). The Determinants of Livestock Prices in Niger. *Journal of African Economies 6*(2), 255–295.

Fafchamps, M. and J. Pender (1997). Precautionary Savings, Credit Constraints, and Irreversible Investments: Evidence from Semi-arid India. *Journal of Business and Economic Statistics 15*, 180–194.

Fafchamps, M., C. Udry, and K. Czukas (1998). Drought and Saving in West Africa: Are Livestock a Buffer Stock? *Journal of Development Economics 55*, 273–305.

FAO (2005). Special Report of the FAO Crop and Food Supply Assessment Mission to Burkina Faso.

FAO (2010). World Food Dietary Assessment System, Version 2.0. Technical report, Food and Agricultural Organization.

FAO (2014a). Climate Impact on Agriculture. Accessed March 10, 2014.

FAO (2014b). Faostat. Accessed March 18, 2014.

Feenstra, R. C., R. Inklaar, and M. P. Timmer (2013). The Next Generation of the Penn World Table. Available at `www.ggdc.net/pwt`. Accessed 02/08/2015.

Foster, J., J. Greer, and E. Thorbecke (1984). A Class of Decomposable Poverty Measures. *Econometrica 52*(3), 761–766.

Fukuyama, F. (2004). The Imperative of State-Building. *Journal of Democracy 15*(2), 17–31.

Galasso, E. and M. Ravallion (2005). Decentralized Targeting of an Antipoverty Program. *Journal of Public Econhomics 89*, 705–727.

Garces, E., D. Thomas, and J. Currie (2002). Longer-Term Effects of Head Start. *American Economic Review 92*(4), 999–1012.

Gelbach, J. and L. Pritchett (2000). Indicator Targeting in a Political Economy: Leakier can be Better. *Journal of Policy Reform 85*, 705–727.

Gelbach, J. and L. Pritchett (2002). Is More for the Poor Less for the Poor? The Politics of Means-Tested Targeting. *The B.E. Journal of Economic Analysis & Policy 2*(1).

Glass, R. I., A. E. Guttmacher, and R. E. Black (2012). Ending Preventable Child Death in a Generation. *Journal of the American Medical Association*.

Global Investment Framework for Women's and Children's Health (2014). Advancing Social and Economic Development by Investing in Women's and Children's Health: A New Global Investment Framework. *Lancet 383*(9925), 1333–1354.

Godfrey, L. (1978). Testing Against General Autoregressive and Moving Average Error Models when the Regressors Include Lagged Dependent Variables. *Econometrica 46*, 1293–1302.

Greene, W. (2005). Functional Form and Heterogeneity in Models for Count Data. *Foundations and Trends in Econometrics 1*(2), 113–218.

Grosh, M. E. and J. L. Baker (1995). Proxy Means Tests for Targeting Social Programs: Simulations and Speculation. *LSMS Working Paper No. 118*.

Haenisch, H. and J. W. Ziegenspeck (1977). *Die Orientierungsstufe*. Weinheim, Germany, and Basel, Switzerland: Beltz Verlag.

Hall, C. (2012). The Effects of Reducing Tracking in Upper Secondary School: Evidence from a Large-Scale Pilo Scheme. *Journal of Human Resources 47*, 237–269.

Handa, S. and B. Davis (2006). The Experience of Conditional Cash Transfers in Latin America. *Development Policy Review 24*(5), 513–536.

Hanushek, E. A. and L. Woessmann (2006). Does Educational Tracking Affect Performance and Inequality? Differences-in-Differences Evidence Across Countries. *The Economic Journal 116*(510), C63–C76.

Harttgen, K., S. Klasen, and S. Vollmer (2013). An African Growth Miracle? Or: What Do Asset Indices Tell Us About Trends in Economic Performance? *Review of Income and Wealth 59*, S37–S61.

Hausman, J., B. H. Hall, and Z. Griliches (1984). Econometric Models for Count Data With An Application to the Patents R&D Relationship. *Econometrica 52*(4), 909–938.

Hayashi, F. (2000). *Econometrics*. Pinceton: Princeton University Press.

High-Level Panel of Emminent Persons (2013). A New Global Partnership: Eradicate Poverty and Transform Economies Through Sustainable Development. Technical report, United Nations.

High-level Panel on Financing Development (2001). Report of the High-level Panel on Financing for Development. Technical report, United National General Assembly.

Holtz-Eakin, D., W. Newey, and H. Rosen (1988). Estimating Vector Auoregressions with Panel Data. *Econometrica 56*(6), 1371–1395.

Hornich, K. (1976). *Praxis der Orientierungsstufe*. Verlag Herder.

IFRC (2005). World Disasters Report 2005.

IMF (2014). World Economic Outlook 2014: Legacies, Clouds, Uncertainties. Technical report, International Monetary Fund, Washington DC.

INSD (2012). Annuaire Statistique 2012. Technical report, Institut National de la Statistique et de la Démographie.

Jahnke, H. E. (1982). *Livestock Production Systems and Livestock Development in Tropical Africa*. Postfach 4403, Kiel, Germany: Kieler Wissenschaftsverlag Vauk.

Jerven, M. (2010). Random Growth in Africa? Lessons from an Evaluation of the Growth Evidence on Botswana, Kenya, Tanzania and Zambia, 1965–1995. *Journal of Development Studies 46*(2), 274–294.

Johannsen, J. (2006). Operational Poverty Targeting in Peru—Proxy Means Testing with Non-income Indicators. *International Poverty Centre Working Paper nNo. 30*.

Johannsen, J. (2008). *Operational Assessment of Monetary Poverty by Proxy Means Tests: The Example of Peru*. Number 65 in Development Economics and Policy. Peter Lang Verlag.

Jürges, H. and K. Schneider (2011). Why Young Boys Stumble: Early Tracking, Age and Gender Bias in the German School System. *German Economic Review 12*(4), 371–394.

Kahneman, D. (2002). Daniel Kahnam—Biographical. `http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2002/kahneman-bio.html`. Accessed: 2015-04-03.

Kassam, A. (1979). Crops of the West African Semi-Arid Tropics. Technical report, International Crop Research Institute for the Semi-Arid Tropics, Hyderabad, India.

Kazianga, H. and C. Udry (2006). Consumption Smoothing? Livestock, Insurance and Drought in Rural Burkina Faso. *Journal of Development Economics 79*, 413–446.

Kerr, S. P., T. Pekkarinen, and R. Uusitalo (2013). School Tracking and Development of Cognitive Skills. *Journal of Labor Economics 31*(3), 577–602.

Kidd, S. and E. Wylde (2011). Targeting the Poorest: An Assessment of the Proxy Means Test Methodology. Published by the Australian Agency for International Development (AUSAid).

Klasen, S. (2012). Policy Note: MDGs Post-2015: What to Do? *Courant Research Centre 'Poverty, Equity and Growth in Developing and Transition Countries' Discussion Papers 123*.

Klasen, S. and S. Lange (2012). Getting Progress Right: Measuring Progress Towards the MDGs Against Historical Trends. *Courant Research Centre 'Poverty, Equity and Growth in Developing and Transition Countries' Discussion Papers 87*.

Kramer, W. (2002). *50 Jahre Schulenwicklung in Niedersachen.* Oldenburg, Germany: Bibliotheks- und Informationssystem der Universität Oldenburg.

Lancet Commission on Investing in Health (2013). Global Health 2035: A World Converging Within a Generation. *Lancet 382*, 1898–1955.

Landau, K., S. Klasen, and W. Zucchini (2012). Measuring Vulnerability to Poverty Using Long-Term Panel Data. *Courant Research Centre Discussion Paper No. 18, Göttingen University*.

Landesamt für Statistik Niedersachen (2014). Amtliche Schulstatistik. Various years.

Lauer, C. and V. Steiner (2000). Returns to Education in West Germany - An Empirical Assessment. *ZWE Discussion Papers No. 00-04*.

Lavy, V., D. Paserman, and A. Schlosser (2011). Inside the Black Box of Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom. *The Economic Journal 122*(559), 208–237.

Lay, J., U. Narloch, and T. O. Mahmoud (2009). Shocks, Structural Change, and the Patterns of Income Diversification in Burkina Faso. *African Development Review 21*(1), 36–58.

Lehmann, R. and R. Peek (1997). Aspekte der Lernausgangslage von Schülerinnen und Schülern der fünften Jahrgangsstufe an Hamburger Schulen. Bericht über die Untersuchung im September 1996.

Lenroot, R. K., N. Gogrtay, D. K. Greenstein, E. M. Wells, G. L. Wallace, L. S. Clasen, J. D. Blumenthal, J. Lerch, A. P. Zijdenbos, A. C. Evans, P. M. Thompson, and J. N. Gieed (2007). Sexual Dimorphism of Brain Developmental Trajectories During Childhood and Adolescence. *NeuroImage 36*, 1065–1073.

Leonard, K. and M. C. Masatu (2007). Variation in the Quality of Care Accessible to Rural Communities in Tanzania. *Health Affairs 26*(3), w380–w392.

Lindert, K., E. Skoufias, and J. Shapiro (2006). Measuring Vulnerability to Poverty Using Long-Term Panel Data. *World Bank Social Protection Discussian Paper No. 0605*.

Lyle, D. S. (2009). The Effects of Peer Group Heterogeneity on the Production of Human Capital at West Point. *American Economic Journal: Applied Economics 1*(4), 69–84.

Magnuson, K. A., C. Ruhm, and J. Waldfogel (2007). Does Prekindergarten Improve School Preparation and Performance. *Economics of Education Review 26*, 33–51.

Malamud, O. and C. Pop-Eleches (2011). School Tracking and Access to Higher Education Among Disadvantaged Groups. *Journal of Public Economics 95*, 1538–1549.

Manning, R. (2010). The Impact and Design of the MDGs: Some Reflections. *IDS Bulletin 41*(1), 7–14.

Manski, C. F. (1993). Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies 60*, 531–542.

Matlon, P. (1988). Burkina Faso Farm Level Studies: Survey Methods and Data Files. Technical report, International Crop Research Institute for the Semi-Arid Tropics, Hyderabad, India.

Matthews, J., C. C. Ponitz, and F. J. Morrison (2009). Early Gender Differences in Self-Regulation and Academic Achievement. *Journal of Educational Psychology 101*(3), 689–704.

McArthur, J. W. (2014). Seven Million Lives Saved: Under-5 Mortality Since the Launch of the Millennium Development Goals. *Brookings Global Economy and Development Working Paper 78*.

McMillan, M. S. and K. Harttgen (2014). What Is Driving the 'African Growth Miracle'? *NBER Working Paper No. 20077*.

McMillan, M. S. and D. Rodrik (2011). Globalization, Structural Change and Productivity Growth. *NBER Working Paper No. 17143*.

Meghir, C. and M. Palme (2005). Educational Reform, Ability, and Family Background. *American Economic Review 95*(1), 414–424.

Miguel, E. and M. K. Gugerty (2005). Ethnic Diversity, Social Sanctions, and Public Goods in Kenya. *Journal of Public Economics 89*, 2325–2368.

Moene, K. O. and M. Wallerstein (2001). Targeting and the Political Support for Welfare Spending. *Economics of Governance 2*(1), 3–24.

Moffitt, R. (1983). An Economic Model of Welfare Stigma. *American Economic Review 73*(5), 1023–1035.

Moll, H. A. (2005). Costs and Benefits of Livestock Systems and the Role of Market and Nonmarket Relationships. *Agricultural Economics 32*, 181–193.

Moratti, M. (2010). Consumption Poverty and Pro-Poor Growth in Bolivia (1999–2007). *University of Sussex Economics Department Working Paper Series No. 13-2010*.

Morris, S. S., P. Olinto, R. Flores, E. Nils, and A. C. Figueiro (2004). Conditional Cash Transfers are Associated with a Small Reduction in the Rate of Weight Gain of Preschool Children in Northeast Brazil. *Journal of Nutrition 134*(9), 2336–2341.

Mühlenweg, A. M. (2008). Educational Effects of Alternative Secondary School Tracking Regimes in Germany. *Schmollers Jahrbuch: Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften 128*(3), 351–379.

Mühlenweg, A. M. and P. A. Puhani (2010). Persistence of the School Entry Age Effect in a System of Flexible Tracking. *Journal of Human Resources 45*, 407–435.

Nickell, S. J. (1981). Biases in Dynamic Models with Fixed Effects. *Econometrica 49*, 1417–1426.

OECD (2003). *Learning for Tomorrow's World: First Results from Pisa 2003.* Paris: OECD (Organization for Economic Cooperation and Development).

OECD (2015). Germany. In *Education Policy Outlook 2015: Making Reforms Happen.* OECD Publishing. `http://dx.doi.org/10.1787/9789264225442-11-en`.

Öhler, H., P. Nunnenkamp, and A. Dreher (2012). Does Conditionality Work? A Test for an Innovative US Aid Scheme. *European Economic Review 56*, 138–153.

Open Working Group (2014). Report of the Open Working Group of the General Assembly on Sustainable Development Goals. Technical report, United National General Assembly.

Park, A. (2006). Risk and Household Grain Management in Developing Countries. *Economic Journal 116*, 1088–1115.

Paxson, C. H. (1992). Using Weather Variability To Estimate the Response of Savings to Transitory Income in Thailand. *American Economic Review 81*(1), 15–33.

Pekkarinen, T., R. Uusitalo, and S. Kerr (2009). School Tracking and Intergenerational Income Mobility: Evidence from the Finnish Comprehensive School Reform. *Journal of Public Economics 93*(7–8), 965–973.

Piopiunik, M. (2013). The Effects of Early Tracking on Student Performance: Evidence from a School Reform in Bavaria. *Ifo Working Paper No. 153*.

Pischke, J.-S. and A. Manning (2006). Comprehensive Versus Selective Schooling in England and Wales: What Do We Know? *NBER Working Paper No. 12176*.

Pritchett, L. and L. H. Summers (1996). Wealthier is Healthier. *The Journal of Human Resources 31*(4), 841–868.

Pritchett, L. and L. H. Summers (2014). Asiaphoria Meets Regression to the Mean. *NBER Working Paper No. 20573*.

Puhani, P. A. and A. M. Weber (2007). Does the Early Bird Catch the Worm? Instrumental Variables Estimates of the Educational Effects of Age at School Entry in Germany. *Empirical Economics 32*, 359–386.

Rajaratnam, J. K., J. R. Marcus, A. Flaxman, H. Wang, A. Levin-Rector, L. Dwyer, M. Costa, A. D. Lopez, and C. J. Murray (2010). Neonatal, Postneonatal, Childhood, and Under-5 Mortality for 187 Countries, 1970–2010: A Systematic Analysis of Progress Towards Millennium Development Goal 4. *Lancet 375*(9730), 1988–2008.

Ravallion, M. (1998). Poverty Lines in Theory and Practice. *LSMS Working Paper No. 133*.

Ravallion, M. (2007). How Relevant is Targeting to the Success of an Antipoverty Program? *World Bank Policy Research Paper No. 4385*.

Ravallion, M. (2008). Miss-targeted or Miss-measured? *Economics Letters 100*(1), 9–12.

Ravallion, M., S. Chen, and P. Sangraula (2007). New Evidence on the Urbanization of Global Poverty. *World Bank Policy Research Working Paper No. 4199*.

Reardon, T., C. Delgado, and P. Matlon (1992). Determinants and Effects of Income Diversification Amongst Farm Households in Burkina Faso. *Journal of Development Studies 28*(2), 264–296.

Reardon, T., P. Matlon, and C. Delgado (1988). Coping with Household-level Food Insecurity in Drought-affected Areas of Burkina Faso. *Wold Development 16*(9), 1065–1074.

Renkow, M., D. G. Hallstronm, and D. d. Karanja (2004). Rural Infrastructure, Transactions Costs, and Market Participataion in Kenya. *Journal of Development Economics 73*(1), 349–367.

Reyburn, H., R. Mbatia, C. Drakeley, I. Carneiro, E. Mwakasungula, O. Mwerinde, K. Saganda, J. Shao, A. Kituna, R. OPlomi, B. M. Greenwood, and C. J. Whitty (2004). Overdiagnosis of Malaria in Patients with Severe Febrile Illness in Tanzania: A Prospective Study. *British Medical Journal 329*(7476), 1–6.

Roodman, D. (2009a). A Note on the Theme of Too Many Instruments. *Oxford Bulletin of Economics and Statistics 71*(1), 0305–9049.

Roodman, D. (2009b). How to xtabond2: An Introduction to Difference and System GMM in Stata. *The Stata Journal 9*(1), 86–136.

Rosenzweig, M. and K. Wolpin (1993). Credit Market Constraints, Consumption Smoothing, and the Accumulation of Durable Production Assets in Low-Income Countries: Investment in Bullocks in India. *Journal of Political Economy 101*, 223–279.

Rösner, E. (1981). *Schulpolitik durch Volksbegehren. Analyse eines gescheiterten Reformversuchs.* Weinheim, Germany: Beltz.

Sacerdote, B. (2001). Peer Effects with Random Assignment: Results for Dartmouth Roommates. *Quarterly Journal of Economics 116*(2), 681–704.

Sakurai, T. and T. Reardon (1997). Insurance in Burkina Faso and its Determinants. *American Journal of Agricultural Economics 79*(4), 1193–1207.

Schady, N. R. (2002). Picking the Poor: Indicators for Geographic Targeting in Peru. *Review of Income and Wealth 48*(3), 417–433.

Schall, T. and G. Smith (2000). Do Baseball Players Regress Toward the Mean? *The American Statistician 54*(4), 231–235.

Schlotter, M. (2011). Age at Preschool Entrance and Noncognitive Skills before School—An Instrumental Variable Approach. *Ifo Institute for Economic Research Working Paper No. 112*.

Schlotter, M. and L. Wößmann (2010). Frühkindliche Bildung und Spätere Nicht-kognitive Fähigkeiten: Deutsche und Internationale Evidenz. *Ifo Institute for Economic Research Working Paper No. 91*.

Schneeweis, N. and M. Zweimüller (2014). Early Tracking and the Misfortune of Being Young. *Scandinavian Journal of Economics 116*(2), 394–428.

Schnepf, S. V. (2002). A Sorting that Fails? The Transition from Primary to Secondary School in Germany. *UNICEF Innocenti Working Paper No. 92*.

Schuchart, C. (2006). *Orientierungsstufe und Bildungschancen: Eine Evaluationsstudie.* Münster, Germany: Waxmann.

Schütz, G., H. W. Ursprung, and L. Wößmann (2008). Education Policy and Equality of Opportunities. *Kyklos 61*(2), 279–308.

Segal, C. (2008). Classroom Behavior. *Journal of Human Resources 43*(4), 783–814.

Sen, A. (1976). Poverty: An Ordinal Approach to Measurement. *Econometrica 844*(2), 219–231.

Sen, A. (1981). *Poverty and Famines.* Oxford: Clarendon Press.

Sen, A. (1987). *Hunger and Entitlement.* Helsinki: World Institute of Develement Economics Research.

Sen, A. (1995). The Political Economy of Targeting. In D. van de Walle and K. Nead (Eds.), *Public Spending and the Poor: Theory and Evidence*, pp. 11–24. John Hopkins University Press.

Sen, A. (1998). Mortality as an Indicator of Economic Success and Failure. *Economic Journal 108*(446), 1–25.

Shanahan, T., J. Overpeck, K. Anchukaitis, J. Beck, J. Cole, D. Dettman, J. Peck, C. Scholz, and J. King (2009). Atlantic Forcing of Persistent Drought in West Africa. *Science 324*(5925), 377–380.

Sivakumar, M. and F. Gnoumou (1987). Agroclimatology of West Africa: Burkina Faso. Technical report, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT).

Skoufias, E., B. Davis, and S. de la Vega (2001). Targeting the Poor in Mexico: An Evaluation of The Selection of Household for PROGRESA. *IFPRI Food Consumption and Nutrition Division Discussion Paper No. 103*.

Stephens, E. C. and C. B. Barrett (2011). Incomplete Credit Markets and Commodity Marketing Behavior. *Journal of Agricultural Economics 62*(1), 1–24.

Stock, J. and M. Yogo (2005). Testing for Weak Instruments in Linear IV Regression. In *Identification and Inference for Econometric Models: Essays in Honor oif Thomas Rothenberg*, Cambridge, pp. 80–108. Cambridge University Press.

Strulik, H. (2008). Geography, Health, and the Pace of Demo-economic Development. *Journal of Development Economics 86*, 61–75.

Sustainable Development Solutions Network (2014). An Action Agenda for Sustainable Development: Report for the UN Secretary-General. `http://unsdsn.org/resources/publications/an-action-agenda-for-sustainable-development/`. Accessed: 2015-03-03.

Svedberg, P. (2002). Undernutrition Overestimated. *Economic Development and Cultural Change 51*(1), 5–36.

Thompson, M. L. and W. Zucchini (1989). On the Statistical Analysis of ROC Curves. *Statistics in Medicine 8*(10), 1277–1290.

Traore, S. and T. Owiyo (2013). Dirty Droughts Causing Loss and Damage in Nothern Burkina Faso. *International Journal of Global Warming 5*(4), 498–513.

Udry, C. (1995). Risk and Savings in Northern Nigeria. *American Economic Journal 85*, 1287–1300.

UNICEF (2013). A Post-2015 World Fit for Children: UNICEF Key Asks on the Post-2015 Development Agenda. `http://www.unicef.org/ceecis/Post_2015_Key_Asks_V01.pdf`. Accessed: 2015-03-03.

USAID (2013). Famine Early Warning Systems Network. Accessed February 21, 2013.

Vandemoortele, J. (2009). The MDG Conundrum: Meeting the Targets Without Missing the Point. *Development Policy Review 27*(4), 355–371.

Veras Soares, F., R. Perez Ribas, and R. Guerreiro Osório (2010). Evaluating the Impact of Brazil's Bolsa Família: Cash Transfer Programmes in Comparative Perspective. *Latin American Research Review 45*(2), 173–190.

Verguet, S., O. F. Norheim, Z. D. Olson, G. Yamey, and D. T. Jamison (2014). Annual Rates of Decline in Child, Maternal, HIV, and Tuberculosis Mortality Across 109 Countries of Low and Middle Income from 1990 to 2013: An Assessment of the Feasibility of Post-2015 Goals. *Lancet Global Health 2*, e698–e709.

Verpoorten, M. (2009). Household Coping in War- and Peacetime: Cattle Sales in Rwanda. *Journal of Development Economics 88*(1), 67–86.

von Friedeburg, L. (1992). *Bildungsreform in Deutschschland: Geschichte und Gesellschaftlicher Widerspruch*. Suhrkamp Verlag.

Wagner, G. G., J. R. Frick, and J. Schupp (2007). The German Socio-Economic Panel Study (SOEP): Scope, Evolution and Enhancements. *Schmollers Jahrbuch 127*(1), 139–169.

Waldinger, F. (2007). Does Ability Tracking Exacerbate the Role of Family Background for Students' Test Scores?

WHO (2014). Children: Reducing Mortality. Fact Sheet No. 178. Available at `http://www.who.int/mediacentre/factsheets/fs178/en/`.

Wodon, Q. (1997). Targeting the Poor Using ROC Curves. *World Development 25*(12), 2083–2092.

Woessmann, L. (2010). Institutional Determinants of School Efficiency and Equity. *Jahrbücher für Nationalökonomie und Statistik 230*(2), 234–270.

Wolpin, K. I. (1982). A New Test of the Permanent Income Hypothesis: The Impact of Weather on the Income and Consumption of Farm Households in India. *International Economic Review 23*(3), 583–594.

Wooldridge, J. M. (1999). Distribution-free Estimation of Some Nonlinear Panel Data Models. *Journal of Econometrics 90*, 77–97.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: The MIT Press.

World Bank (2004). World Development Report 2004: Making Services Work for Poor People. Technical report, The World Bank, Washington DC.

World Bank (2011). Global Monitoring Report 2011: Improving the Odds of Achieving the MDGs. Technical report, The World Bank and the International Monetary Fund, Washington DC.

World Bank (2013). World Development Indicators 2013.

Wouterse, F. (2011). Social Services, Human Capital, and Technical Efficiency of Smallholders in Burkina Faso. *International Food Policy Research Institute Discussion Paper 01068*.

Young, A. (2012). The African Growth Miracle. *Journal of Political Economy 120*(4), 696–739.

Zeldes, S. (1989). Optimal Consumption with Stochastic Income: Deviations from Certainty Equivalence. *Quarterly Journal of Economics 104*(2), 275–298.

Ziegenspeck, J. W. (2000). *Handbuch Orientierungsstufe: Sachstandsbericht und Zwischenbilanz.* Bad Heilbrunn, Germany: Klinkhardt.

Zimmerman, F. J. and M. R. Carter (2003). Asset Smoothing, Consumption Smoothing and the Reproduction of Inequality Under Risk and Subsistence Constraints. *Journal of Development Economics 71*, 233–260.